Sankhyā: The Indian Journal of Statistics
Special issue on Sample Surverys
1999, Volume 61, Series B, Pt. 1, pp. 1-57

SOME CURRENT TRENDS IN SAMPLE SURVEY THEORY AND METHODS*

By J.N.K. RAO
Carleton University, Ottawa

SUMMARY. Beginning with the pioneering contributions of Neyman, Hansen, Mahalanobis and others, a large part of sample survey theory has been directly motivated by practical problems encountered in the design and analysis of large scale sample surveys. Major advances have taken place in handling both sampling and nonsampling errors as well as data collection and processing. In this paper, some current trends in sample survey theory and methods will be presented. After a brief discussion of developments in survey design and data collection and processing, issues related to inference from survey data, resampling methods for analysis of survey data and small area estimation will be studied. Advantages of a conditional design-based approach to inference that allows restricting the set of samples to a relevant subset will be demonstrated. Quasi-score tests based on the jackknife method will be presented. Finally, issues related to model-based methods for small area estimation will be discussed.

1. **Introduction**

The principal steps in a sample survey are survey design, data collection and processing, estimation and analysis of data. We have seen major advances in all those areas, especially in developing methods to handle sampling errors and in the analysis of complex survey data taking account of the design features such as clustering, stratification and unequal selection probabilities. This paper attempts to appraise some current trends in those areas.

Section 2 deals with survey design issues. The need for a total survey design approach is emphasised as well as the need for replicate observations to obtain bias-adjusted estimators of distribution functions and quantiles. Advantages of compromise sample size allocations to satisfy reliability requirements at the

AMS (1991) *subject classification.* 62D05, 62G09.

Key words and phrases. Survey design, data collection, inferential issues, resampling methods, small area estimation.

* Research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This paper was presented as the 8th Annual Morris Hansen Lecture, October 19, 1998.

provincial (or state) level and the subprovincial level are pointed out. Data collection and processing is considered in Section 3. Topics covered include telephone surveys, split questionnaires, ordering of sensitive questions and application of artificial neutral network technology to editing of survey data. Section 4 studies basic inferential issues. Advantages of a conditional design-based approach to inference are demonstrated. Quasi-score tests of hypotheses based on the jackknife method are presented in Section 5. These tests have the advantage that only the reduced model under the hypothesis needs to be fitted and they are invariant to parameter transformations. Finally, issues related to model-based method for small area estimation are discussed in Section 6.

2. **Survey Design**

Survey samplers have paid a lot of attention to sampling errors, and developed numerous methods for optimal allocation of resources to minimize the sampling variance associated with an estimated total (or a mean). But they have given much less attention to total survey error arising from both sampling and nonsampling errors. Fellegi and Sunter (1974), Linacre and Trewin (1993) and Smith (1995) among others emphasised the need for a total survey design approach by which resources are allocated to those sources of error where error reduction is most effective, thus leading to superior survey designs. Linacre and Trewin (1993) applied this approach to the design of a Construction Industry Survey. They entertained 176 resource allocation options and obtained measures of total collection cost and total error based on the root mean squared error (RMSE) criterion for each of the options. A plot of RMSE versus total cost for those options revealed seven cost effective options. Among those, one option provided a good balance between cost and error: RMSE of 3.54% and a cost of \$1.05m compared to the 1984-85 option with RMSE = 3.12% and cost = \$2.63m and the option used in 1988-89 with RMSE= 5.98% and cost = \$1.05m. It is also remarkable that the scatter plot indicated a wide spread with options costing \$0.3m having the same RMSE as options costing ten times as much. Similarly, options with RMSE of 3% can have the same cost as options with RMSE three times as much.

Smith (1995) proposed the sum of component MSEs, rather than the MSE of the estimated total, as a measure of the total error which may be written as $\sum e_j$ with e_j denoting the error from source j. This measure, called total MSE, seems to be more appropriate than the MSE of the estimated total, as demonstrated by Smith. Whichever measure is used, the total survey design approach is not feasible without some knowledge of variances, biases and marginal costs associated with the different error sources. Evaluation studies can be very useful in getting estimates of desired parameters, as demonstrated by Linacre and

Trewin, and thus arriving at a cost effective resource allocation plan. It is important therefore to allocate resources to evaluation studies at the design stage.

It is customary to study the effect of measurement errors in the context of estimating a population total (or a mean). For this case, usual estimators are design-unbiased and consistent under the assumption of zero mean measurement errors. Moreover, customary variance estimators remain valid provided the sample is in the form of interpenetrating subsamples, that is, independent subsamples each providing a valid estimator of the true population total (Mahalanobis, 1946). However, these nice features no longer hold in the case of distribution function, quantiles and some other complex parameters, as demonstrated by Fuller (1995). The usual estimators are biased and inconsistent and thus can lead to erroneous inferences. Bias-adjusted estimates can be obtained if estimates of measurement error variance, σ^2, are available. But customary survey designs do not permit the estimation of σ^2 due to lack of replication. It is important, therefore, to allocate resources at the design stage to estimate σ^2 through replicate observations for a subsample of the sampled units. Fuller (1995) obtained bias-adjusted estimators under the assumption of independent, normally distributed errors. Eltinge (1998) extended Fuller's results to the case of nonnormal errors, using small σ approximations. He applied the methods to data from the U.S. Third National Health and Nutrition Examination Survey (NHANES) to estimate low-hemoglobin prevalence rate among white women aged 20-49 and the fifth percentile of the population hemoglobin distribution.

As noted above, interpenetrating subsamples provide a valid estimate of the total variance of an estimated total in the presence of measurement errors. But such designs are not often used, at least in North America, due to cost and operational considerations. Hartley and Rao (1978) and Hartley and Biemer (1981) provided interviewer and coder assignment conditions that permit the estimation of total variance and its components (sampling, interviewer and coder variances) directly from stratified multistage surveys that satisfy the estimability conditions. Moreover, the theory is applicable to general mixed linear models, thus permitting realistic modelling of interviewer, coder and other effects. For example, Groves (1996) noted that the customary one-way random effect interviewer variance model may be unrealistic because it fails to reflect "fixed" effects of interviewer attributes such as race, age and gender that can affect the responses. Mixed models with both fixed and random effects of interviewers and other sources of error can be accommodated in the Hartley-Rao framework. Unfortunately, current surveys are often not designed to satisfy the estimability conditions and even if they do the required information on interviewer and coder assignments is seldom available at the estimation stage.

In practice, resources are allocated to minimize the sampling variance at the national or provincial (state) level. But such allocations may produce unreliable estimates at the subprovincial level due to small sample sizes at the lower level.

Compromise sample size allocations are often needed to satisfy both reliability requirements. Singh *et al.* (1994) presented an excellent illustration of compromise allocation used in the redesign of the Canadian Labour Force Survey. With a monthly sample of 59,000 households, optimizing at the provincial level yields a coefficient of variation (CV) for "unemployed" as high as 17.7% for Unemployment Insurance (UI) regions. On the other hand, a two-step allocation with 42,000 households allocated in the first step to get reliable provincial estimates and the remaining 17,000 households allocated in the second step to produce best possible UI region estimates reduce the worst case of 17.7% CV for UI regions to 9.4% at the expense of a small increase in CV at the provincial and national levels: CV for Ontario increases from 2.8% to 3.4% and for Canada from 1.36% to 1.51%.

3. **Data Collection and Processing**

3.1 *Data collection.* Telphone surveys have become popular in recent years, at least in developed countries, due to dramatic increases in the telephone coverage rates of household populations and reduced costs of data collection by telephone compared to face-to-face interviewing. Also, Computer Assisted Telephone Interviewing (CATI) has helped to reduce both measurement and data processing errors. Monthly surveys where a sample dwelling remains in the sample for several consecutive months often combine face-to-face contact in the first interview with telephone interviewing in the subsequent months.

Random digit dialing (RDD) methods provide coverage of both listed and unlisted telephone households. Also, ingenious methods, such as the two-stage Mitofsky-Waksberg technique and its refinements (Casady and Lepkowski, 1993), are designed to increase the proportion of eligible numbers in the sample and thus reduce data collection costs. Dual frame approaches are also used to obtain more efficient estimates by combining a sample selected from an incomplete directory (or commercial) list frame with another sample selected by random digit dialing (Groves and Lepkowski, 1986). Costs are also reduced because of the higher proportion of eligible numbers in the list sample. Despite these impressive advances, it is unlikely that telephone sampling will be used in the near future in developing countries, such as India, for household surveys because of the poor telephone coverage rates.

Many large-scale surveys, especially surveys on family expenditure and health, use long questionnaires for data collection. But such surveys can lead to high nonresponse rates and decrease in response quality. This problem may be remedied by splitting the long questionnaire into two or more blocks and administering the blocks to subsamples of sampled households. Wretman (1995) gives an example of a split questionnaire design where the questionnaire is divided into five nonoverlapping blocks, 0-4, of questions and the sample is partitioned at random into four subsamples, 1-4. The questions in block 0, containing questions about basic variables, are administered to all the sampled individuals while questions in block i are administered only to the individuals in subsample $i(= 1-4)$. This design is similar to two-phase sampling so that more efficient estimates can be obtained by using the sample data from block 0 as auxiliary information in the estimation procedure (Wretman, 1995; Renssen and Nieuwenbroek, 1997). Raghunathan and Grizzle (1995) used a multiple imputation approach to obtain estimates of totals and their standard errors, by replacing each "missing" value by two or more imputed values and thus creating multiple completed data sets.

For surveys dealing with sensitive questions, the quality of responses and response rates might depend on the ordering of the questions in the questionnaire. A sensible approach is to order the questions so that most sensitive questions are relegated to the last part of the questionnaire. Cross-over designs can be used in evaluation studies to estimate the residual effects of questions and then order the questions according to increasing size of residual effects. Lakatos (1978) developed relevant theory for cross-over designs, assuming undiminished residual effects which may be more appropriate in the survey context than the usual assumption of only first (or second) order residual effects. Further work on ordering sensitive questions using cross-over and related designs would be useful.

3.2 *Data processing.* It is often necessary to edit the data collected from a survey or a census. The aim of editing is to determine which records are unacceptable, then identify the values of such records that need to be corrected and then correct those values using imputation. It is important to ensure that the editing procedure changes as few values as possible. Fellegi and Holt (1976) and others developed methods for automatic editing of survey data using high-speed computers and assuming that the edit specifications would be given explicitly by subject matter experts. Automatic editing procedures have been implemented by many survey agencies, including the Generalized Edit and Imputation System (GEIS) at Statistics Canada.

Recently, Nordbotten (1995) applied the Artificial Neural Network (ANN) technology to editing of survey data. He reformulated editing as a "feed-forward" neural network with a "hidden" layer. Such a two-layer ANN can be "trained" to edit and impute from a sample of records edited by experts rather than using explicit edit and imputation rules. Nordbotten conducted empirical experiements to study the performance of ANN for editing sample records. He partitioned a sample of n records, each represented as a $1 \times p$ vector of binary elements, at random into two sets: a training set of observed and edited (true) records represented by $n_1 \times p$ matrices R_1 and T_1 and a test set of observed and edited records represented by $n_2 \times p$ matrices R_2 and $T_2(n_1 + n_2 = n)$. The training set $\{R_1, T_1\}$ was used to train the ANN using "Back Propagation" learning algorithm. The trained ANN was then used to transform R_2 to a new matrix T_2^*. The performance of ANN was judged by comparing T_2^* to T_2. The trained ANN was able to edit correctly most of the records in the training set, and the results for the test set were "encouraging". Roddick (1993) earlier proposed a similar edit system based on ANN.

As noted by Nordbotten, not much theoretical support exists on choice of a good ANN "architecture" for an editing model, including the specification of initial values for the weights that represent the "memory" of the model, "learning" rate and the number of "neurons" of the hidden layer. The performance of ANN also may depend on the choice of n_1, the size of the training set. Model training can be resource intensive and time consuming, but commercial programs for neural networks are available. Clearly, further research, both theoretical and empirical, is needed before ANN can be adopted with confidence for editing of sample records.

4. Inferential Issues

Traditional sample survey framework for handling sampling errors uses a design-based approach which leads to valid repeated sampling inferences regardless of the population structure, at least for large samples. "Working" models are often used to choose good designs and efficient design-consistent estimators, but the inferences remain design-based and hence "robust" to model misspecifications. This "model-assisted" approach is discussed in detail by Sarndal *et al.* (1991).

Model-dependent approaches have also been advanced to handle sampling errors (Brewer, 1963; Royall, 1970). The population structure is assumed to follow a specified model, and the model distribution yields valid inferences referring to the particular sample of units that has been drawn. Such conditional inferences are more appealing than repeated sampling inferences, but model-dependent strategies can perform poorly in large samples under model misspecifications, as demonstrated by Hansen *et al.* (1983). By introducing a model misspecification to the assumed model that is not detectable through significance tests for samples as large as 400, they showed that the design-coverage rate of model intervals on the parameter can be substantially smaller than the nominal level and that the coverage rate becomes worse as the sample size increases. On the other hand, design-based intervals performed well with design-coverage rates close to the nominal level. But this result is not surprising because the model-dependent estimator is not design-consistent under their stratified sampling design so that the design performance of the associated model intervals deteriorates as the sample size, n, increases. On the other hand, the design-based estimators they used are design-consistent with improved performance of the associated intervals as n increases. Proponents of the model-dependent approach would argue that design-based comparisons are irrelevant from a conditional inference viewpoint. To address this criticism, one should compare the performances under a conditional inference framework.

To handle nonsampling errors (measurement errors, nonresponse), models are necessary even in the design-based approach. That is, inferences refer to both the design and the assumed model; for example, zero-mean measurement errors or missing at random (MAR) response mechanism. In the case of small areas, sample sizes are too small for traditional estimators to provide reliable precision. By using implicit or explicit models that provide a link to related small areas, increased precision of estimators may be achieved.

To simplify the discussion, suppose that we are estimating the population total of a characteristic of interest, y. A sample s is selected according to a specified sampling design $p(s)$ from a population of size N and the sample data $\{(i, y_i), i \in s\}$ are collected, assuming nonsampling errors are absent. The basic problem of inference is to obtain an estimator $\hat{Y}$, its standard error $s(\hat{Y})$ or coefficient of variation $c(\hat{Y}) = s(\hat{Y})/\hat{Y}$ and associated normal theory intervals, $\hat{Y} \pm z_{\alpha/2}s(\hat{Y})$, on Y from the sample data, assuming n is large enough to justify the normal approximation, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of a $N(0,1)$ variable.

4.1. *Design-based approach.* We assume that the sample design $p(s)$ ensures positive first order inclusion probabilities, π_i, and also positive second order inclusion probabilities, π_{ij}. Such designs permit design-unbiased estimators $\hat{Y}$ and variance estimators $s^2(\hat{Y})$. The basic design-unbiased estimator of $\hat{Y}$ is given by

$$\hat{Y}_{\text{NHT}} = \sum_{i \in s} y_i/\pi_i = \sum_{i \in s} w_i y_i \qquad \ldots (4.1)$$

(Narain, 1951; Horvitz and Thompson, 1952). The coefficient w_i in (4.1) are called the basic design weights. The estimator $\hat{Y}_{\text{NHT}}$ is admissible for *any* design and sample size n, but it can lead to absurd results if the π_i are unrelated to the y_i, as demonstrated by the well-known Basu's (1971) circus elephants example with $n = 1$. Unfortunately, some main stream statisticians believe that Basu's "counter-example" essentially "destroys frequentist sample survey theory" (Lindley, 1996). But this is far from the truth. First, for large n, $\hat{Y}_{\text{HT}}$ is design-consistent even if the π_i are inappropriately chosen (Ghosh, 1992). Secondly, auxiliary information, $\mathbf{x}_i$, with known totals $\mathbf{X}$ and closely related to y_i should be used at the estimation stage through ratio or regression estimation. For example, if $\mathbf{x}_i$ is a scalar, then it is a common practice to use the ratio estimator

$$\hat{Y}_r = (\hat{Y}_{\text{NHT}}/\hat{X}_{\text{NHT}})X, \qquad \ldots (4.2)$$

where $\hat{X}_{\text{NHT}} = \sum_s x_i/\pi_i$. Basu's circus statistician could have saved his job by using the weights of elephants, x_i, in the previous census through (4.2). For $n = 1$, $\hat{Y}_r$ reduces to N times the current weight of the average elephant Sambo selected in the sample. This estimator is identical to the one originally proposed by the circus manager! In the context of surveys with multiple characteristics with some of the y_i unrelated to π_i, Rao (1966) in fact proposed such an estimator and showed that it can be considerably more efficient than $\hat{Y}_{\text{HT}}$. Hájek (1971) suggested the choice $x_i = 1$ in (4.2) for this case which again reduces to N (Sambo's weight) in Basu's example. If the π_i and y_i are weakly correlated and no auxiliary data x_i is available, then Amahia *et al.* (1989) proposed

$$\hat{Y}_A = (1-\rho)\hat{Y}_{\text{NHT}} + \rho\hat{Y}_R \qquad \ldots(4.3)$$

where $\hat{Y}_R = N$ (sample mean) is Rao's estimator and ρ is the correlation coefficient between y_i and π_i. A good guess of ρ is necessary in using $\hat{Y}_A$.

Model-assisted approach provides a formal framework for using auxiliary information, $\mathbf{x}_i$, at the estimation stage. Suppose the "working" model is

$$y_j = \mathbf{x}_j' \boldsymbol{\beta} + e_j, \quad j = 1, \ldots, N \qquad \ldots(4.4)$$

with mean zero, uncorrelated errors e_j and model variance $V_m(e_j) = \sigma^2 q_j$, where the q_j are known constants. The generalized regression (GREG) is then given by

$$\hat{Y}_{gr} = \sum_{i \in s} w_i^* y_i, \qquad \ldots(4.5)$$

where $w_i^* = w_i g_i$ with

$$g_i = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{\text{NHT}})'\hat{\mathbf{T}}^{-1}\mathbf{x}_i/q_i \qquad \ldots(4.6)$$

and $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x}_i \mathbf{x}_i'/q_i$ (Fuller, 1975; Sarndal *et al.*, 1992). The estimator $\hat{Y}_{gr}$ also covers poststratification according to one or more poststratifiers with known marginal counts $\mathbf{X}$ (e.g., projected demographic counts) by using indicator auxiliary variables $\mathbf{x}_i$ to denote the categories of the post-stratifiers. An advantage of $\hat{Y}_{gr}$ is that a single set of weights w_i^* is used for all variables y_i, thus ensuring consistency of figures when aggregated over variables y_i. Also, it is a calibration estimator in the sense $\sum_s w_i^* \mathbf{x}_i = \mathbf{X}$, that is, it ensures consistency with the known totals $\mathbf{X}$. Note that only the totals $\mathbf{X}$ are needed to implement $\hat{Y}_{gr}$, not the individual population values $\mathbf{x}_j$. The estimator $\hat{Y}_{gr}$ provides a unified approach to estimation using auxiliary information. A Generalized Estimation System (GES) based on $\hat{Y}_{gr}$ has been developed at Statistics Canada. Note that the ratio estimator $\hat{Y}_r$ is obtained from (4.5) and (4.6) by choosing $q_i = x_i$ in (4.6); g_i reduces to $X/\hat{X}_{\text{NHT}}$. A drawback of $\hat{Y}_{gr}$ is that some of the weights w_i^* can be negative if n is not large, and such weights may not be acceptable to some users.

The GREG estimator (4.5) can also be motivated, without appealing to a model, by noting that the adjusted weights w_i^* are obtained by minimizing the chi-squared distance $\sum_s (w_i^* - w_i)^2 q_i/w_i$ subject to the benchmark constraints (or calibration equations) $\sum_s w_i^* \mathbf{x}_k = \mathbf{X}$ (Deville and Sarndal, 1992). Other distance measures may also be used. For example, the measure $\sum_s [w_i^* \log(w_i^*/w_i) - w_i^* + w_i]$ leads to positive weights w_i^* but some of the w_i^* can be extreme. Several iterative methods that attempt to meet both benchmark constraints and range restrictions on the weights w_i^* have been proposed in the literature, but a solution may not always exist. Rao and Singh (1997) proposed a method based on ridge regression that forces convergence for a given number of iterations by using a built-in tolerance specification procedure to relax some benchmark constraints while satisfying the range restrictions. The range restrictions on the weights can be further relaxed if necessary to meet lower tolerance levels on the benchmark constraints.

4.2. *Conditional design-based approach.* A conditional design-based approach to inference from survey data has also been proposed. This approach allows us to restrict the set of samples to a "relevant" subset and leads to conditionally valid inferences in the sense that the conditional bias ratio (i.e., ratio of conditional bias to conditional standard error) goes to zero as n increases. Approximately 100 $(1-\alpha)\%$ of the realized confidence intervals in repeated sampling from the conditional set will contain the unknown total Y. Thus the approach attempts to combine the conditional features of the model-dependent approach with the model-free features of the design-based approach.

Holt and Smith (1979) provided compelling arguments for conditional design-based inference, even though this discussion was confined to one-way post-stratification of a simple random sample, in which case it is natural to make inferences conditional on the realized strata sample sizes. When only the over-all totals $\mathbf{X}$ of auxiliary variables, $\mathbf{x}$, are known from external sources, conditioning on $\hat{\mathbf{X}}_{\mathbf{NHT}}$ may be justified because $\hat{\mathbf{X}}_{\mathbf{NHT}}$ is "approximately" an ancillary statistic when $\mathbf{X}$ is known and $\hat{\mathbf{X}}_{\mathbf{NHT}} - \mathbf{X}$ provides a measure of imbalance in the realized sample. Rao and Liu (1992), Rao (1994) and Casady and Valliant (1993) studied conditional inference when only $\mathbf{X}$ is known. They showed that the optimal regression estimator leads to conditionally valid inferences given $\hat{\mathbf{X}}_{\mathbf{NHT}}$. This estimator is given by

$$\hat{Y}_{\text{opt}} = \hat{Y}_{\text{NHT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\text{NHT}})'\hat{\mathbf{B}}_{\text{opt}}, \qquad \ldots(4.7)$$

where $\hat{\mathbf{B}}_{\text{opt}} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$ with $\hat{\boldsymbol{\Sigma}}_{xx}$ and $\hat{\boldsymbol{\Sigma}}_{xy}$ denoting the estimated covariance matrix of $\hat{\mathbf{X}}_{\text{NHT}}$ and the estimated covariance of $\hat{\mathbf{X}}_{\text{NHT}}$ and $\hat{Y}_{\text{NHT}}$ (Montanari, 1987). The estimator $\hat{Y}_{\text{opt}}$ can also be written in the form $\sum_s \tilde{w}_i y_i$, so that a single set of weights $\tilde{w}_i$ is used for all variables y_i, as in the case of $\hat{Y}_{\text{gr}}$. For example, under stratified simple random sampling without replacement, the basic design weights are given by $w_{hi} = N_h/n_h$, where n_h and N_h denote the hth stratum sample size and population size, and $\hat{Y}_{\text{opt}} = \sum_s \tilde{w}_{hi} y_{hi}$ with $\tilde{w}_{hi} = w_{hi} g_{hi}$ and

$$\tilde{g}_{hi} = 1 + \frac{N_h(1-f_h)}{n_h - 1}(\mathbf{X} - \hat{\mathbf{X}}_{\text{NHT}})'\hat{\boldsymbol{\Sigma}}_{xx}^{-1}(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h), \qquad \ldots(4.8)$$

where

$$\boldsymbol{\Sigma}_{xx} = \sum_h \frac{N_h^2(1-f_h)}{n_h(n_h-1)} \sum_i (\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)',$$

$f_h = n_h/N_h$ and $\bar{\mathbf{x}}_h = \sum_i \mathbf{x}_{hi}/n_h$. The optimal estimator is also a calibration estimator in the sense of $\sum_s \tilde{w}_{hi}\mathbf{x}_{hi} = \mathbf{X}$, and only totals $\mathbf{X}$ are needed to implement $\hat{Y}_{\text{opt}}$, as in the case of $\hat{Y}_{\text{gr}}$. A drawback of $\hat{Y}_{\text{opt}}$ is that $\hat{\mathbf{B}}_{\text{opt}}$ can become unstable if the degrees of freedom associated with $\hat{\mathbf{\Sigma}}_{xx}$ is small. For example, in the case of stratified multistage sampling, the degrees of freedom, ν, is usually taken as the total number of sampled primary units (clusters) minus the number of strata so that $\hat{\mathbf{\Sigma}}_{xx}^{-1}$ becomes unstable if ν is not large relative to p, the number of anxiliary variables. On the other hand, $\hat{\mathbf{T}}^{-1}$ in the GREG weights g_i, given by (4.6), remains stable because the total number of elements n is much larger than p. The GREG estimator, however, does not lead to conditionally valid inferences except for simple random sampling in which case $\hat{Y}_{\text{opt}} = \hat{Y}_{\text{gr}}$; that is, the conditional bias ratio of $\hat{Y}_{\text{gr}}$ does not go to zero as $n \to \infty$.

Rao (1997) studied the conditional relative bias of $\hat{Y}_{\text{opt}}$, $\hat{Y}_{\text{NHT}}$ and the combined ratio estimator $\hat{Y}_r$ and its effect on conditional coverage rates of confidence intervals, using stratified simple random sampling with $L = 2$ strata and assuming that only the total X is known. In this case, the model-assisted approach would often use the ratio model $y_{hj} = \beta x_{hj} + e_{hj}$ with $V_m(e_{hj}) = \sigma^2 x_{hj}$ and a common slope $\beta_h = \beta$ across strata. This leads to $\hat{Y}_r$, a special case of $\hat{Y}_{gr}$. We cannot use a separate ratio estimator, obtained from a ratio model with different strata slopes, because the individual strata totals X_h are unknown. Note that the estimators $\hat{Y}_{\text{NHT}}$, $\hat{Y}_r$ and $\hat{Y}_{\text{opt}}$ are all inferentially valid in the unconditional design-based framework, that is, the unconditional bias ratio tends to zero as $n \to \infty$, even though $\hat{Y}_r$ and $\hat{Y}_{\text{opt}}$ are preferred over $\hat{Y}_{\text{NHT}}$ on efficiency grounds.

Assuming that the true model is a ratio model with different strata slopes $\beta_1 = 3$ and $\beta_2 = 1$, 10,000 stratified random samples from the true model, with strata sample sizes $n_1 = n_2 = 100$ and strata weights $W_1 = 0.2$, $W_2 = 0.8$, were generated. The simulated samples were then ordered according to their $\hat{X}_{\text{NHT}}$-values and divided into 10 groups, each with 1000 samples. This set-up mimics conditioning on $\hat{X}_{\text{NHT}}$ because samples within a group have similar $\hat{X}_{\text{NHT}}$-values.

The conditional values of bias ratio (BR), coverage rate (C) and lower (L) and upper (U) error rates of 0.90 level normal theory confidence intervals were calculated for each group, using the jackknife variance estimator for $\hat{Y}_{\text{opt}}$ and $\hat{Y}_r$, and the usual variance estimator for $\hat{Y}_{\text{NHT}}$. The results from the simulation study indicated that $\hat{Y}_{\text{NHT}}$ performs very poorly with conditional BR ranging from -133% to 152% and conditional coverage rates (C) as low as 60% for group 1, 70% for group 10 and significantly larger than the nominal rate of 90% for groups 3 to 8 (95.2% to 99.6%). Moreover, the conditional error rates L and U exhibited a clear trend across groups with L ranging from 40% to 0% and U from 0% to 31% compared to nominal rate of 5% in each tail. The combined ratio estimator $\hat{Y}_r$ exhibited significant positive or negative conditional BR for the extreme groups (47% for group 1 and -39% for group 10). It performed generally well in terms of conditional coverage rate, but the conditional error rates exhibited a trend across groups with L ranging from 1.7% to 10% and U from 10.6% to 2.1%. On the other hand, the optimal estimator $\hat{Y}_{\text{opt}}$ led to small conditional bias ratio ($< 10\%$), performed well in terms of conditional coverage rate and exhibited no visible trends in conditional error rates with both L and U closer to the nominal 5%. The optimal estimator is clearly preferable to the ratio estimator in controlling both error rates which is desirable and also necessary if one wishes to employ one-sided confidence intervals.

We have also conducted a similar simulation study using the Hansen *et al.* (1983) model misspecification and their stratified random design with near-optimal allocation and total sample size n. The population x-values are known here and 10 strata were formed on the basis of x-values such that the x-totals are approximately the same for each stratum. Note that such an x-stratification is not possible for the preceding simulation study because only the total X is assumed to be known. Following Hansen *et al.* (1983), we considered the ratio model-based estimator, $\hat{Y}_m$ say, and two-standard design based estimators, $\hat{Y}_{\text{NHT}}$ and the combined ratio estimator $\hat{Y}_r$, and associated variance estimators. The unconditional coverage rate of model-based confidence intervals is about 70% which is substantially less than the nominal 95%, whereas the design-based estimators performed well with unconditional coverage rates of 94.8% and 94.4% for $\hat{Y}_{\text{NHT}}$ and $\hat{Y}_r$. These values are very close to these reported by Hansen *et al.* (1983).

Table 1. Conditional Coverge Rate (%)

	Group									
	1	2	3	4	5	6	7	8	9	10
$\hat{Y}_m$	74.0	73.0	73.0	75.0	74.0	75.0	76.0	74.0	73.0	74.0
$\hat{Y}_{\text{NHT}}$	92.0	93.6	93.2	94.7	94.7	96.1	95.7	94.4	94.3	94.3
$\hat{Y}_r$	94.4	92.7	93.7	94.0	93.8	95.2	95.1	95.0	93.7	94.8

Table 1 gives conditional coverage rates for the 10 groups formed according to $\hat{X}_{\text{NHT}}$- values of the simulated samples. It is evident from Table 1 that the model-based intervals perform poorly in terms of conditional coverage rate: 73% to 75% compared to nominal 95%. On the other hand, both design-based intervals performed well in terms of conditional coverage rate: 92.0% to 96.1% for $\hat{Y}_{\text{NHT}}$ and 92.7% to 95.2% for $\hat{Y}_r$. Because of efficient x-stratification, most of the simulated samples have $\hat{X}_{\text{NHT}}$-values close to the total X, that is, the design leads to mostly balanced samples unlike in the previous simulation study where x-stratification is not possible. For nearly balanced samples, $\hat{Y}_{\text{NHT}}$ and $\hat{Y}_r$ give similar conditional results whereas $\hat{Y}_{\text{NHT}}$ performed poorly relative to $\hat{Y}_r$ in the preceding simulation study. In any case, this simulation study demonstrates that the conclusions of Hansen *et al.* (1983) are also valid under a conditional inference framework, provided the design leads to nearly balanced samples.

Empirical likelihood methods, recently introduced by Owen (1988), in the context of independent, identically distributed random variables, provide a systematic nonparametric approach to utilizing auxiliary information in making inference on the parameters of interest. Hartley and Rao (1968) gave the original idea of empirical likelihood in the context of sample surveys, using their "scale-load" approach. Under simple random sampling, they obtained the maximum empirical likelihood estimator of the total Y when only X is known, and showed that it is asymptotically equivalent to the customary regression estimator. Chen and Qin (1993) extended these results to cover the distribution function of y-values, $F_y(t) = \sum I(y_j \leq t)/N$, and associated quantiles M_q, $0 < q < 1$, where $I(y_j \leq t)$ is the indicator variable. Zhong and Rao (1996) studied the empirical likelihood under stratified random sampling when only X is known. They showed that the empirical maximum likelihood estimator is asymptotically equivalent to the optimal regression estimator, and also obtained the empirical maximum likelihood estimator, $\hat{F}_y(t)$, of $F_y(t)$ and associated quantiles. The optimal regression estimator of $F_y(t)$ is asymptotically equivalent to $\hat{F}_y(t)$ but it may not be monotone unlike $\hat{F}_y(t)$; monotonicity of the estimated distribution function is necessary for calculating estimators of quantiles. The previous results suggest that empirical likelihood methods in the sample survey context should lead to conditionally valid inferences, but much work remains to be done including extensions to complex sampling designs.

5. Resampling Methods

In recent years, considerable attention has been given to the estimation of complex parameters, θ, such as ratio and regression coefficients which can be expressed as smooth functions of totals of suitably defined variables, and to functions of $F_y(t)$ such as the median $M_{\frac{1}{2}}$, low-income proportion $F_y(M_{\frac{1}{2}})$ and measures of income inequality (Lorenz curve ordinate and the Gini index). Analyses of survey data, such as multi-way tables of estimated counts, linear and logistic regression, and multivariate analysis, also received a lot of attention. Standard software packages, such as SAS, use the survey weights properly for estimation but at the analysis stage the "design effects" are ignored; see Brogan (1998) for a discussion of pitfalls of using standard packages for survey data. Packages tailor-made for analysis of survey data take proper account of design effects, for example SUDAAN and PC CARP based on the Taylor linearization method and WESVAR based on resampling methods (jackknife, balanced repeated replication(BRR)). Consequences of ignoring design effects include underestimation of standard errors, inflated test levels and erroneous model diagnostics.

An advantage of resampling methods is that a single standard error formula is used for all statistics, unlike the linearization method which requires the derivation of a separate formula for each statistic. Moreover, the linearization method can become cumbersome in handling post-stratification and nonresponse adjustments, whereas it is relatively straightforward with resampling methods. For example, PC CARP and SUDAAN currently cannot handle complex analyses such as logistic regression with post-stratified weights, unlike WESVAR. Several statistical agencies in North America and Europe have adopted the jackknife, BRR or some other resampling methods for variance estimation and analysis of survey data. Resampling methods are applicable to stratified random sampling, commonly used in establishment surveys, as well as to stratified multistage sampling commonly used in large socio-economic surveys. In the latter case, resampling methods are valid provided the sample clusters are selected with replacement or the first-stage sampling fraction is negligible. In this section we focus on the preceding stratified multistage design with large number of strata, L, and relatively few primary sampling units (clusters), $m_h (\geq 2)$, sampled within each stratum $h(= 1, \ldots, L)$. We assume that subsampling within sampled clusters i $(= 1, \ldots, m_h)$ is performed to ensure unbiased estimation of cluster totals. The basic design weights w_{hik} attached to the sample elements hik are adjusted for poststratification (and unit nonresponse) to get the adjusted weights w^*_{hik}; see (4.5).

Many parameters of interest, $\boldsymbol{\theta}$, can be formulated as the solutions to the "census" equations

$$\mathbf{S}(\boldsymbol{\theta}) = \sum \mathbf{u}(\mathbf{y}_{hik}, \boldsymbol{\theta}) = \mathbf{0}, \qquad \ldots (5.1)$$

where the summation is over the population elements hik (Binder, 1983). For example, (i) $u(y, \theta) = y - \theta$ give the population mean $\bar{Y}$, (ii) $u(y, \theta)I(y < \theta) - \frac{1}{2}$ give the population median and (iii) $\mathbf{u}(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}(\mathbf{y} - \mathbf{x}^T\boldsymbol{\theta})$ give the population linear regression coefficient. The ℓ-th component of $\mathbf{u}(y, \boldsymbol{\theta})$ for a generalized linear model with $E(y) = \mu = \mu(\mathbf{x}, \boldsymbol{\theta})$ and a "working" variance $V(y) = V_0 = V_0(\mu)$ is given by

$$u_\ell(y, \theta) = \frac{\partial \mu}{\partial \theta_\ell} \frac{(y - \mu)}{V_0}. \qquad \ldots (5.2)$$

Linear regression is a special case of (5.2) with $\mu = \mathbf{x}'\boldsymbol{\theta}$ and $V_0 = \sigma^2$, a constant not depending on μ. Logistic regression is obtained by taking $\log\{\mu/(1 - \mu)\} = \mathbf{x}'\boldsymbol{\theta}$ and $V_0 = \mu(1 - \mu)$.

Since $\mathbf{S}(\boldsymbol{\theta})$ is a population total, the GREG estimator of $\mathbf{S}(\boldsymbol{\theta})$ is

$$\hat{\mathbf{S}}_r(\boldsymbol{\theta}) = \sum_{hik \in s} w^*_{hik} \mathbf{u}(\mathbf{y}_{hik}, \boldsymbol{\theta}) \qquad \ldots (5.3)$$

and the solution of $\hat{\mathbf{S}}_r(\boldsymbol{\theta}) = \mathbf{0}$ gives the estimator $\hat{\boldsymbol{\theta}}_r$ of $\boldsymbol{\theta}$.

For smooth statistics $\hat{\boldsymbol{\theta}}_r$, the delete-one-cluster jackknife method readily provides a variance estimator of $\hat{\boldsymbol{\theta}}_r$ as well as tests of hypotheses on $\boldsymbol{\theta}$. To implement this method, we need to recalculate the adjusted weights w^*_{hik} each time a sample cluster (ℓj) is deleted. This is done in two steps: (1) change the basic weights

w_{hik} to the jackknife weights $w_{hik(\ell j)} = w_{hik}b_{\ell j}$, where $b_{\ell j} = 0$ if $(hi) = (\ell j)$; $= m_\ell/(m_\ell - 1)$ if $h = \ell$ and $i \neq j$; $= 1$ if $h \neq \ell$. (2) Replace w_{hik} by $w_{hik(\ell j)}$ in the poststratification adjustment process to get the adjusted jackknife weights $w^*_{hik(\ell j)}$. Now replace w^*_{hik} in (5.3) by $w^*_{hik(gj)}$ to get the estimator $\hat{\boldsymbol\theta}_{r(\ell j)}$ when the sample cluster (ℓj) is deleted. A jackknife variance estimator of $\hat{\boldsymbol\theta}_r$ is then given by

$$\mathrm{var}_J(\hat{\boldsymbol\theta}_r) = \hat{\boldsymbol\Sigma}_J = \sum_{\ell=1}^{L} \frac{m_\ell - 1}{m_\ell} \sum_{j=1}^{n_\ell} (\hat{\boldsymbol\theta}_{r(\ell j)} - \hat{\boldsymbol\theta}_r)(\hat{\boldsymbol\theta}_{r(\ell j)} - \hat{\boldsymbol\theta}_r)'. \qquad \ldots(5.4)$$

If the solution of estimating equations (5.2) requires iterations, as in the case of logistic regression for example, then the jackknife implementation can be simplified by doing only a single iteration with $\hat{\boldsymbol\theta}_r$ as the starting value to get $\bar{\boldsymbol\theta}_{r(\ell j)}$, say, and then replacing $\hat{\boldsymbol\theta}_{r(\ell j)}$ in (5.4) by $\bar{\boldsymbol\theta}_{r(\ell j)}$.

Wald tests of hypotheses on $\boldsymbol\theta$ can be readily obtained using $\hat{\boldsymbol\theta}_r$ and $\mathrm{var}_J(\hat{\boldsymbol\theta}_r)$. A drawback of Wald tests is that the full model should be fitted to get $\hat{\boldsymbol\theta}_r$ which can be computer-intensive when the dimension of $\boldsymbol\theta$ is large. Also, Wald tests are not invariant to reparametrizations. To overcome these drawbacks, Rao, Scott and Skinner (1998) proposed quasi-score tests using the jackknife. Suppose $\boldsymbol\theta = (\boldsymbol\theta_1', \boldsymbol\theta_2')'$ and the hypothesis of interest is of the form $H_0 : \boldsymbol\theta_2 = \boldsymbol\theta_{20}$ where $\boldsymbol\theta_2$ is $q \times 1$ and $\boldsymbol\theta_{20}$ is specified. Let $\hat{\mathbf{S}}_r(\boldsymbol\theta) = \hat{\mathbf{S}}_r = (\hat{\mathbf{S}}'_{1r}, \hat{\mathbf{S}}'_{2r})'$ be the partition of $\hat{\mathbf{S}}_r$ corresponding to the partition of $\boldsymbol\theta$, and $\tilde{\boldsymbol\theta}_r(\tilde{\boldsymbol\theta}'_{1r}, \boldsymbol\theta'_{20})'$ be the solution of $\hat{\mathbf{S}}_{1r}(\boldsymbol\theta_1, \boldsymbol\theta_{20}) = \mathbf{0}$ which is much simpler to obtain than $\hat{\boldsymbol\theta}_r$ if the dimension of $\boldsymbol\theta_2$ is large. For example, $\boldsymbol\theta_2$ might contain several interaction effects and we are interested in testing whether a simpler model with zero interactions might fit the data. To get a jackknife quasi-score test, we calculate the score vector $\tilde{\mathbf{S}}_{2r} = \hat{\mathbf{S}}_{2r}(\tilde{\boldsymbol\theta}_r)$ and the jackknife score vectors $\tilde{\mathbf{S}}_{2r(\ell j)}$ which are obtained in the same manner as $\tilde{\mathbf{S}}_{2r}$, but using $w^*_{hik(\ell j)}$ in place of w^*_{hik}. The jackknife variance estimator of $\tilde{\mathbf{S}}_{2r}$, denoted by $\tilde{\boldsymbol\Sigma}_{2SJ}$, is simply obtained from (5.4) by changing $\hat{\boldsymbol\theta}_{r(\ell j)}$ and $\hat{\boldsymbol\theta}_r$ to $\tilde{\mathbf{S}}_{2r(\ell j)}$ and $\tilde{\mathbf{S}}_{2r}$. The jackknife quasi-score test is now given by

$$Q_J = \tilde{\mathbf{S}}'_{2r} \tilde{\boldsymbol\Sigma}^{-1}_{2SJ} \tilde{\mathbf{S}}_{2r} \qquad \ldots(5.5)$$

which is treated as a chi-squared variable with q degrees of freedom under H_0.

If the effective degrees of freedom, $\Sigma m_h - L$, is not large relative to q, the dimension of $\boldsymbol\theta_2$, then Q_J can become unstable. Wald tests also suffer from the instability problem. Alternative tests based on the Rao-Scott (1984) corrections to the naive tests that ignore the design feature may be used to ovecome the degrees of freedom problem. These corrections use "generalized design effects" which can be calculated using the jackknife. The Rao-Scott corrected tests are also useful for secondary analyses from published tables providing information on design effects; note that to implement Q_J we need microdata. Bonferroni procedures in the context of a Wald test (Korn and Graubard (1990)) or a quasi-score test (Rao, Scott and Skinner, 1996) can also be used. In the latter case, the elements of $\tilde{\mathbf{S}}_{2r}$ and associated standard errors (obtained from the diagonal elements of $\tilde{\boldsymbol{\Sigma}}_{2SJ}$) are used to construct q t-statistics before applying the Bonferroni procedure.

Hinkins, Oh and Scheurin (1997) proposed drawing subsamples from the original complex sample so as to yield at the end multiple simple random samples. It is not always possible to invert a original sample to get a simple random sample, but Hinkins *et al.* discussed several practically important special cases. This method looks promising for generating public-use microdata files that preserve confidentiality, and for analysing such files by standard methods, including model diagnostics. But much more work remains to be done in terms of feasibility and performance relative to the methods that take account of design features.

The delete-one-cluster jackknife may not perform well for nonsmooth statistics such as the median; for simple random sampling the delete-one-element jackknife is known to be inconsistent for the median. Balanced half-samples (BHS) method for the important special case of $m_h = 2$ sample clustes per stratum works for both smooth and nonsmooth statistics. The design weights w_{hik} are changed to $2w_{hik}$ or 0 according as the (hi)-th cluster is selected or not in the half-sample. Using these weights in place of the jackknife weights, we follow the preceding steps to get $\hat{\boldsymbol{\theta}}_r^{(a)}$ and $\tilde{\mathbf{S}}_{2r}^{(a)}$ for the a-th half sample $(a = 1, \ldots, A)$. The BHS variance estimator of $\hat{\boldsymbol{\theta}}_r$ is then given by

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{BHS}} = \frac{1}{A}\sum_{a=1}^{A}(\hat{\boldsymbol{\theta}}_r^{(a)} - \hat{\boldsymbol{\theta}}_r)(\hat{\boldsymbol{\theta}}_r^{(a)} - \hat{\boldsymbol{\theta}}_r)'. \qquad \ldots(5.6)$$

The BHS variance estimator of $\tilde{\mathbf{S}}_{2r}$ is obtained from (5.6) by changing $\hat{\boldsymbol{\theta}}_r^{(a)}$ and $\hat{\boldsymbol{\theta}}_r$ to $\tilde{\mathbf{S}}_{2r}^{(a)}$ and $\tilde{\mathbf{S}}_{2r}$.

A drawback of BHS is that some of the replicate estimators may become extreme because only a half sample is used, that is, the weights w_{hik} are sharply perturbed. The jackknife avoids this problem by dropping only one cluster at a time. Robert Fay of the US Census Bureau provided a solution to this problem by using a gentler perturbation of the weights: w_{hik} is changed to $1.5w_{hik}$ or $0.5w_{hik}$ according as the (hi)-th cluster is selected in the a-th half-sample. Thus, the full sample is used to construct the estimators $\hat{\boldsymbol{\theta}}_r^{(a)}\left(\frac{1}{2}\right)$ and $\tilde{\mathbf{S}}_{2r}^{(a)}\left(\frac{1}{2}\right)$ using Fay's weights. The variance estimator of $\hat{\boldsymbol{\theta}}_r$ is given by

$$\hat{\boldsymbol{\Sigma}}_{BHS\left(\frac{1}{2}\right)} = \frac{4}{A}\sum_{a=1}^{A}\left[\hat{\boldsymbol{\theta}}_r^{(a)}(1/2) - \hat{\boldsymbol{\theta}}_r^{(a)}\right]\left[\hat{\boldsymbol{\theta}}_r^{(a)}(1/2) - \hat{\boldsymbol{\theta}}_r^{(a)}\right]'. \qquad \ldots(5.7)$$

Similarly, the variance estimator of $\tilde{\mathbf{S}}_{2r}$ is obtained.

Unfortunately, it is difficult to construct balanced replicates for arbitrary m_h. Bootstrap offers a way out and leads to valid inferences for both smooth and nonsmooth statistics. A bootstrap sample is obtained by drawing a simple random sample of $m_h - 1$ clusters with replacement from the m_h sample clusters, independently for each h. We select a large number, B, of bootstrap samples independently which can be represented in terms of the bootstrap frequencies $m_{hi}(b)$ = number of times (hi)-th sample cluster is selected in the b-th bootstrap sample $(b = 1, \ldots, B)$. The bootstrap design weights are simply given by $w_{hik}(b) = w_{hik}[m_h/(m_h - 1)]m_{hi}(b)$. Using these weights we proceed as before to get the bootstrap estimators $\hat{\boldsymbol{\theta}}_r(b)$ and $\tilde{\mathbf{S}}_{2r(b)}$. A bootstrap variance estimator of $\hat{\boldsymbol{\theta}}_r$ is given by

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{BOOT}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}_r(b) - \hat{\boldsymbol{\theta}}_r)(\hat{\boldsymbol{\theta}}_r(b) - \hat{\boldsymbol{\theta}}_r)'. \qquad \ldots(5.8)$$

Similarly, the variance estimator of $\tilde{\mathbf{S}}_{2r}$ is obtained.

Various extensions of the preceding results have been studied in recent years. Kott and Stukel (1998) extended the jackknife to two-phase sampling which uses stratified multistage sampling in the first phase, then restratifies the first-phase sample elements and selects simple random samples from each second-phase stratum. Lohr and Rao (1997) extended the jackknife to handle dual-frame surveys where samples are drawn independently from two overlapping frames that together cover the population. Rao and Shao (1992) and Rao (1996) studied jackknife variance estimation for imputed survey data, while Shao and Sitter (1996) used the bootstrap and Rao and Shao (1999) used Fay's BHS for imputed data. The reader is referred to Shao and Tu (1996) for an excellent account of resampling methods and their theoretical properties. Rust and Rao (1996) summarized the properties of resampling methods for surveys and presented several examples from the literature of the use of resampling to analyse data from large complex surveys.

We have focussed on the "census" parameters $\boldsymbol{\theta}$, but often we are interested in making inferences on the parameters of the superpopulation model that is assumed to generate the finite population. If the finite population is assumed to be a random sample from the hypothetical superpopulation and the overall sampling fraction n/N is negligible, then the finite population inferences remain valid for the superpopulation. But this assumption is somewhat unrealistic as it ignores the finite population structure such as clustering. More careful modelling is needed and the preceding results have to be modified (Korn and Graubard, 1998). But validating such models can be difficult in practice.

6. Small Area Estimation

As explained in Section 2, preventive measures should be taken at the design stage, whenever possible, to ensure adequate precision for subgroups (or domains) and smaller geographical areas like the UI region. But even after taking such measures sample sizes may not be large enough for direct estimators to provide adequate precision for small areas (domains). Sometimes, the survey is deliberately designed to oversample specific areas at the expense of small samples or even no samples in other areas. For example, in the US Third National Health and Nutrition Examination Survey, states with larger black and hispanic populations were oversampled at the expense of small samples or no samples in the other states (e.g., Nebraska). Yet reliable estimates are needed for all the states and areas within states.

Small sample sizes in small areas (even large areas as in the preceding example) make it necessary to "borrow strength" from related areas to find indirect estimators that increase the effective sample size and thus increase the precision. It is now generally accepted that one should use indirect estimators based on explicit models that relate the small areas through supplementary data such as administrative records and recent census counts. An advantage of the model approach is that it permits validation of models from the sample data.

Small area models may be broadly classified into two types. In the first type, area specific auxiliary data, $\mathbf{x}_i$, are available for the areas $i = 1, \ldots, m$. The population small area mean $\bar{Y}_i$ (or total Y_i) or some suitable function $\theta_i = g(\bar{Y}_i)$ is assumed to be related to $\mathbf{x}_i$ through a linear model with random small area effects v_i:

$$\theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \qquad \ldots (6.1)$$

where $\boldsymbol{\beta}$ is the vector of regression parameters and the v_i's are uncorrelated with mean zero and variance σ_v^2; normality of the v_i is also often assumed. In the second type of models, unit y-values, y_{ij}, are assumed to be related to auxiliary values $\mathbf{x}_{ij}$ through a nested error regression model

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m \qquad \ldots (6.2)$$

where the $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ are independent of $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ and N_i is the number of population units in the i-th area. The model (6.2) is appropriate for continuous valued variables y. To handle count or categorical (e.g., binary) variables y, generalized linear models with random small area effects are often used. Ghosh *et al.* (1998) assumed models of the following form: (i) Given θ_{ij}'s, the y_{ij}'s are independent and the probability density of y_{ij} belongs to the exponential family with canonical parameter θ_{ij}; (ii) $h(\theta_{ij}) = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i$ where $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$, and $\mathbf{x}_{ij}$ does not include 1 and $h(\cdot)$ is a strictly increasing function. The model (6.2) is a special case with $h(a) = a$; the logistic function $h(a) = \log[a/(1-a)]$ is often used for binary y. The parameters of interest are the small area means $\bar{Y}_i$ or totals Y_i.

In the case of type 1 models, we assume that direct survey estimators $\widehat{\bar{Y}}_i$ are available whenever the sample sizes $n_i \geq 1$. It is customary to assume that

$$\hat{\theta}_i = \theta_i + e_i, \qquad \ldots (6.3)$$

where $\hat{\theta}_i = g(\hat{\bar{Y}}_i)$ and the sampling errors e_i are independent $N(0, \psi_i)$ with known ψ_i. Combining this sampling model with the linking model (6.1), we get

$$\hat{\theta}_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i + e_i \qquad \ldots (6.4)$$

which is a special case of the standard mixed linear model. Note that (6.4) involves both design-based random variables e_i and model based random variables v_i. If the sampling variances ψ_i are unknown, smoothing of estimated variances $\hat{\psi}_i$ is often done to get stable estimates ψ_i^* which are treated as the true ψ_i. Recently, models of the form (6.4) with $\theta_i = \log Y_i$ have been used to produce model-based county estimates of poor school-age children in the United States (National Research Council, 1998). Using these estimates, the US Department of Education allocated over \$7 billion of federal funds to counties, and the states distributed these funds among school districts in each county. The difficulty with unknown ψ_i was handled by using a model of the form (6.4) for the census year 1990, for which reliable estimates $\hat{\psi}_{ic}$ of sampling variances, ψ_{ic}, are available, and assuming the census small area effects v_{ic} follow the same distribution as v_i. Under this assumption, an estimate of σ_v^2 was obtained from the census data assuming $\hat{\psi}_{ic} = \psi_{ic}$ and then used in (6.4), assuming $\psi_i = \sigma_e^2/n_i$, to get an estimate of σ_e^2. The resulting estimate $\tilde{\psi}_i = \tilde{\sigma}_e^2/n_i$, was treated as the true ψ_i in developing small area estimates $\tilde{\theta}_i$ of θ_i. The small area totals Y_i were then estimated as $\tilde{Y}_i = \exp(\tilde{\theta}_i)$.

Noting that (6.4) is a special case of the standard mixed linear model, we can appeal to general results on best linear unbiased prediction (BLUP) estimation of a linear combination of fixed and random effects. The BLUP estimator of θ_i is given by

$$\tilde{\theta}_i(\sigma_v^2) = \gamma_i \hat{\theta}_i + (1 - \gamma_i)\mathbf{x}_i' \tilde{\boldsymbol{\beta}} (\sigma_v^2), \qquad \ldots (6.5)$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \psi_i)$ and $\tilde{\boldsymbol{\beta}} (\sigma_v^2)$ is the weighted least squares estimator of β with weights $(\sigma_v^2 + \psi_i)^{-1}$. It follows from (6.5) that the BLUP estimator is a weighted combination of the direct estimator $\hat{\theta}_i$ and the regression "synthetic" estimator $\mathbf{x}_i' \tilde{\boldsymbol{\beta}}$. It gives more weight to $\hat{\beta}_i$ when ψ_i is small and moves towards $\mathbf{x}_i' \tilde{\boldsymbol{\beta}}$ as σ_v^2 decreases. Note that (6.5) does not depend on the normality of v_i and e_i. Replacing σ_v^2 by a suitable estimator $\hat{\sigma}_v^2$ we obtain a two-stage estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$ which is called empirical BLUP or EBLUP estimator in analogy with the empirical Bayes (EB) estimator. One could use either the method of fitting constants (not requiring normality) or the restricted maximum likelihood (REML) method to estimate σ_v^2.

The EB approach applied to model (6.4) gives an estimator $\tilde{\theta}_i^{\mathrm{EB}}$ identical to the EBLUP estimator $\tilde{\theta}_i$, but it is more generally applicable. The conditional distribution of θ_i given the model parameters and the data $\hat{\theta}_i$ is first obtained.

The model parameters are estimated from the marginal distribution of $\hat{\theta}_i$'s, and inferences are then based on the estimated conditional distribution of θ_i. EB is essentially frequentist because it uses only the model (6.4) which can be validated from the data; no prior distribution on the model parameters, $\boldsymbol{\beta}$ and σ_v^2, is assumed. Linear EB approach is similar to EBLUP in the sense that it is also distribution-free; under normality it agrees with EB.

In the case of nonlinear functions $\theta_i = g(\bar{Y}_i)$, the assumption $E(e_i|\theta_i) = 0$ in the sampling model (6.3) may not be valid if the sample size n_i is unduly small, even if the direct estimator $\hat{\bar{Y}}_i$ is design-unbiased, i.e., $E(\hat{\bar{Y}}_i|\bar{Y}_i) = \bar{Y}_i$. A more realistic sampling model is given by

$$\hat{\bar{Y}}_i = \bar{Y}_i + e_i^* \qquad \ldots (6.6)$$

with $E(e_i^*|\bar{Y}_i) = 0$. In this case, we cannot combine (6.6) with the linking model (6.1) to produce a linear mixed model in order to be able to use the EBLUP approach. The EB approach also runs into difficulties because the conditional distribution of $\bar{Y}_i$ given the model parameters and $\hat{\bar{Y}}_i$ no longer has a closed-form. Note that the conditional distribution of θ_i is normal under the sampling model (6.3).

A measure of variability associated with the estimator $\tilde{\theta}_i$ is given by the mean squared error, MSE $(\tilde{\theta}_i) = E(\tilde{\theta}_i - \theta_i)^2$, but no closed form exists except in some special cases. Under normality of the errors, an accurate approximation to MSE, for large m, can be obtained as

$$\text{MSE}(\tilde{\theta}_i) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) \qquad \ldots (6.7)$$

where

$$\begin{aligned} g_{1i}(\sigma_v^2) &= \gamma_i \psi_i, \\ g_{2i}(\sigma_v^2) &= (1-\gamma_i)^2 \mathbf{x}_i' [\textstyle\sum_i \mathbf{x}_i \mathbf{x}_i'/(\sigma_v^2 + \psi_i)]^{-1} \mathbf{x}_i, \end{aligned} \qquad \ldots (6.8)$$

$$\begin{aligned} g_{3i}(\sigma_v^2) &= [\psi_i^2/(\sigma_v^2+\psi_i)^4] E(\hat{\theta}_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \bar{V}(\hat{\sigma}_v^2) \qquad \ldots (6.9) \\ &= [\psi_i^2/(\sigma_v^2+\psi_i)^3] \bar{V}(\hat{\sigma}_v^2) \qquad \ldots (6.10) \end{aligned}$$

and $\bar{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$. The leading term $g_{1i}(\sigma_v^2)$ is of order $O(1)$ whereas $g_{2i}(\sigma_v^2)$, due to estimating β, and $g_{3i}(\sigma_v^2)$, due to estimating σ_v^2, are both of order $O(m^{-1})$, for large m. Note that the leading term shows that MSE $(\tilde{\theta}_i)$ can be substantially smaller than MSE $(\hat{\theta}_i) = \psi_i$ under the model (6.4) when γ_i is small or the model variance σ_v^2 is small relative to the sampling variance ψ_i. The success of small area estimation, therefore, largely depends on getting good auxiliary information $\{\mathbf{x}_i\}$ that leads to a small model variance σ_v^2 relative to ψ_i. Of course, one should also make a thorough evaluation of the assumed model.

An estimator of MSE $(\tilde{\theta}_i)$, correct to the same order approximation as (6.7), is given by

$$mse(\tilde{\theta}_i) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \qquad \ldots (6.11)$$

i.e., the bias of (6.11) is of lower order than m^{-1} for large m. The approximation (6.11) is valid for both the method of fitting constants estimator and the REML estimator, but not for the maximum likelihood (ML) estimator of σ_v^2 (Prasad and Rao, 1990; Datta and Lahiri, 1997). Using the fitting-of-constants estimator, Lahiri and Rao (1995) showed that (6.11) is robust to nonnormality of the small area effects v_i in the sense that the preceding approximate unbiasedness remains valid. Note that the normality of the sampling errors is still assumed but it is less restrictive due to the central limit theorem's effect on the direct estimators $\hat{\theta}_i$.

A criticism of (6.11) is that it is not area-specific in the sense that it does not depend on $\hat{\theta}_i$ although $\mathbf{x}_i$ is involved. But it is easy to find other choices using the form (6.9) for $g_{3i}(\sigma_v^2)$. For example, we can use

$$\begin{aligned} mse_1(\tilde{\theta}_i) &= g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2) \\ &+ [\psi_i^2/(\hat{\sigma}_v^2+\psi_i)^4](\hat{\theta}_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}})^2 h_i(\hat{\sigma}_v^2), \end{aligned} \qquad \ldots (6.12)$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_v^2)$ and $h_i(\sigma_v^2) = \bar{V}(\hat{\sigma}_v^2) = 2m^{-2}\sum_i(\sigma_v^2+\psi_i)^2$ for the fitting-of-constants estimator $\hat{\sigma}_v^2$ (Rao, 1998).

The customary EB approach uses the estimated conditional distribution of θ_i for inference, so that the variability of $\tilde{\theta}_i^{\text{EB}} = \tilde{\theta}_i$ is given by $g_{1i}(\hat{\sigma}_v^2)$ which leads to severe underestimation of the true variability as measured by MSE. Laird and Louis (1997) proposed a parametric bootstrap method to account for the variability in $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$. By deriving an analytical approximation to the bootstrap measure of variability, Butar and Lahiri (1997), however, showed that it is not second-order correct, i.e., its bias involves terms of order m^{-1} unlike the bias of (6.12). By correcting this bias, they obtained an estimator which is identical to the area-specific MSE estimator (6.12).

Hierarchical Bayes (HB) approach has also been used for small area estimation. A prior distribution on the model-parameters is specified, and $\bar{Y}_i$ is then estimated by its posterior mean given the sample data, and its variability is measured by its posterior variance. For the simple area-level model given by (6.1) and (6.3) the posterior mean and posterior variance of θ_i are obtained in two stages. In the first stage, we obtain $E(\theta_i|\hat{\boldsymbol{\theta}}, \sigma_v^2)$ and $V(\theta_i|\hat{\boldsymbol{\theta}}, \sigma_v^2)$ for a given σ_v^2, assuming an improper prior, $f(\boldsymbol{\beta}) \propto$ const., on $\boldsymbol{\beta}$ to reflect absence of prior information on $\boldsymbol{\beta}$, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$. The posterior mean, given σ_v^2, is identical to the BLUP estimator $\tilde{\theta}_i(\sigma_v^2)$ and the posterior variance given σ_v^2 is equal to $g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)$. At the second stage, we take account of the uncertainty about σ_v^2 by first calculating its posterior distribution $f(\sigma_v^2|\hat{\boldsymbol{\theta}})$, assuming a prior distribution on σ_v^2 and prior independence of $\boldsymbol{\beta}$ and σ_v^2. The posterior mean and variance are then obtained as

$$E(\theta_i|\hat{\boldsymbol{\theta}}) = E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}[\tilde{\theta}_i(\sigma_v^2)] \qquad \ldots (6.13)$$

and

$$V(\theta_i|\hat{\boldsymbol{\theta}}) = E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}[g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)] + V_{\sigma_v^2|\hat{\boldsymbol{\theta}}}[\tilde{\theta}_i(\sigma_v^2)], \qquad \ldots (6.14)$$

where $E_{\sigma_v^2|\hat{\boldsymbol{\theta}}}$ and $V_{\sigma_v^2|\hat{\boldsymbol{\theta}}}$ denote the expectation and variance with respect to $f(\sigma_v^2|\hat{\boldsymbol{\theta}})$. No closed form expressions for (6.13) and (6.14) exist, but they can be evaluated numerically using only one-dimensional integration. If the assumed prior $f(\sigma_v^2)$ is proper and informative, the HB approach is straightforward and no difficulties are encountered. On the other hand, an improper prior $f(\sigma_v^2)$ could lead to an improper posterior (Hobert and Casella, 1996). In the latter case, we cannot avoid the difficulty by choosing a diffuse proper prior because we will be simply approximating an improper posterior by a proper posterior.

Unit level models of the linear model form (6.2) and the generalized linear model form have also been studied in the literature. The sample data $\{y_{ij}, \mathbf{x}_{ij}, j = 1, \ldots, n_i; i = 1, \ldots, m\}$ is assumed to obey the model used for the population values, i.e., selection bias is absent. For the linear case, the EBLUP, EB and HB approaches have been used. We refer the reader to Ghosh and Rao (1994) for further details.

Jiang and Lahiri (1998) obtained the EB estimator and an approximation to its MSE correct to order m^{-1} for the case of binary y variables and logistic linear model with a random effect. They call the EB estimator as empircal best predictor (EBP) which may be more appropriate because no priors on model parameters are involved. Booth and Hobert (1988) argued that the conditional MSE of the EBP given the i-th area data $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ is more relevant as a measure of variability than the unconditional MSE because it is area-specific. Fuller (1989) earlier suggested a similar criterion in the context of linear models. But we have already shown that it is possible to obtain area-specific estimators of the unconditional MSE, at least for the linear model. Also, it is not clear how one should proceed with the conditioning when two or more small area estimators need to be aggregated to obtain an estimator for a larger area. How should one define the conditional MSE of the larger area estimator?

Ghosh *et al.* (1998) applied the HB approach to generalized linear models with random effects. Employing diffuse gamma proper priors on the inverse of the variance components, they provide conditions for the posterior to be proper. These conditions require the scale parameter of the gamma distribution, say d, to be strictly greater than zero. But as d approaches zero to reflect lack of prior information, we are essentially approximating an improper posterior by a proper posterior.

For complex models like the generalized linear model, implementation of HB involves high dimensional integration. Markov Chain Monte Carlo (MCMC) integration methods, such as the Gibbs sampler, seem to overcome the computational difficulties. Moreover, software called BUGS is also readily available for implementing MCMC. But extreme caution should be exercised in using these methods. For example, Hobert and Casella (1996) demonstrated that the Gibbs sampler could lead to seemingly reasonable inferences about a nonexistent posterior distribution. This happens when the posterior is improper and yet the Gibbs conditionals are all proper. Another difficulty with MCMC is that the diagnostic tools used to assess the convergence of the MCMC can fail to detect the sorts of convergence failure that they were designed to identify (Cowles and Carlin, 1996). All in all, HB methods can be very promising in solving even complex problems of small area estimation, provided the preceding difficulties can be taken care of in implementing the MCMC methods.

The model-based estimators under the unit level models do not use the survey weights unlike the area level models. As a result, the estimators are not design-consistent. Kott (1989) and Prasad and Rao (1998) obtained design-consistent estimators for a simple area level model by considering a reduced model based on the survey weights. Prasad and Rao (1998) also obtained a model-based estimator of MSE.

Various extensions of the basic models have been proposed to handle correlated sampling errors, spatial dependence, multivariate responses, time series and cross-sectional data, etc. Also modifications to ensure that the model-based estimators add up to a reliable direct estimator at a large area level as well as constrained estimators whose variation matches the variation of the small area means have been proposed. Ghosh and Rao (1994) provide some details of these developments up to 1992 or 1993.

7. Concluding Remarks

In this paper, I have presented some current trends in sample survey theory and methods, covering issues related to survey design, data collection and processing, and inference from survey data. I have not been able to cover many other important topics of current interest and I hope that the discussants of my paper will fill in this gap to some extent by covering some of the omitted topics.

References

Amahia, G.N., Chaubey, Y.P. and Rao, T.J. (1989). Efficiency of a new estimator in PPS sampling for multiple characteristics, *J. Statist. Planning. Inf.*, **21**, 75-84.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Intl. Statist. Rev.*, **51**, 279-292.

Booth, J.G. and Hobert, J.P. (1998). Standard errors of predictors in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 262-272.

Brewer, K.R.W. (1963). Ratio estimators and finite populations: some results deducible from the assumption of an underlying stochastic process. *Austr. J. Statist.*, **5**, 93-105.

Brogan, D.J. (1998). Pitfalls of using standard statistical software packages for sample survey data. In *Encyclopedia of Biostatistics* (P. Armitage and T. Colton ed.), Wiley.

Butar, F.B. and Lahiri, P. (1997). On the measures of uncertainty of empirical Bayes small-area estimators. *Tech. Rep.*, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

Casady, R.J. and Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, **19**, 103-113.

Casady, R.J. and Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodol.*, **19**, 183-192.

Chen, J. and Qin, J. (1993). Empirical likelihood estimator for finite populations and the effective use of auxiliary information. *Biometrika*, **80**, 107-116.

Cowles, M.K. and Carlin, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.*, **91**, 883-904.

Datta, G.S. and Lahiri, P. (1997). A unified measure of uncertainty of estimated best linear unbiased predictor in small-area estimation problems. *Tech. Rep.*, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

Eltinge, J.L. (1998). Accounting for non-Gaussian measurement error in complex survey estimators of distribution functions and quantiles. *Statist. Sinica*, **8**, (in press).

Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic editing and imputation. *J. Amer. Statist. Assoc.*, **71**, 17-35.

Fellegi, I.P. and Sunter, A.B. (1974). Balance between different sources of survey errors – some Canadian experiences. *Sankhyā*, **C36**, 119-142.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhya*, **C37**, 117-132.

–––– (1989). Prediction of true values for the measurement error model. Paper presented at the *Conf. Statist. Analysis of Measurement Error Models*, Humboldt State University.

– – –– (1995). Estimation in the presence of measurement error. *Intl. Statist. Rev.*, **63**, 121-147.

Ghosh, J.K. (1992). The Horvitz-Thompson estimate and Basu's circus revisited. *Proc. Indo-US Bayesian Workshops*, (P.K. Goel *et al.* eds.), pp. 225-228.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statist. Sci.*, **9**, 55-93.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *J. Amer. Statist. Assoc.*, **93**, 273-282.

Groves, R.M. (1996). Nonsampling error in surveys: the journey toward relevance in practice. *Proc. Statist. Can. Symp.* **96**, 7-14.

Groves, R.M. and Lepkowski, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proc. Sec. Survey Res. Meth.*, American Statistical Association, 340-345.

Hajek, J. (1971). Comments on a paper by Basu, D. In *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), Holt, Rienhart and Winston, Toronto.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *J. Amer. Statist. Assoc.*, **78**, 776-793.

Hartley, H.O. and Biemer, P.P. (1978). The estimation of nonsampling variances in current surveys. *Proc. Sec. Survey Res. Meth.*, American Statistical Association, 257-262.

Hartley, H.O. and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

– – –– (1978). The estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement.* (N.K. Namboodiri, ed.), Academic Press, New York, 35-43.

Hinkins, S., Oh, H.L. and Scheurin, F. (1997). Inverse sampling

design algorithms. *Survey Methodology.*, **23**, 11-21.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.*, **91**, 1461-1479.

Holt, D. and Smith, T.M.F. (1979). Post-stratification. *J. Roy. Statist. Soc.*, **A142**, 33-46.

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.

Jiang, J. and Lahiri, P. (1998). Empirical best prediction for small area inference with binary data. *Tech. Rep.*, Deparment of Mathematics and Statistics, University of Nebraska-Lincoln.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t statistics. *Amer. Statist.*, **44**, 270-276.

− − −− (1998). Variance estimation for superpopulation parameters. *Statist. Sinica*, **8**, 1131-1151.

Kott, P. (1989). Robust small domain estimation using random effects modelling. *Survey Methodology*, **15**, 1-12.

Kott, P. and Stukel, D. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, **23**, 81-89.

Lakatos, E. (1978). Undiminished residual effects designs and their applications. *Ph.D. thesis*, University of Maryland.

Lahiri, P. and Rao. J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **90**, 758-766.

Laird, N.M. and Louis, T.A. (1987). Empirical Bayes confidence interevals based on bootstrap samples. *J. Amer. Statist. Assoc.*, **82**, 739-750.

Lindley, D.V. (1996). Letter to Editor, *Amer. Statist.*, **50**, p. 197.

Linacre, S.J. and Trewin, D.J. (1993). Total survey design-application to a collection of the construction industry. *J. Off. Statist.*, **9**, 611-621.

Lohr, S.L. and Rao, J.N.K. (1998). Jackknife variance estimation in dual frame surveys. *Tech. Rep.*, Laboratory for Research in Statistics and Probability, Carleton University.

Mahalanobis, P.C. (1946). On large-scale sample surveys. *Phil. Transac. Roy. Soc.*, London, **B231**, 324-351.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large sample surveys. *Intl. Statist. Rev.*, **55**, 191-202.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Ind. Soc. Agric. Statist.*, **3**, 169-174.

National Research Council (1998). *Small Area Estimates of School-Age Children in Poverty.* Interim Report 2, National Research Council, Washington, D.C.

Nordbotten, S. (1995). Editing statistical records by neural networks. *J. Off. Statist.*, **11**, 391-411.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

− − −− (1998). On robust small area estimation using a simple random effects model. *Tech Rep.*, Laboratory for Research in Statistics and Probability, Carleton University.

Raghunathan, T.E. and Grizzle, J.E. (1995). A split questionnaire survey deisgn. *J. Amer. Statist. Assoc.*, **90**, 54-63.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, **A28**, 47-60.

− − −− (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Statist.*, **10**, 153-165.

− − −− (1996). On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.*, **91**, 499-506.

− − −− (1997). Developments in sample survey theory: an appraisal. *Can. J. Statist.*, **25**, 1-21.

− − −− (1998). EB and EBLUP in small area estimation. *Tech. Rep.*, Laboratory for Research in Statistics and Probability, Carleton University.

Rao, J.N.K. and Liu, J. (1992). On estimating distribution functions from sample survey data using supplementary information at the estimation stage. In *Nonparametric Statistics and Related Topics* (A.K.Md.E. Saleh ed.), Elsevier.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hotdeck imputation. *Biometrika*, **79**, 811-822.

– – – – (1998). Modified balanced repeated replication for complex survey data. *Biometrika*, **86** (in press).

Rao, J.N.K. and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proc. Sec. Survey Res. Meth.*, American Statistical Association, pp. 57-65.

Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statist. Sinica*, **8**, 1059-1070.

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *J. Amer. Statist. Assoc.*, **92**, 368-374.

Roddick, L.H. (1993). Data editing using neural networks. *Tech. Rep.*, Systems Development Division, Statistics Canada.

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

Rust, K. and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statist. Meth. Med. Res.*, **5**, 283-310.

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

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, **20**, 3-22.

Smith, T.M.F. (1995). Problems of resource allocation. *Proc. Stat. Can. Symp.*, **95**, Statistics Canada, 107-114.

Wretman, J. (1995). Split questionnaires. Paper presented at the *Conference on Methodological Issues in Official Statistics*, Stockholm.

Zhong, C.X. and Rao, J.N.K. (1996). Empirical likelihood inference under stratified random sampling using auxiliary information. *Proc. Sec. Survey Res. Meth.*, American Statistical Association, pp. 798-803.

J.N.K. Rao
School of Mathematics and Statistics
Carleton University
Ottawa, ON K1S 5B6
Canada
e-mail : jrao@math.carleton.ca

Discussion of the paper
Some Current Trends in Sample Survey Theory and Methods
By J.N.K. Rao

Discussant : Arijit Chaudhuri
Indian Statistical Institute

Within a very short space this review paper covers appreciably wide varieties of current topics on survey sampling. Professor Rao's well-known mastery of the subject is clearly reflected in this write-up. Encouraged by Professor Rao's suggestions, I would like to raise the following questions and issues.

First, in developing small domain statistics, on a client's demand, which approach is user-friendly in practice ? Should one go for an EBLUP with a frequentist measure of error or follow the EB or HB approaches with a fully Baysian interpretation? Does it make sense to start with a GREG predictor for a domain total and follow Fay and Herriot (JASA, 1979) to improve upon it by an EB estimator and compare the estimated measures of errors of both?

Are the design-cum-model based interpretations on the confidence intervals using the two valid? The Indian National Sample Survey Organisation (NSSO), like the US Current Population Survey (CPS) and the Canadian Labour Force Survey (LFS), has been producing national estimates of comparable parameters for a long time using a complex survey design.

Thus it seems appropriate to apply time series methods like Kalman filtering in deriving improved current estimates utilising the past estimates. Reaserch has been conducted on this topic (see, for example, Binder and Hidiroglou (1988), Meinhold and Singpurwalla (1983), among many others), but unfortunately, a detailed account of this topic is currently not available in a standarad text book in sampling.

Professor Rao mentioned about an interesting approach of ordering sensitive questions using cross-over and related designs. However, another way of handling sensitive questions is to use randomized response technique. For a comprehensive treatment of the subject, the readers are refered to Chaudhuri and Mukerjee(1988).

Network sampling and adaptive sampling are two promising methods described in two recent text books by Thompson(1992) and Thompson and Seber(1996). These methods are useful especially when serviceable frames are not available. It will be interesting to investigate how they complement 'Dual Frames'and telephone surveys.

References

Binder, D.A. and Hidiroglou, M.A. (1988). Sampling in time. In *Handbook of Statistics* **6**, ed, Krishnaiah, P.R. and Rao, C.R. Elsevier Science Publishers,B.V. 187-211.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response Theory and Techniques.* Marcel Dekker, Inc, New York.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places:an application of James-Stein procedures to Census data. *Jour. Amer. Stat. Assoc.* **74**, 269-277.

Meinhold, R.J. and Singpurwalla, N.D. (1983). Understanding the Kalman filter. *The Amer. Stat.* **37**, 123-127.

Thompson, S.K. (1992). *Sampling.* John Wiley and Sons,Inc,New York.

Thompson, S.K. and Seber, G.A.F (1996). *Adaptive Sampling.* John Wiley and Sons,Inc,New York.

Discussant : J.L. Eltinge
Texas A $ M University

Professor Rao offers a fascinating perspective on the current status of sample survey theory and methods, and raises many issues that warrant further discussion. In the interest of space, I will limit my comments to two areas: total survey design and inference from complex survey data.

1. Total Survey Design. Section 2 of the main paper highlights the importance of a total survey design approach to assessment of survey errors and survey costs. Implementation of this approach involves a mixture of technical and administrative considerations. The main paper and the general survey error literature (e.g., Andersen *et al.*, 1979; and Groves, 1989) lead to the following three points:

1.1 Links between mathematical statistics and empirical effects. Implementation of a total survey design approach involves a careful balance of results from mathematical statistics and empirical studies. In particular, mathematical statistics provides the essential underlying framework and offers important general insights into development and evaluation of proposed methods. However, the practical impact of these insights depends heavily on the combined empirical characteristics of the sample design, the underlying population, the data collection process, and the selected analytic methods. Some examples include main effects of, and interactions among, the following factors:

−− Classical sampling errors and components thereof (e.g., selection probabilities; and population-level variability among and within strata and clusters).

−− Nonsampling errors and efforts to reduce the effect of nonsampling errors (e.g., the magnitudes of measurement error components; the correlation of response probabilities with survey items Y; the magnitude of error reduction achieved through specific modifications in the questionnaire or in interviewer training; the number of replicate measurements taken; or the number of nonrespondent callbacks).

−− Specific estimation and inference methods used (e.g., uses of auxiliary data, measurement error adjustments, imputation methods, and variance estimation and confidence interval methods). The remainder of this discussion will use the term "empirical effects" collectively for all of these main effects and interactions.

1.2 Utility of observed empirical effects: Predictability, homogeneity and additivity. Implementation of a total survey design approach requires information regarding the relative magnitudes of the abovementioned empirical effects. Because it is often hard to obtain this information, an important question is the extent to which empirical effects in one survey may apply to another survey. To explore this, let M be the objective function of interest. In many cases, M will be a transformation of mean squared error or the total MSE (Smith, 1995). For example, M may be a generalization of design effect or "rate-of-homogeneity" measures considered by Kish (1965, 1995) and others. In addition, let x be a vector representing the abovementioned factors, using indicators x_i for classificatory factors and appropriately scaled x_i for continuous factors. Note that x includes descriptors of the population, design, data collection methods and analysis methods. This is in keeping with the customary view that each of these are integral components of a unified survey procedure and its operating characteristics. Now consider the model,

$$M = (\text{Main effects for } x) \ + \ (\text{Interactions for } x) \ + \ (\text{Equation error}). \quad \ldots(1)$$

One may use model (1) to explore total survey design in a class C, say, of important variables, subpopulations and surveys. Of special interest is the degree of predictability, homogeneity and additivity displayed by model (1) within class C. We will say that the values of M are highly predictable within C if the equation errors in model (1) make a relatively small contribution to the overall variability of M. Also, homogeneity refers to the extent to which the same coefficients in model (1) apply to many or all of the variables and subpopulations within our class C. Finally, we will say that model (1) has a high degree of additivity within C if the interaction effects are relatively small, so that the practical impact of a given factor is well approximated by its associated main effect. Note that the ideas of predictability, homogeneity and additivity are related to the idea of "portability" of rate-of-homogeneity measures discussed in empirical studies of design effects; see, e.g., Kish (1995) and references cited therein.

Ideally, we would like to work with a well-defined class C for which: (i) model (1) displays a high degree of predictibility, homogeneity and additivity; and (ii) relevant high-quality data are available for estimation of the parameters of model (1). For example, under condition (i), model (1) would be dominated by a relatively small number of main effects. In that case, approximately optimum survey procedures follow from minimization of the main-effects sum in (1), subject to cost constraints. Also, for some groups of surveys, there has been extensive empirical work with components believed to dominate model (1). See, e.g., the case studies for incomplete data in Madow *et al.* (1983); work with measurement error in Biemer *et al.* (1991); and the Groves and Couper (1998) empirical study of nonresponse in seven household surveys.

However, practical applications often fall short of the idealized conditions (i)-(ii). For example, many of the sources of variability in model (1) are social characteristics (e.g., of interviewers or respondents). These social characteristics often do not display high levels of predictibility, even at the levels of aggregation in model (1). Also, if model (1) has large interaction terms, then practical implementation of total survey design reduces to the case-specific approach of Linacre and Trewin (1993, p. 620, point (ii)). In that case, model (1) results from one empirical study may shed relatively little light on design of other surveys; cf. Groves (1989). This may be especially problematic if our survey is not a revision of an ongoing series.

Despite these limitations, empirical results (e.g., Linacre and Trewin, 1993, Figure 1) indicate that total survey design methods can lead to substantial improvements in survey efficiency. Thus, it would be useful to expand currently available methods to accomplish the following:

−− Quantify the abovementioned limitations. This would allow assessment of the uncertainty of efficiency gains or losses anticipated for specific proposed changes in design. In addition, this may indicate the extent to which empirical results from other surveys may be applicable to design of a new survey.

−− Identify specific classes C and models (1) that display relatively high levels of predictibility, homogeneity and additivity, based on currently available data. This would help a statistical agency to set priorities, given limited resources available for total survey design work.

−− Identify additional information (e.g., specific new experiments or observational studies) that would be most likely to improve the efficacy of model (1) in total survey design work.

In a formal sense, one may view the three abovementioned tasks in terms of construction of confidence bounds and sensitivity analyses for the response surface defined by model (1). The bounds and sensitivity analyses would be intended to reflect the combined effects of parameter estimation error, equation error, and potential overfitting of model (1). The resulting methods, and a formal asymptotic framework to justify them, would be in part natural extensions of meta-analysis ideas used previously to combine results of multiple studies of empirical effects, e.g., Singer *et al.* (1998); and experimental methods to explore response surfaces, e.g., Box *et al.* (1978) and Myers and Montgomery (1995).

1.3 Formal statement of generally accepted survey procedures. Section 2 of the main paper notes several aspects of sample design on which there is a strong consensus regarding criteria for good survey practice. Two examples are the importance of using a total survey design approach to develop a "cost effective resource allocation plan"; and use of replicate measurements to assess the effects of measurement error. A similar consensus may apply to some of the data collection and processing ideas covered in Section 3. However, the main paper also alludes to important cases in which survey sampling has not succeeded in translating consensus into customary routine practice. To some degree, this may be attributable to the fact that a detailed assessment of good survey practice is somewhat technical in nature, and thus is not immediately accessible to nonspecialists, e.g., nonstatisticians in survey agencies or other funding organizations; journalists; and the general public. In the absence of broadly accessible justification, recommendations for good survey practice can be hampered by budgetary pressures, institutional inertia and information-content variants on Gresham's Law.

Consequently, a natural question is whether communication of technical standards to nonspecialists could be improved by a formal systematic statement of generally accepted survey procedures. To some degree, this would be a natural outgrowth of extensive previous work in, e.g., widely accepted textbooks, or agency statements and manuals on policies and procedures such as U.S. Bureau of Labor Statistics (1992) or "Source and Accuracy Statements" from the U.S. Bureau of the Census (Alexander, 1994, p. 25). However, a thorough treatment of generally accepted survey procedures would be somewhat more comprehensive and more operationally oriented than a textbook; and would potentially be applicable well beyond a specific agency. The result (both in its potential benefits and its potential limitations) might be somewhat analogous to Generally Accepted Accounting Principles used by accountants, or measurement and manufacturing standards used by engineers. Also, in parallel with the accounting and engineering examples, formal statements of standards would be appropriate primarily: (a) in sub-areas that have reached a relatively stable point of development (cf. the comments in Section 1.2 above); and (b) when stated in a form that is sufficiently flexible to reflect the variety and complexity of problems encountered in applied survey work.

2. Inference from Complex Survey Data

2.1 Assessment of inferential robustness and efficiency: power curves and related surfaces. To expand on the Section 4 discussion of inferential issues, note that the sample survey literature uses several criteria to evaluate the performance of point estimation and inference methods. In some cases (e.g., Section 4.2 of the main paper), principal interest focuses on the robustness of a given method against deviations from idealized conditions (e.g., specific levels of conditioning; assumed relationships between survey characteristics and auxiliary variables; or magnitudes of nonresponse or measurement error effects). Common measures of robustness include the biases of point estimators and the true coverage rates of nominal $(1-\alpha)100\%$ confidence intervals.

In other cases, it is also important to consider efficiency, as measured, e.g., by point estimator variances and confidence interval widths. The survey literature reflects differing opinions regarding the relative importance of robustness and efficiency; see, e.g., Hansen *et al.* (1983), Smith (1994) and references cited therein. At a minimum, however, it is useful to examine potential trade-offs between robustness and efficiency in specific analyses. For point estimators, we often assess these trade-offs with conditional or unconditional mean squared errors. For standard pivotal-quantity-based inference methods, we can evaluate related trade-offs through hypothesis-test power curves. To illustrate this idea, consider first a univariate parameter θ, a point null hypothesis $\mathrm{H}_0 : \theta = \theta_0$, and an associated test statistic $t_0 = (\hat{\theta} - \theta_0)/se(\hat{\theta})$. A plot of the power of this test against various values of θ gives an indication of the extent to which t_0 is useful in identifying deviations from H_0 that are considered of substantive importance for scientific or policy work.

Now consider several test statistics $t_1, t_2, \ldots$, derived from competing methods, e.g., based on different uses of auxiliary information. Overlaid power curves (where in each case power is evaluated with respect to the sample design and relevant models for nonsampling error) for $t_1, t_2, \ldots$ then give some indication of the practical impact of robustness-efficiency trade-offs for these competing methods. For example (cf. Section 2 of the main paper), in some applications small amounts of measurement error may produce small biases in estimators of regression coefficients or quantiles. The resulting tests will have corresponding distortions that may seriously inflate the true type I error rate, and may also reduce the power to identify small-to-moderate deviations from H_0. Measurement-error adjustment methods may reduce or remove these distortions, but (primarily due to variance inflation) may seriously reduce the power to identify moderate-to-large deviations from H_0 for samples of moderate size.

Next, note that these power curves generally will depend on the values of specific parameters. For instance, in our measurement-error example, the location and shape of each power curve depend on the magnitude of measurement error variances, relative to true-value variances and effective sample sizes. Expansion of a two-dimensional power curve into a third dimension representing measurement error variance thus defines a power-curve surface. Comparison of these surfaces for test methods $t_1, t_2, \ldots$ then will give a somewhat richer indication of robustness-efficiency trade-offs. Similar comments apply to other (non-measurement-error) problems for which robustness and efficiency properties depend heavily on the empirical characteristics of the population or procedure. Consequently, many of the empirical-effect-assessment ideas from Section 1.2 of this discussion may also apply to the power-curve questions considered here. In essence, information regarding the magnitudes of specific empirical effects gives an indication of the approximate location of a given analysis problem on the abovementioned multidimensional power-curve surface. Details for computation of these surfaces can be nontrivial, and will be considered elsewhere.

2.2 Inference for superpopulation parameters. Section 5 of the main paper ends with a brief discussion of inference for superpopulation parameters, and of the recent work by Korn and Graubard (1998) on inference for superpopulations that are stratified or clustered. Two related points are as follows. First, the Korn and Graubard (1998) results indicate that ignoring superpopulation-level stratification and clustering effects will be problematic primarily when sample fractions are not small or intracluster correlations are strong. Consequently, it would be of interest to develop empirical information (beyond the example in Korn and Graubard, 1998) identifying some cases with nontrivial superpopulation-level mean patterns and intracluster correlations. To some degree, this would be an extension of previous empirical work with intracluster correlations (for design-determined clusters) and classical design effects; see, e.g., Verma *et al.* (1980), Verma and Le (1996) and references cited therein.

Second, for models like those considered by Korn and Graubard (1998), it is natural to view superpopulation inference roughly in the framework of two-phase sampling. In this setting, the first phase is generation of a finite population through a stratified and clustered superpopulation model; and the second phase is the customary selection of a stratified multistage sample. Note especially that for some of the models considered by Korn and Graubard (1998), superpopulation-level stratification or cluster boundaries are not necessarily identical to the boundaries (often arising from somewhat artificial administrative or geographical classifications) imposed by the sample design. This has parallels in standard two-phase designs where the second phase design may cut across first-phase cluster or stratum boundaries.

Acknowledgements. The author thanks Barry Graubard and Eleanor Singer for providing preprints of Korn and Graubard (1998) and Singer *et al.* (1998), respectively. This research was supported in part by the U.S. National Center for Health Statistics. The views expressed here are those of the author and do not necessarily represent the policies of the U.S. National Center for Health Statistics.

References

Alexander, C.H. (1994), Discussion of Paper by Smith. *International Statistical Review* **62**, 21-28.

Andersen, R., Kasper, J. and Frankel, M. (1979). *Total Survey Error: Applications to Improve Health Surveys.* Jossey-Bass, San Francisco.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (1990). *Measurement Errors in Surveys.* Wiley, New York.

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building.* Wiley, New York.

Bureau of Labor Statistics (1992). *BLS Handbook of Methods*, Bulletin 2414, United Stated Department of Labor. Washington, DC: U.S. Government Printing Office.

Groves, R.M. (1989). *Survey Errors and Survey Costs.* Wiley, New York.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys.* Wiley, New York.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). Rejoinder to Comments by Royall, Little, Dalenius, Smith and Rubin, *Journal of the American Statistical Association* **78**, 805-807.

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

– – –– (1995). Methods for Design Effects, *Journal of Official Statistics* **11**, 55-77.

Madow, W.G., Nisselson, J. and Olkin, I., eds. (1983). *Incomplete Data in Sample Surveys (Volume 1): Report and Case Studies.* Academic Press, New York.

Myers, R.H. and Montgomery, D.C. (1995), *Response Surface Methodology: Process and Produce Optimization Using Designed Experiments.* New York: Wiley.

Singer, E., Gebler, N., Raghunathan, T., Van Hoewyk, J. and McGonagle, K. (1998). The effect of incentives on response rates in face-to-face and telephone surveys, *J. Off. Statist.*, to appear.

Smith, T.M.F. (1983). Comment on Paper by Hansen, Madow and Tepping, *Jour. Amer. Statist. Assoc.* **78**, 801-802.

– – –– (1994). Sample Surveys 1975-1990: An Age of Reconciliation? (with discussion). *International Statistical Review* **62**, 5-34.

Verma, V. and Le, T. (1996). An Analysis of Sampling Errors for the Demographic and Health Surveys. *International Statistical Review* **64**, 265-294.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample Designs and Sampling Errors for the World Fertility Survey (with discussion), *Jour. Royal Statist. Soc.,* Series A **143**, 431-473.

Discussant : Robert E. Fay
U.S. Bureau of the Census

1. **Introduction**

From its very beginning, the Hansen Memorial Lecture Series has been a remarkable institution of the Washington Statistical Society. The breadth of topics addressed in the series, each a reflection of Morris Hansen's wide contributions and interests, has been a fitting tribute to his professional life.

The presentation by Prof. J.N.K. Rao and the paper on which it was based honor Hansen's contributions to the foundations of probability sampling in survey research, by showing just how vital and dynamic the field remains. Rao himself is a significant source of this vitality, but his paper also summarizes the contributions of many other researchers to the continued growth of the field.

A number of participants at the session have been able to reflect widely and with detailed anecdotes about Hansen's life. I believe that many current members of our profession, however, did not have the opportunity to meet him. I must confess that my own contact with him was limited – enough to be on a first name basis and to have enjoyed a number of conversations – but much less than the close working association that many of my Westat colleagues enjoyed. Nonetheless, I can attest to his generosity, kindness, creativity, and enthusiasm for his profession. I recall his thoughtful presence and contributions to Census Advisory Committees. I also remember with gratitude the personal interest he expressed in my early work some years ago, which was generally characteristic of his treatment of younger professionals.

Rao's paper does what a good review paper should do, surveying the history of the development of the field, summarizing the state of current research, and suggesting possible emerging directions. Such a paper provides almost unlimited points of departure for a discussant. To narrow the scope, if only a bit, this discussion will attempt to relate the paper to Hansen's contributions and interests.

2. Foundations

The first sentence of the paper's abstract, "Beginning with the pioneering contributions of Neyman, Hansen, Mahalanobis and others, a large part of sample survey theory has been directly motivated by practical problems encountered in the design and analysis of large scale sample surveys." recognizes Hansen's fundamental contributions to the field. More specifically, however, the paper cites Hansen *et al.* (1983) for its restatement of the inferential basis of the classical theory of design-based inference. In large part, Hansen *et al.* (1983) was a defense of this approach and emphasis of its strengths relative to model-dependent approaches (Brewer 1963) advanced particularly pointedly by Royall (1970) and others. The example discussed by Hansen and his colleagues illustrated a potential dilemma faced in applications of model-based inference; in the example, the observed data typically provided insufficient evidence to reject a presumed model, yet inferences under the model were severely affected by the undetected departures. On the other hand, the design-based inference was consistently robust. Hansen *et al.* (1983) were clear, however, that the design-based approach was a large-sample theory, generally suitable to the moderate to large samples found in survey practice, but not confidently extended to problems involving very small samples.

Rao's paper revisits one of the potential limitations of design-based inference. The problem of Basu's (1971) circus, which included the notorious elephant, is reviewed in the same section. Basu argued that sampling the population of animals in a circus posed an inherent problem for design-based inference, since confidence intervals based on observed samples would be too short in samples that excluded the elephant. More generally, characterizations of design-based inference, including Hansen *et al.* (1983), often emphasize unconditional properties, yet one of the criticisms of Royall was that unconditional inference ignored potentially useful auxiliary information from the realized sample.

Significantly, Rao's paper presents an advance in design-based inference, showing how it can be extended to conditional inference.

3. The Theory in Practice

Many, including Hansen, have recognized that the design-based approach does not exclude models from important roles in survey practice. It is well recognized that models, whether explicit or implicit, underlie virtually all adjustments for nonresponse and other forms of missing data. Hansen participated in a early survey using regression analysis to form small domain estimates, an area of application that has grown to merit its own section in Rao's paper. The appropriate balance and interaction of the design-based theory with presumed models has posed interesting issues in both of these fields.

Many other aspects of survey research are based on informal extensions of this theory. Use of sampling designs for which there is no consistent estimator of the sampling variance, such as the usual form of systematic sampling, is, in principle, a problem for the theory. On the other hand, practitioners, largely through inductive rather than deductive reasoning, have grown quite comfortable with such practical extensions.

Practitioners may often not distinguish between unconditional and conditional inference. For example, in some settings, replication methods such as the jackknife provide an estimate of a conditional variance.

The practice of variance generalization provides another example of an informal extension. Originally, computer resources severely limited the number of direct variance estimates that could be computed from a large survey, so that prediction of variance through a model was the only practical option. As computing power has grown exponentially, computing large numbers of direct variance estimates has increasingly become feasible, so the issue of direct vs. model-based variance estimates has the same trade-off of variance vs. bias as small domain estimation. I have observed many colleagues to prefer variance generalizations to direct variance estimates without formal consideration of the trade-off, yet such a preference lacks a clear theoretical foundation.

Recently, Lee (1998) provided a careful review of the treatment of outliers in survey data. He stressed the importance of domain specific information in developing strategies to handle outliers, making a unified approach to outlier adjustment unlikely. Within the design-based approach, there can be available strategies to limit the effect of outliers at the design stage, if appropriate auxiliary information is available. Many strategies at the estimation stage, however, involve departures, to varying degrees, from the design-based framework. For example, treating Basu's elephant simply as an outlier to be downweighted may appear superficially to be an adequate solution, rationalized by parallels to data analysis, but it departs from Hansen's basic program.

Rao's review of inferential issues brings to attention the distinct boundaries of this theory and the extent to which current practice may extend practice beyond its theoretical foundations. Many of these extensions may be quite useful and may have unstated rationales outside of the core theory, but Rao's summary serves as a reminder of where the boundary is.

Remarks on Other Sections. The review of resampling methods demonstrates that this approach continues to hold promise as a general strategy for variance estimation in many complex situations. One of the potential strengths of this approach is the possible extension to a number of missing data situations. Rao provides a few relevant citations, including Rao and Shao (1992), but some additional advances have been made and many more may await discovery. The potential is for a statistical treatment of missing data with closer integration to design-based theory than offered by multiple imputation.

As a related issue, the literature on replication methods generally considers a set of discrete strategies. Faced with a highly complex design, however, there are instances in which hybrid strategies may be more appropriate; for example, by using some resampling methods for some sampling strata and others for other strata. Although the consistency of the variance estimate for linear estimates and smooth nonlinear estimates may be clear, other theoretical results, such as the consistency of the variance estimate for medians, may not be. A similar problem occurs for replication methods modified to account for missing data variance, where the statement of existing asymptotics may not explicitly include these methods. A future direction for the replication literature, then, would be to identify classes of replication methods, perhaps based on the order of variation of replicates about the sample estimate, which could then be perhaps shown to have common properties as classes.

Rao encourages more explicit consideration of both bias and variance in designing surveys, citing the application by Linacre and Trewin (1993) as an example. Although this seems a rational direction, the discussant's experience is that enormous difficulties lie ahead for most applications of this idea. Rao recognizes the necessity to allocate resources to necessary evaluation studies, but in many evaluation studies, if based on reinterview or other survey-based approaches, often are subject to their own biases. This area is one than will require attention to the survey methodology and other aspects of the empirical science, as well as to the mathematical framework for the effort.

The author is to be congratulated on this excellent review.

References

Lee, Hyunshik (1998), Outliers in survey sampling, presentation to the *Washington Statistical Society*, April 29, 1998, Washington, DC.

Discussant : J.K. Ghosh
Indian Statistical Institute and *Purdue University*

What a splendid collection of ideas, methods and a balanced view of what‘s happening in the area of survey sampling. I find so much that is new to me but interesting, useful and relevant for the entire statistical profession that I cannot but mourn the fact that much of the material referred to in the text is not easily accessible. Moreover, most of the material is fairly new and only available in journals. Hopefully, someone will write a new book that makes all this material easily accessible.

However, having listened to the lecture on which it is based, one cannot but miss the humor and passion that rocked one with laughter or made one wince when a stab went home.

The whole paper is a veritable feast but there are many special delights some of which I proceed to list. Cost is reduced by taking into account nonsampling errors and then choosing the best option. Adjustment for bias due to measurement error is available for non-linear parameters like quantiles through modelling the error of distribution. Even without interpenetrating samples one can adjust for measurement error for the total by modelling effects of interviewers, coder and so on. Even the choice of sample size can take into account the multiple objectives of accuracy at different levels of disaggregation. At the level of data processing, there are possibilities of developing and training neural networks to edit as well or nearly as well as experts. This may improve existing automatic editing methods. Conditioning in the form of post stratification seems to have become an accepted practice. (Surely, D. Basu's polemical essays (Basu, 1988) had something to do with this?) The possible absurdities in the Horvitz-Thompson estimate - renamed NHT honouring another pioneer Narain - are at least partly resolved by multiplying a plausible correction factor or adjusting through regression. (But one can still construct counter examples.) If I do not include in this list the use of resampling or model based small area estimation it is only because I was somewhat familiar with the progress in these areas.

A number of interesting issues arise as one goes over the list. On the one hand model based sampling, once advocated by Godambe, Royall and others, does not seem to have done well. But model based techniques, in the form of hierarchical Bayes or empirical Bayes analysis or use of modelling to provide estimates at the level of small areas or adjust for bias or nonsampling error are both in vogue and apparently stand up well to scrutiny. Is it the case that one resorts to modelling or Bayesian ideas when nothing else is available but design based methods, adjusted for possible biases under conditioning, remain trusted mainstream tools? Nonetheless success of modelling for small area estimates and its failure at higher level of aggregation call for more analysis. I will return to this point later.

The acceptance of the conditionality principle and the resulting adjustments for the Narain-Horvitz-Thompson estimate warm my heart. From frequent references made to Basu's circus, I presume Basu was at least partly responsible for this change. I note that Professor Rao had indeed pointed out much earlier than Basu that the estimate can misbehave if Y and the probabilities of selection are uncorrelated. But Basu pointed out the disaster that takes place when a large value gets associated with a small probability - rare but possible. The circus statistician lost his job because he was prepared to estimate the total weight of 50 elephants by (4900). (weight of Jumbo), Jumbo being the heaviest elephant.

What puzzles me in the course of events since that example is the estimate has been adjusted but not abandoned. I understand the pps scheme is minimax under certain assumptions but apparently the reason for the popularity of such a design and the NHT estimate lies not in such optimality results but rather in considerations of convenience such as self-weighing and so on. But in these days of enormously powerful computing and surveys by telephone is such convenience really as compelling as before?

Professor Rao refers to a modest contribution of mine (Ghosh, 1992) and points out that I proved the consistency of the estimate even when the probabilities and the response are not positively correlated. I myself found this result somewhat surprising but realized later that in such cases as Basu's circus it can be easily shown that the rate of convergence is appallingly slow because the variance has been inflated to produce an unbiased estimate when a small bias would have been more judicious. The adjustments proposed introduce some bias even though they perform well in the simulations reported in the paper. Incidentally, a very thorough study of the asymptotic distribution of the NHT estimate is available in Gordon (1981). Some of the asymptotics about NHT estimates has been useful in the entirely different area of forecasting discovery of oil reserves.

In the rest of this discussion I will focus on the possibilities of model based inference and methods in large scale surveys, pps sampling and the NHT estimate.

First, the apparent failure of model based or Bayesian inference which are treated as essentially the same thing below. Can it be that it has worked in small area estimation but not large surveys simply because modelling must be much more complex when the sample size is large? It is a trivial fact that any simple model would prove to be false in large samples. Consequently what would be needed are methods of nonlinear and nonparametric regression rather than simple linear normal models. Alternatively, one can try linear models whose dimension or complexity increases with sample size rather like exponential families of increasing dimension suggested by Efron, Hjort and others for density estimation. I wonder if anything in this direction has been tried and found to fail in the context of survey sampling. If not, that's an area worth exploring. I vaguely recall T. M. F. Smith reported some preliminary work of this sort at the conference for Godambe at Waterloo in 1991. That had seemed quite successful.

In the case of pps sampling, if convenience or optimality requires that it it still necessary to use it, one would want to see if it is possible to manage with a few probabilities of selection corresponding to a few size classes rather than a probability corresponding to every size. The size classes would then be much easier to deal with and the asymptotic behavior of the estimate less of a mystery than now and one can also think of alternative estimates. This is what I had done in (Ghosh, 1992). Probably others have done too.

The NHT estimate itself has acquired new significance in Bayesian analysis. An example due to Robins and Ritov (1997) has been described by Wasserman (1998) as follows. The auxiliary random variable X_i has known density f_X. Given X_i one observes Y_i with probability π_i and does not observe with probability $1 - \pi_i$. The π_i's are a known function $g(X_i)$ of X_i. There is an indicator R which tells you if Y_i is observed or not. The Y_i's are binary random variables, X_i's live in a high dimensional Euclidian space and the object of interest is the unknown $P(Y = 1|X) \equiv h(X)$. Ritov and Robins (Robins *et al.*, 1997) have shown that a Bayesian analysis with a prior for $h(\cdot)$ provides a poor estimate for $E(Y) = \int h(x) f_X(x) dx$ which is what one wants to estimate. It is clearly very similar to, if not identical with, the problem of estimating the population mean in the presence of pps sampling. Robins and Ritov (Robins *et al.*, 1997), provide an essentially NHT estimate and show it has $n^{-\frac{1}{2}}$ rate of convergence and other good properties.

The story seems to have implication for estimating a parameter in the presence of a high dimensional nuisance parameter. Here all of $h(\cdot)$ that one can abstract away after fixing $E(Y)$ is a nuisance parameter. In such problems the prior on the high dimensional nuisance parameter can often cause problems in estimating the parameter of interest as in Neyman - Scott problems which Joydeep Bhanja had studied in his ISI thesis in the early nineties. These problems are better handled by conditional or partial likelihood or semiparametric methods. This suggests that at least in some cases even flexible high dimensional modelling that I suggested earlier may not work. One may have to do something like what is suggested below in the Bayesian digression.

Let me round off this discussion with a bit of Bayesian digression on the problem of Robins and Ritov $[Y]$ as posed by Wasserman (Wasserman, 1998). The "observed" Y_i's are *i.i.d* Bernoulli variables with $\theta = E(Y)$ as the parameter if one ignores the high dimensional co-variate. Natural $n^{-\frac{1}{2}}$ consistent Bayes estimates are then easy to construct though by virtue of ignoring the co-variates they will not be optimal in a frequentist sense. A part or all of the co-variates may be kept if more information is available. It seems to me such estimates are natural for a Bayesian when modelling through co-variates leads to a high dimensional nuisance parameter and extracts an enormous price unless the function $h(\cdot)$ is nice. Similar phenomena were observed by Mark Berliner in Bayesian prediction when data has a chaotic component and one uses the entire past data to predict. Prediction based on the immediate past is better and more natural.

Lack of frequentist optimality is a price one pays for avoiding an intuitively strange estimate. In any case it is interesting to see the old NHT estimate reappearing in a new dress to puzzle Bayesians once more.

References

Basu, D. (1988). *A Collection of Critical Essays* (J. K. Ghosh, Ed) Springer, New York.

Ghosh, J. K. (1992). The Horvitz-Thompson estimate and Basu's circus revisited, *Proc. of the Indo - US Bayesian Workshop* (Eds P. K. Goel *et al*), 225-228.

Gordon, L. (1983). Successive sampling in large finite populations, *Ann. Statist.*, **12**, 702-706.

Robins, J. and Ritov, Y. (1997). Towards a curse of dimensionality appropriate asymptotic theory for semiparametric models. To Appear in *Statistics and Medicine.*

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures, in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds D. Dey *et al*, 293-302, Springer, New York.

Discussant : Malay Ghosh
University of Florida

We are indeed fortunate to have such a comprehensive review from one of the maestros of survey sampling. Professor Rao's pioneering contributions to this field, spanning over more than three decades, has made a lasting impact on both methodologists and practitioners alike.

In this review article, Professor Rao has touched upon many aspects of survey sampling, beginning with the important issues of how to design a survey followed by collection and processing of data. Clearly, these are important practical issues, often neglected by theorists, but cannot be brushed aside, especially by those who are engaged in conducting the actual survey. Currently, at the University of Florida, we are involved in designing a survey targeted to find estimates of the proportions of people with no health insurance in the different districts of Florida. Random digit dialing (RDD) will be used to collect the data. The task is to find a suitable design, not necessarily "optimal" according to some textbook criterion, but one which can guard against large potential nonresponse, and especially is capable of oversampling the underprivileged sectors of the community. For a survey of this type, the notion of total (sampling + nonsampling) survey error is all the more important.

Next we turn to inferential issues. Model-versus design-based estimation in survey sampling has been a topic of heated debate for several decades. Fortunately, there is a growing realization now that model-assisted design-unbiased or design-consistent estimators enjoy the best of both worlds. On one hand, they guard against any possible model specification, while on the other, they are "optimal" or "near-optimal" under the assumed model. The generalized regression estimator $\hat{Y}_{gr}$ introduced in (4.5) and (4.6) is a example of this type.

One way to see this optimality is via estimating functions introduced by Godambe (1960), and extended to survey sampling by Godambe and Thompson (1986). Consider a finite population U with units labeled $1, \cdots, N$. With each individual j, there is an unknown characteristic of interest y_j, and a known vector of covariates $\boldsymbol{x}_j = (x_{j1}, \cdots, x_{jp})^T$. It is assumed that $\boldsymbol{y} = (y_1, \cdots, y_N)^T$ is generated from a distribution ξ belonging to a class C. The class C is usually referred to as a *superpopulation model.*

Letting E_ξ denote expectation with respect to the distribution ξ, consider those distributions ξ under which

(i) $E_\xi(y_j) = \boldsymbol{x}_j^T \boldsymbol{\beta}$, $(j = 1, \cdots, N)$;

(ii) $E_\xi(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})^2 = \sigma^2 q_j (j = 1, \cdots, N)$;

(iii) $E_\xi[(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})(y_{j'} - x_{j'}^T\boldsymbol{\beta})] = 0 \quad (1 \le j \ne j' \le N)$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ $(p < N)$ is the vector of regression parameters. A vector-valued estimating function $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta})$ is said to be *linearly unbiased* in $y_1, \cdots, y_N$ if it has the form

$$\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{j=1}^{N} (y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{a}_j(\boldsymbol{\beta}). \qquad \ldots (1)$$

We restrict attention only to those $\boldsymbol{g}$ for which $\boldsymbol{V}(\boldsymbol{g}) = \sigma^2 \sum_{j=1}^{N} q_j \boldsymbol{a}_j(\boldsymbol{\beta})\boldsymbol{a}_j^T(\boldsymbol{\beta})$ is positive definite. Let $\boldsymbol{D}_{\boldsymbol{g}} = E_{\boldsymbol{\xi}}\left(\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\beta}}\right) = \boldsymbol{X}^T\boldsymbol{a}(\boldsymbol{\beta})$, where $\boldsymbol{X}^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ and $\boldsymbol{a}^T(\boldsymbol{\beta}) = (a_1(\boldsymbol{\beta}), \cdots, a_N(\boldsymbol{\beta}))$. Following Chandrasekar and Kale (1984), an estimating function $\boldsymbol{g}^*(\boldsymbol{y}, \boldsymbol{\beta})$ is said to be "linearly optimal" if $\boldsymbol{g}^*(\boldsymbol{y}, \boldsymbol{\beta})$ is linearly unbiased and $\boldsymbol{D}_{g^*}\boldsymbol{V}^{-1}(\boldsymbol{g}^*)\boldsymbol{D}_{g^*}^T \ge \boldsymbol{D}_g\boldsymbol{V}^{-1}(\boldsymbol{g})\boldsymbol{D}_g^T$. From Godambe and Thompson (1986), under (i)-(iii), the linearly optimal estimating function $\boldsymbol{g}^*$ is given by

$$\boldsymbol{g}^*(\boldsymbol{y}, \boldsymbol{\beta}) = \sigma^{-2} \sum_{j=1}^{N} q_j^{-1}(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j. \qquad \ldots (2)$$

Also, following these authors, a survey population parameter $\boldsymbol{\beta}_N$ is a solution of the linearly optimal estimating equation $\boldsymbol{g}^*(\boldsymbol{y}, \boldsymbol{\beta}) = \boldsymbol{0}$, and is given by $\boldsymbol{\beta}_N = (\boldsymbol{X}^T\boldsymbol{Q}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Q}^{-1}\boldsymbol{y}$.

In practice, it is more important to estimate the survey population parameter $\boldsymbol{\beta}_N$ from a sample drawn from the survey population. The following approach due to Godambe and Thompson (1986) shows how the estimating function theory can be used here.

A sample s is a subset of $\{1, \cdots, N\}$. Let $S = \{s\}$ denote the set of all possible samples. A *sampling design* p is a probability distribution on S such that $p(s)\epsilon[0, 1]$ and $\sum_{s\epsilon S} p(s) = 1$. The data when a sample s is drawn, and the corresponding y-values are observed will be denoted by $\mathcal{X}_s = \{(j, y_j) : j\epsilon s\}$. Let $\pi_j = \sum_{j\epsilon s} p(s)$ denote the inclusion probability of the jth population unit $(j = 1, \cdots, N)$. Now a function $\boldsymbol{h}(\mathcal{X}_s, \boldsymbol{\beta})$ is said to satisfy the criterion of design unbiasedness under ξ if

$$E_\xi[\boldsymbol{h}(\mathcal{X}_s, \boldsymbol{\beta})] = \sigma^{-2} \sum_{j=1}^{N} q_j^{-1}(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j \qquad \ldots (3)$$

for each $\boldsymbol{y}$ and $\boldsymbol{\beta}$. The optimal $\boldsymbol{h}$ as found in Godambe and Thompson (1986) is given by $\sigma^{-2}\sum_{j\epsilon s} w_j(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j$; where $w_j = \pi_j^{-1}$. The corresponding optimal estimator of $\boldsymbol{\beta}$ from the sample data is obtained by solving $\sum_{j\epsilon s} a_j^{-1} w_j (y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j = \boldsymbol{0}$ and is given by $\hat{\boldsymbol{\beta}}_s = (\boldsymbol{X}^T(s)Q^{-1}\boldsymbol{X}^T(s))^{-1}\boldsymbol{X}^T(s)\boldsymbol{Q}^{-1}\boldsymbol{y}_s$, where $\boldsymbol{X}(s)$ and $\boldsymbol{y}_s$ are sample analogs of $\boldsymbol{X}$ and $\boldsymbol{y}$ based on the $(y_j, \boldsymbol{x}_j; j\epsilon s)$. It is assumed here that rank $(\boldsymbol{X}(s)) = p < n(s), n(s)$ being the size of the sample s. Plugging in the solution $\hat{\boldsymbol{\beta}}_s$ for $\boldsymbol{\beta}$, an "optimal" estimator for the finite population total $\sum_{j=1}^N y_j = \sum_{j=1}^N \boldsymbol{x}_j^T\boldsymbol{\beta} + \sum_{j=1}^N (y_j - \boldsymbol{x}_j^T\boldsymbol{\beta})$ is given by $\sum_{j=1}^N \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_s + \sum_{j\epsilon s} w_j(y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}_s)$ which equals $\hat{Y}_{gr}$. We may add here that in the current state of the art, the sampling design is usually taken to be noninformative. This need not always be so in real surveys. For informative sampling, the work of Pfeffermann and his colleagues should be relevant. In particular, informative sampling will change the likelihood, and may yield different results especially for the estimates of the variances of estimates. A good source of references in this area is Pfeffermann (1993). It may be useful to come up with estimators alternative to $\hat{Y}_{gr}$ under different informative sampling schemes.

My final comment relates to small area estimation. We all have benefited from Professor Rao's fundamental contributions to this topic. This review article contains a short and yet informative summary of what has so far been accomplished. I mention an important real life situation which requires a slight extension of the models that have been discussed here.

To be specific, consider stratified two-stage sampling. The local areas consist of primary sampling units which cut across stratum boundaries. As an example, one may consider stratified two-stage sampling, where the strata are different counties in the state of Florida, the primary sampling units are the census tracts within these counties, and the secondary units or the subunits are individuals within these census tracts. One of the issues of critical importance is to estimate the poverty rate for different local areas within the state of Florida. These local areas need not always be geographic areas, but are formed by crossing several ethnic and demographic categories. For example, one may be interested in finding the proportion of non-Hispanic white males in the age-group 20-24 who are below the poverty level, proportion of Hispanic females in the age-group 40-44 who are below poverty level etc. Direct survey estimates are very unreliable here because the original survey was targeted at the national level, and the state level estimate for any individual category is usually accompanied by large standard error or coefficient of variation.

To formulate the problem in the context of finite population sampling, suppose there are m local areas. The ith local area consists of M_i primary units which cut across the different strata. For the jth primary unit in the ith local area, there are N_{ij} subunits. We denote by y_{ijq} the value of the characteristic of interest for the qth subunit belonging to the jth primary unit within the ith local area. Also, subunit specific auxiliary characteristics $\boldsymbol{x}_{ijq}$ are available. In special cases, such auxiliary characteristics may not depend on the subunits, for example $\boldsymbol{x}_{ijq} = \boldsymbol{x}_{ij}$. Suppose the number of strata is K. The objective is to estimate the local area means $\sum_{j=1}^{M_i}\sum_{q=1}^{N_{ij}} y_{ijq} / \sum_{j=1}^{M_i} N_{ij}$ based on a sample of m_i primary units, and a sample of size n_{ij} for the jth selected primary unit. For notational simplicity, the selected primary units within the ith local area are denoted by $1, \cdots, m_i$, while the sampled units for the jth selected primary unit are denoted by $1, \cdots, n_{ij}$.

We begin with a generalized linear model for the y_{ijq} given by

$$f(y_{ijq}|\theta_{ijq}) = \exp[\{y_{ijq}\theta_{ijq} - \psi(\theta_{ijq})\}/\phi_{ijq} + c(y_{ijq}, \phi_{ijq})],$$

where the ϕ_{ijq} are known. At the next stage, we model the θ_{ijq} as

$$\theta_{ijq} = \boldsymbol{x}_{ijq}^T \boldsymbol{b} + \alpha_i + \eta_{ij} + \sum_{k=1}^{K} \gamma_k \boldsymbol{I}_{[j\epsilon k]} + \sum_{k=1}^{K} \xi_{ik} \boldsymbol{I}_{[j\epsilon k]} + e_{ijq},$$

where $\boldsymbol{b}$ denotes the vector of regression coefficients, α_i is the effect of the ith local area, η_{ij} is the effect of the jth primary unit within the ith local area, γ_k is the effect of the kth stratum, $\boldsymbol{I}$ is the usual indicator function, ξ_{ik} is the interaction effect between the ith local area, and the kth stratum, e_{ijq} are the errors assumed to be iid $N(0, \sigma^2)$. It is also assumed that the $\alpha_i, \eta_{ij}, \gamma_k$, and ξ_{ik} are mutually independent with the α_i iid $N(0, \sigma_\alpha^2), \eta_{ij}$ iid $N(0, \sigma_\eta^2), \gamma_k$ iid $N(0, \sigma_\gamma^2)$, and ξ_{ik} iid $N(0, \sigma_\xi^2)$. At this point, it is possible to adopt either a Bayesian or a frequentist approach for estimating the small area parameters θ_{ijq}.

References

Chandrasekar, B. and Kale, B.K. (1984). Unbiased statistical estimation functions in the presence of nuisance parameters. *Journal of Statistical Planning and Inference*, **9**, 45-54.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208-1212.

Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population, their relationship and estimation. *International Statistical Institute Review*, **54**, 127-138.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Institute Review*, **61**, 317-337.

Discussant : P.Lahiri
University of Nebraska-Lincoln

Professor Rao's expertise on survey sampling has been clearly reflected in this excellent review paper. I would like to congratulate Professor Rao for nicely covering a wide variety of research areas in survey sampling within such a short space.

I will narrowly focus my discussions on small-area estimation. An accurate estimation of the MSE of EBLUP which captures all sources of variabilities has been a major research activity for the last decade. Professor Rao has nicely summarized the work in this area. Most of the papers available in the literature on this important topic are normality-based although the normality assumption is not needed to justify EBLUP. Thus, it seems natural to develop MSE estimator of EBLUP whose validity does not require the normality assumption.

For the simplicity of exposition, let us consider model (6.4) of Professor Rao's paper. As mentioned in the paper, Lahiri and Rao (1995) replaced the normality of the random effects v_i in the model by certain moment assumptions and showed the robustness of the normality-based Prasad-Rao MSE estimator. However, the preliminary research conducted by Professor Rao and myself indicates that the Prasad-Rao MSE estimator is not robust under other important models such as the nested error regression model considered by Battese *et al.* (1988). Certain kurtosis terms of the distribution of random effects and errors appear in the MSE approximation and it certainly becomes a messy problem.

Note that EBLUP is well justified under the assumption of *posterior linearity* since in this case EBLUP is empirical best predictor (EBP) or, in the Bayesian language, linear empirical Bayes (LEB) estimator. For a good discussion on posterior linearity, the reader is referred to the recent monograph by Ghosh and Meeden (1997). If one does not believe in the posterior linearity assumption, one should not even consider EBLUP in the first place since EBLUP will not then be EBP. Thus, the posterior linearity assumption seems to be a very reasonable non-normal assumption for the purpose of obtaining an accurate MSE estimator of EBLUP.

In the context of estimation of drug prevalence for small-areas, my colleagues and I (see Chattopadhayay *et al.* 1999) developed a jackknife estimator of MSE of LEB estimator (identical with EBLUP). Let me explain the procedure for model (6.4). Let $\tilde{\theta}_i(\hat{\theta}_i;\phi)$ denote the BP (also the linear Bayes estimator) of θ_i, where $\phi = (\beta, \sigma_v^2)$. Let $\hat{\phi}$ be a consistent estimator of ϕ and $\tilde{\theta}_i(\hat{\theta}_i;\hat{\phi})$ be the corresponding EBP of θ_i. Here EBP (also LEB) and EBLUP are identical.

Under the posterior linearity assumption, one gets

$$MSE[\tilde{\theta}_i(\hat{\theta}_i;\hat{\phi})] \quad = \quad MSE[\tilde{\theta}_i(\hat{\theta}_i;\phi)] + E[\tilde{\theta}_i(\hat{\theta}_i;\hat{\phi}) - \tilde{\theta}_i(\hat{\theta}_i;\phi)]^2\ . \qquad \ldots(1)$$

Note that the above decomposition is a standard identity in the Bayesian calculations and is different from that of Kacker and Harville (1984) who validated their identity only for *normal* mixed linear models. As a result, all the subsequent papers on MSE estimators of EBLUP which used the Kackar-Harville decomposition are normalily-based.

It is easy to check that $MSE[\tilde{\theta}_i(\hat{\theta}_i;\phi)] = g_{1i}(\sigma_v^2)$. Prasad and Rao (1990) (also see Lahiri and Rao 1995) noted that the bias in $g_{1i}(\hat{\sigma}_v^2)$ is of order $O(m^{-1})$ and corrected the bias by a Taylor series method. One can instead use a jackknife method and propose a bias-corrected estimator of the first term in the right hand side of (1) as:

$$mse_i^{BP} \;=\; g_{1i}(\hat{\sigma}_v^2) - \frac{m-1}{m}\sum_{u=1}^{m}\{g_{1i}[\hat{\sigma}_v^2(-u)] - g_{1i}(\hat{\sigma}_v^2)\}\ , \qquad \ldots(2)$$

where $\hat{\sigma}_v^2(-u)$ is an estimator of σ_v^2 calculated similarly as $\hat{\sigma}_v^2$ with data from all but the uth small-area. The second term in (1) is estimated by

$$\hat{E}_i \;=\; \frac{m-1}{m}\sum_{u=1}^{m}\{\tilde{\theta}_i[\hat{\theta}_i;\hat{\phi}(-u)] - \tilde{\theta}_i(\hat{\theta}_i;\hat{\phi})\}^2\ , \qquad \ldots(3)$$

where the definition of $\hat{\phi}(-u)$ is similar to that of $\hat{\sigma}_v^2(-u)$. Note that to jackknife the second term in the right hand side of (1), $\hat{\theta}_i$ is held fixed - this is quite intuitive since the variability of $\hat{\theta}_i$ has been already taken care of by the first term of (1). Note that as in (6.12) of Professor Rao's paper (see also Butar and Lahiri 1997) the term $\hat{E}_i$ is area-specific through $\hat{\theta}_i$. Thus, a jackknife estimator of $MSE[\tilde{\theta}_i(\hat{\theta}_i;\hat{\phi})]$ is $mse_i^{EBLUP} = mse_i^{BP} + \hat{E}_i$.

For a special case of model (6.4) with $\psi_i = \psi$ and $x_i'\beta = \mu$ $(i = 1,\cdots,m)$, a simplification of the expression for mse_i^{EBLUP} shows that, unlike Lahiri and Rao (1995), the formula involves both estimated skewness and kurtosis terms of the marginal distribution of $\hat{\theta}_i$. When normality is assumed in (6.4), the jackknife formula is asymptotically (in the second order sense) equivalent to that of Butar and Lahiri (1997) or (6.12) of Professor Rao's paper.

It is to be noted that the above jackknife method is different from the one proposed by Prasad and Rao (1986). Unlike our method, they deleted the "Henderson's residuals" to jackknife certain functions of the variance components in the context of nested error regression model. However, they did not consider jackknifying in prediction problems.

In many small-area applications, the posterior linearity assumption does not hold. For example, in the nested error regression model the distributions of the random effects and the errors may not induce the posterior linearity. The first step would then be to study the distribution of the random effects so that the BP can be obtained. The literature in this area of research is, however, not very rich. There are some work on informal check of normality of random effects (*e.g.*, Lange and Ryan 1989; Calvin and Sedransk 1991). Very recently, Jiang, Lahiri and Wu (1998) proposed a formal test for checking arbitrary distributional assumptions on the random effects and errors.

If the BP is nonlinear, how does one proceed to estimate the MSE of the corresponding EBP? As mentioned by Professor Rao, Jiang and Lahiri (1998) proposed a Taylor series method for a simple mixed logistic model. My recent work with Professor Jiang indicates that the Taylor series method works for generalized linear mixed models. Interestingly, the jackknife recipe described earlier can be extended to cover a very general class of models which include generalized linear mixed models and mixed linear models. Unlike the Taylor series method, the jackknife method does not require the tedious calculations of derivatives and so it should be appealing to the practitioners. For a recent theoretical development on jackknife in the prediction context and and the related variance component problems, the reader is referred to the very recent paper by Jiang, Lahiri and Wan (1998).

I totally agree with Professor Rao that his (6.6) is a more realistic sampling model. I will now show that it is feasible to propose a frequentist's alternative to the Bayesian solution. It can be shown that the BP of $\bar{Y}_i$ is given by

$$E(\bar{Y}_i|\hat{\bar{Y}}_i;\phi) \quad = \quad \frac{E[h(x_i'\beta+\sigma_v Z)u_i(Z;\hat{\bar{Y}}_i;\phi)]}{E[u_i(Z;\hat{\bar{Y}}_i;\phi)]}\,, \qquad \ldots(4)$$

where the expectations in the right hand side of (4) are taken over $Z \sim N(0,1)$. In the above, the function h is the inverse function of g, i.e, $h(x_i'\beta+\sigma_v Z) = \bar{Y}_i$ and $u_i(Z;\hat{\bar{Y}}_i;\phi) = exp[-\frac{1}{2\psi}\{\hat{\bar{Y}}_i - h(x_i'\beta+\sigma_v Z)\}^2],\ i = 1,\cdots,m$. Thus, to calculate the BP, one just need to calculate one-dimensional integrals. One can actually approximate the above expectations by a Monte Carlo method. To get EBP, ϕ needs to be replaced by a consistent estimator $\hat{\phi}$ (e.g., the method of simulated moments as described in Jiang and Lahiri 1998). The approaches of Jiang and Lahiri (1998) and Jiang, Lahiri and Wan (1998) can be pursued to get a measure of uncertainty of EBP.

Professor Rao discussed the issue of estimated ψ_i in model (6.4). This issue has generated some recent research activities. Since the ψ_i's contain certain design information (*e.g.,* sample size), it does not seem to be appropriate to model the ψ_i's directly. One way to handle this problem is to first assume that $\psi_i = \sigma_i^2 C_i$, where C_i contains certain design information and then model the σ_i^2's (assuming that some information is available on σ_i^2 from the sample survey). Arora and Lahiri (1997) considered a hierarchical Bayes estimation of consumer expenditures for small-areas using such a model. Incidentally, as a by product, their procedure also produces estimators which are design consistent. Bell (1995) used similar ideas in estimating poverty related statistics for small-areas. It would be interesting to develop frequentist's method for such random sampling variance models. Kleffe and Rao (1992) proposed a random sampling variance model without any specific distributional assumption on σ_i^2. They considered EBLUP for their point estimator and showed such a random sampling variance model affects the MSE estimator of EBLUP. It may be worthwhile to assume a specific distribution (*e.g.,* inverted gamma) for σ_i^2 in their model and develop a frequentist's method to address both point estimation and the MSE of the point estimator.

In many small-area applications, one has information from previous time points. Thus, one has basically three options: (i) use cross-sectional data only, (ii) use time series data only and (iii) use both the time series and cross-sectional data. The literature on (i) is vast. Recently Tiller (1992) considered approach (ii). Pfeffermann and Bluer (1992), Rao and Yu (1994), Ghosh *et al.* (1996), Datta *et al.* (1998), among others, considered approach (iii). One interesting issue here is how does one systematically determine the relative contributions from the time series and the cross-sectional parts? In my recent work with my colleagues (see Datta *et al.* 1998), we found out that it is impossible to go for approach (ii) because of a huge number of model parameters to be estimated and the approach (iii) really helps. Even when we eliminated some model parameters for comparison's sake, approach (iii) turns out to provide better results based on certain recently developed model diagnostic criteria for hierarchical models.

Acknowledgments. The research was supported, in part, by National Science Foundation Grant SBR-9705574.

References

Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica,* **7**, 1053-1063.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **80**, 28-36.

Bell, W.R. (1995). Bayesian sampling error modeling with application. *Proc. Sec. Survey Meth.,* Statistical Society of Canada annual Meeting.

Butar, F. and Lahiri, P. (1997). On the measures of uncertainty of empirical Bayes small-area estimators. *Tech. Report*, Div. of Stat., University of Nebraska-Lincoln.

Calvin, J.A. and Sedransk, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *J. Amer. Statist. Assoc.*, **86**, 36-48.

Chattopadhyay, M., Lahiri, P., Larsen, M., and Reimnitz, J. (1996). Composite estimation of drug prevalences for the substate areas. To appear in *Survey Methodology.*

Datta, G.S., Lahiri, P., Maiti, T. and K.L. Lu (1998). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *unpublished manuscript.*

Fay, R. E., and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269-277.

Jiang, J., and Lahiri, P. (1998). Empirical best prediction for small area inference with binary data. *Tech. Report*, Div. of Stat., University of Nebraska-Lincoln.

Jiang, J., Lahiri, P. and Wan, S. (1998). Jackknifing the mean squared error of empirical best predictor. *Tech. Report*, Div. of Stat., University of Nebraska-Lincoln.

Jiang, J., Lahiri, P. and Wu, C. (1998). On Pearson-χ^2 testing with unobservable cell frequencies and mixed model diagnostics. *Tech. Report*, Div. of Stat., University of Nebraska-Lincoln.

Ghosh, M. and Meeden, G. (1997). *Bayesian methods for Finite Population sampling.* Chapman and Hall.

Ghosh, M., Nangia, N., and Kim, D. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *J. Amer. Statist. Assoc.* , **91**, 1423-1431.

Kacker, R.N., and Harville, D.A. (1984). Approximations for the standard errors of estimation of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.*, **79**, 853-862.

Kleffe, J. and Rao, J.N.K. (1992). Estimation of mean square error of empirical best unbiased predictors under a random error variance. *J. Mult. Anal.*, **43**, 1-15.

Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *J. Amer. Statist. Assoc.* **90**, 758-766.

Lange, N., and Ryan, L. (1989). Assessing normality in random effects models. *Ann. Statist.*, **17**, 624-642.

Pfeffermann, D., and Bleuer (1992). Robust joint modeling of labour force series of small areas. *Survey Methodology*, **19**, 149-163.

Prasad, N.G.N., and Rao, J.N.K. (1986). Discussion on "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis," by C.F.J. Wu, *Ann. Statist.*, **14**, 1320-1322.

– – –– (1990). The estimation of mean squared errors of small area estimators. *J. Amer. Statist. Assoc.* **85**, 163-171.

– – –– (1998). On robust small area estimation using a simple random effects model. *Tech. Report*, Carleton University.

Rao, J.N.K., and Yu, M. (1994): Small area estimation by combining time series and cross-sectional data. *Can. J. Statist.*, **22**, 511-528.

Tiller, R. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *J. Off. Statist.*, **8**, 149-166.

Discussant : Danny Pfeffermann
Hebrew University

This excellent review focuses on the new developments in sample survey methodology over the last decade. In fact, out of the seventy-two references listed in the paper, only twenty-two are from years before 1989. One article I am missing though is the paper by Rao and Bellhouse (1990) which reviews the main developments in earlier decades. Taken together, these two articles provide a very comprehensive and thorough discussion of the foundations and developments in survey sampling methodology from its birth until our days.

In the concluding remarks of the present article, the discussants are asked to review and discuss the (few) topics not covered in the paper. At least in my case, I was asked explicitly to discuss also my own work, a request I find somewhat embarrassing and yet hard to refuse to. My discussion will focus on inference about super-population models which has implications to small area estimation, a topic considered at length in the paper.

Statistical models are often fitted to data collected from survey samples. Family expenditure surveys are used for the computation of income elasticities, health surveys are used extensively for analyzing the relationships between risk factors and health measures, whereas labor force surveys are used for studying labor market dynamics. In this kind of studies, the target of inference is the model parameters, or functions of these parameters, rather than the prediction of finite population means where the estimation of the model parameters is only an intermediate step. This change in the target parameters can affect the choice of weights and it opens the way for new inference methods.

A fundamental question, underlying the fitting of statistical models to survey data is whether weighting is at all needed. In Pfeffermann (1993), I argue (following earlier discussions) that a major reason for wanting to weight the sample observations is to protect against informative sampling. The sampling design is informative when the model holding for the sample data differs from the model holding in the population. This happens when the sample selection probabilities are related to the response variable even after conditioning on the model covariates. Ignoring the sample selection effects in such cases can bias the inference with the risk of reaching misleading conclusions.

For models that contain only fixed effects, weighting the sample observations by the standard sampling weights yields estimators that are design consistent for the population quantities under the randomization (repeated sampling) distribution. Since the latter are model consistent for the corresponding model parameters, the weighted estimators are likewise consistent for the model parameters under the mixed design and model distribution. For example, the weighted vector regression estimator, $b_w = (\sum_{i \in s} w_i x_i x_i')^{-1}(\sum_{i \in s} w_i x_i y_i)$ is under general conditions design consistent for the population vector , $B = (\sum_{i=1}^{N} x_i x_i')^{-1}(\sum_{i=1}^{N} x_i y_i)$ which in turn is consistent under the standard regression model for the model vector coefficients β. The convergence of the weighted estimators to their finite population counterparts is not dependent on model assumptions, implying some robustness against model misspecification. Specifically, the weighted estimators may have a meaningful interpretation and hence be useful for inference even if the working model fails to hold.

Standard weighting of the sample observations does not yield consistent estimators for the model parameters under the mixed linear model (6.2) of Rao's paper, which is in wide use for small area estimation. Pfeffermann et. al. (1998) propose a two-step weighting procedure that overcomes this problem. For the special case of non-informative sampling within the small areas, (but informative sampling of the small areas), the use of scaled weights provides consistent estimators for the model parameters even when the sample sizes within the small areas are bounded. Notice that under informative sampling of the small areas, the model holding for the sampled areas is different from the population model (6.2) (see below). Pfeffermann, Feder and Nathan (1998) apply these results to models that permit the random effects to evolve stochastically over time. Such models are useful for analyzing longitudinal data.

The consistency of the weighted estimators explains their broad use by data analysts but the utilization of the randomization distribution for inference has some major drawbacks discussed in Pfeffermann (1993). As a result, new approaches have been proposed to deal with these drawbacks. One of these approaches is to model the joint population distribution of the response variable and all the design variables employed for the sample selection (or adequate summaries of them), and then integrate over the response values of units outside the sample. The book edited by Skinner, Holt and Smith (1989) contains several examples. A major limitation of this approach, however, is that the population values of the design variables are often not available or that it is too complicated to model them. Chambers, Dorfman and Wang (1998) consider ways to deal with this problem.

Skinner (1994) proposes to extract the population model from models fitted to the sample data. The basic steps underlying this approach are summarized in the paper by Pfeffermann and Sverchkov (1999, hereafter PS) which appears in this volume. The appealing feature of this approach is that it permits the use of standard (efficient) model fitting procedures. An interesting question underlying its application is how to select sample models that yield an acceptable population model.

A different approach, described in Pfeffermann, Krieger and Rinott (1998, hereafter PKR), consists of extracting the sample probability density function (pdf) from the population pdf and the first order sample inclusion probabilities. For pairs of vector random variables (u, v), the sample pdf is defined as,

$$f_s(u_i|v_i) = f(u_i|v_i,\ i \in s) = \frac{E_p(\pi_i|u_i, v_i) f_p(u_i|v_i)}{E_p(\pi_i|v_i)} \qquad \ldots(1)$$

where the π_i's are the sample inclusion probabilities and the subscript p designates the population pdf. It follows from (1) that the sample and population distributions are different, unless $E_p(\pi_i|u_i, v_i) = E_p(\pi_i|v_i)$ for all i, in which case the sampling scheme is non-informative. Notice that the sample pdf is defined in conditional terms, which as discussed by Rao, is generally complicated in a randomization distribution framework. PKR establish asymptotic independence of the sample values with respect to the sample distribution for independent population measurements under commonly used schemes for single stage sampling with unequal probabilities. Hence, the sample pdf lends itself to the use of standard inference procedures such as maximum likelihood estimation or residual analysis applied to the sample measurements. It permits also the application of classical re-sampling methods such as the bootstrap with no further modifications, which is often problematic under the randomization distribution framework. Krieger and Pfeffermann (1997) use the sample pdf for testing hypotheses on the population pdf.

PS further explore the relationship between the sample and population distributions by comparing the corresponding moments. The basic equations are,

$$E_p(u_i|v_i) = E_s(w_i u_i|v_i)/E_s(w_i|v_i); \quad E_p(\pi_i|v_i) = 1/E_s(w_i|v_i) \qquad \ldots(2)$$

where E_p and E_s denote expectations under the population and sample pdf's and $w_i = 1/\pi_i$ are the sampling weights. Equation (2) and another relationship shown in PS induce the following new semi-parametric estimator for the regression vector coefficients under the classical regression model,

$$b_{sp} = (X'QX)^{-1}X'QY = (\sum_{i\in s} q_i x_i x_i')^{-1} \sum_{i\in s} q_i x_i y_i; \quad q_i = w_i/\hat{E}(w_i|x_i). \quad \ldots(3)$$

The difference between the estimator b_{sp} and the probability weighted estimator b_w defined before is in the weighting method. As noted by Chris Skinner (private communication), the weights q_i correct for the net sampling effects on the conditional distribution of $y|x$, whereas the weights w_i control for the sampling effects on the joint distribution of (y, x). The estimators $\hat{E}(w_i|x_i)$ can be obtained by regressing w against x; see PS for examples with simulated and real data. In these examples, the estimator b_{sp} and the maximum likelihood estimator obtained from (1), (assuming normality for the population residuals) outperform the estimator b_w decisively.

Can the weights q_i be utilized for estimating finite population means ? We can express $\bar{Y}$ as $E_p(Y_i)$ where the pdf "p" assigns the probability mass 1/N for every value Y_i. By (2), $E_p(Y_i) = E_s(w_i Y_i)/E_s(w_i)$ so that an application of the method of moments yields $\hat{\bar{Y}} = (\sum_{i\in s} w_i Y_i)/(\sum_{i\in s} w_i)$, the estimator proposed by Hajek in response to the circus dilemma. We may also write

$$\overline{Y} = E_p[E_p(Y_i|w_i)] = E_s[\frac{w_i}{E_s(w_i)} E_p(Y_i|w_i)] = E_s[\frac{w_i}{E_s(w_i)} E_s(Y_i|w_i)] \qquad \ldots (4)$$

where the second equality follows from a result in Skinner (1994). For $E_s(Y_i|w_i) = kw_i$, with $k = E_s(Y_i w_i)\big/E_s(w_i^2)$, an application of the method of moments yields again Hajek's estimator. For other relationships, the use of (4) yields different estimators. We are currently investigating ways of incorporating known values of auxiliary variables into the estimation process. Notice that the variances obtained from employing the sample distribution are different in principle from the randomization variances. As mentioned before, one feasible way to obtain variance estimators is by application of classical re-sampling procedures. See PS for illustrations.

Next consider the issue of small area estimation and suppose that the sample selection within the small areas is informative. For a model like in (6.1) and (6.3) of Rao's paper, replacing the unweighted area mean estimators by weighted estimators can control the sampling effects. This is the basic idea behind the estimators proposed by Kott (1989), and I presume Prasad and Rao (1998). When the target response variable is categorical however, such that the model has to be formulated at the unit level like, for example, in Ghosh et al. (1998, see paragraph below equation 6.2 of Rao's paper), there seems to be no obvious way to account for the sampling effects within the randomization distribution framework. The use of the sample distribution offers a principled solution to this problem by first extracting the distribution of the sample measurements utilizing (1), and then combining the resulting model with the model assumed for the true small area means, (say, the model defined by 6.1). This paradigm can be extended to cover situations where not all the small areas are represented in the sample and the selection of areas to the sample is informative. Say, areas with larger true means have higher probabilities of being selected. Work in these directions is in planning.

My final comment relates to the use of time series models for small area estimation. As is well known, many of the large surveys conducted in practice are repeated at fixed time intervals with possibly overlapping samples. For such designs, the direct small area estimators can be strengthened by borrowing information across time instead of, or in addition to borrowing information cross-sectionally. Models that account for the time series variation of the direct small area estimators typically combine a time series model for the true area means and another model that accounts for the correlation structure of the survey errors. In the U.S., all the major labor force statistics for all the states are produced from such models, see Tiller (1992). Pfeffermann, Feder and Signorelli (1998) propose a simple method for estimating the autocorrelations of the survey errors for panel surveys and use these estimates for fitting time series models in small regions of Sydney, Australia. This paper references other studies that employ time series models for small area estimation. Some of these models account also for the cross-sectional (spatial) dependencies between the area means. The use of time series models does not circumvent the problem discussed extensively by Rao of producing MSE estimates that account for parameter estimation. Quenneville and Singh (1998) handle this problem from a Bayesian perspective.

References

Chambers, R. L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Jour. Royal Statist. Soc.*, Series B, **60**, 397-411.

Krieger, A. M. and Preffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, **13**, 123-142.

Quenneville, B. and Singh, A. C. (1998). Bayesian prediction MSE for state space models with estimated parameters. Technical report, Business Survey Methods Division, Statistics Canada.

Preffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 317-337.

Preffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087-1114.

Preffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Jour. Bus. Econ. Statist.*, **16**, 339-348.

Preffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *Jour. Royal Statist. Soc.*, Series B, **60**, 23-56.

Preffermann, D., Feder, M. and Nathan, G. (1998). Time series multilevel modeling of longitudinal data from complex surveys. Paper presented at the annual meeting of the American Statistical Association, Dallas, Texas.

Preffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B*, (In this issue).

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

Skinner, C. J. (1994). Sample models and weights. In Proceedings of the Section on Survey Research Methods, American Statistical Association, 133-142.

Skinner, C. J., Holt, D. and Smith, T. M. F. (Eds.) (1989). *Analysis of Complex Surveys.* Wiley, New York.

Tiller, R. B. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, **8**, 149-166.

Rejoinder : J.N.K. Rao
Carteton University

I am thankful to the discussants of my paper for their insightful comments and for mentioning some important developments not covered in my paper. In particular, Dr. Pfeffermann gives a brief account of his joint work on informative sampling when the target of inference is the model parameters, Dr. Lahiri discusses his joint work on empirical best prediction (EBP) of small area means under generalized linear mixed models, Dr. Malay Ghosh proposes realistic models to handle small area estimation from stratified two stage samples when the areas cut across the primary units and stratum boundaries, and Dr. Chaudhuri refers to randomized response, network sampling and adaptive sampling.

As noted by Dr. Pfeffermann, my paper focussed on developments over the last decade or so. As a result, I did not attempt to cover much of the important work in earlier decades, and even failed to mention my own review paper, Rao and Bellhouse (1990)! I thank Dr. Pfeffermann for his kind remark: "Taken together, these two articles provide a very comprehensive and thorough discussion of the foundations and developments in survey sampling methodology from the birth until our days".

I will now try to respond to the discussant's comments topic-wise: survey design, data collection and processing, inferential issues, resampling methods and small area estimation.

Survey design. I agree with Dr. Eltinge that the case-specific total survey design approach of Linacre and Trewin (1993) sheds relatively little light on the design of other surveys, especially if the survey to be designed is not a revision of an ongoing series. But case-specific approaches often can be implemented and can lead to substantial improvements in survey efficiency, as demonstrated by Linacre and Trewin. Dr. Eltinge proposes to model a transformation of total MSE, such as a generalization of Kish design effect, as a function of various factors including sampling errors and components thereof, nonsampling errors and efforts to reduce the effect of nonsampling error, and specific estimation and inference methods used. This ambitious proposal should be the goal of total survey design and Dr. Eltinge's suggestions for implementing such an approach, even partially, are very useful. Many agencies find comprehensive quality guidelines helpful in implementing a total survey design approach; for example, Statistics Canada (1998) Quality Guidelines. The meta analysis approach of Singer *et al.* (1998) for combining the results of multiple studies of empirical effects also looks promising.

Dr. Fay is correct in saying that many evaluation studies for measuring different error sources are subject to their own biases, for example, reinterviews. I agree with his comment "This area is one that will require attention to the survey methodology and other aspects of the empirical science, as well as to the mathematical framework for the effort". Dr. Malay Ghosh gives an interesting application of total survey error.

Data collection and processing. Dr. Chaudhuri notes that randomized response (RR) technique can be used to handle sensitive questions. Cross-over designs for ordering questions, mentioned in my paper, are, however, meant for less sensitive questions than those used with random response designs. In this connection, finding appropriate measures of privacy for RR is important. Ljungquist (1993) shows that most measures of privacy proposed in the literature are inconsistent with the underlying rationale for the randomized response procedure.

Network sampling is a fairly established method for sampling rare populations (Sirken, 1970). Adaptive sampling is of more recent vintage and, as noted by Chaudhuri, it can be useful when serviceable frames are not available. Thompson's (1992) book gives a lucid account of many nonconventional sampling methods, including adaptive sampling, line-intercept and capture-recapture sampling which are especially useful for sampling ecological populations.

Inferential issues. Dr. J.K. Ghosh notes that modelling has been successful for small area estimation but failed at higher level of aggregation. For small area estimation, models are necessary because of very small samples or even no samples in some small areas. Similarly, models are necessary to handle response errors and nonresponse. At an aggregate level, sample sizes are sufficiently large to make the direct design-based estimators reliable in terms of sampling error. The latter estimators can be model-assisted but remain design-consistent to avoid difficulties with model-dependent estimators for large samples under model misspecification. Model-based estimators of small area totals are in fact adjusted to add up to direct estimators at aggregate levels, as noted in Section 6 of my paper.

I agree with Dr. J.K. Ghosh that Basu's circus elephants example is most compelling. In fact, it was instrumental in putting a stop to publication of papers (in prestigious journals) claiming optimality of the NHT-estimator for any design. But the example does not destroy frequentist sampling theory, as demonstrated in subsection 4.1. Dr. Ghosh is also correct in saying that Basu is at least partially responsible for the acceptance of conditionality principle in survey sampling. In fact, for simple random sampling with replacement Basu (1958) suggested that for the purpose of inference one should condition on the number of distinct units in the sample.

Dr. J.K. Ghosh suggests that for model-based inference much more complex models than the simple linear model are needed when the sample size is large. But I am doubtful if this approach would resolve the difficulties with the model approach pointed out by Hansen *et al.* (1983). In the Hansen example, model deviations are not detectable for samples as large as 400 and yet a large model bias is induced through their stratified design with disproportionate sample allocation. If the design is changed to simple random sampling, then the model bias is small under the same model deviations. The model bias is given by $\alpha(\bar{X}/\bar{x}-1)$, where α is the intercept therm. $\bar{x}$ is the unweighted sample mean and $\bar{X}$ is the known population mean. It is clear from this expression that model bias is small under simple random sampling because $\bar{x} \approx \bar{X}$ for large samples. On the other hand, the model bias can be substantial under disproportionate sampling even for small α because $\bar{x}$ deviates substantially from $\bar{X}$ for large samples. Chambers *et al.* (1993) proposed nonparametric regression to define a bias-adjusted version of the model-based estimator. It would be interesting to investigate if the proposed method can lead to valid inferences for large samples under the Hansen *et al.* set up.

Applications of NHT-type estimators in other areas, mentioned by Dr. J.K. Ghosh, look interesting. He notes that some of the asymptotics for NHT-estimators has also been useful in forecasting discovery of oil reserves. In this connection, it is interesting to note that Andreatta and Kaufman (1986) developed asymptotic expansions of Murthy's (1957) estimator to make inference on the size distribution of oil and gas fields in a petroleum basin.

Dr. Malay Ghosh gives a nice justification of the generalized regression estimator, $\hat{Y}_{gr}$, via the estimating functions approach of Godambe. But I do not agree with him that $\hat{Y}_{gr}$ is optimal or near optimal under the assumed model. To see this, consider the model $E(y_j)\mu$ with uncorrelated errors $e_j y_j - \mu$ and $E_\xi(e_j^2)\sigma^2$. This model is relevant under noninformative sampling with inclusion probabilities π_j unrelated to y_j (Rao, 1966). Under this model, the model-assisted estimator is the Hajek estimator $\hat{Y}_{gr} N \Sigma_s \tilde{w}_j y_j$ with $\tilde{w}_j w_j/\Sigma_s w_j$ and $w_j \pi_j^{-1}$. On the other hand, the best linear unbiased predictor is $\hat{Y}_m N\bar{y}$, where $\bar{y}$ is the sample mean. It is easy to see that $E_\xi(\hat{Y}_{gr}-Y)^2 - E_\xi(\hat{Y}_m - Y)^2 N^2 \Sigma_s(\tilde{w}_j - \frac{1}{n})^2$ which can be substantial when the normalized weights $\tilde{w}_j$ vary considerably. Thus $\hat{Y}_{gr}$ can be considerably less efficient than $\hat{Y}_m$ under the assumed model. In fact, my 1966 Sankhya paper proposed $\hat{Y}_m$ for design-based inference when y_j and π_j are nearly unrelated because its bias is small and it is considerably more efficient than the NHT-estimator in such cases.

I agree with Dr. Eltinge that it would be useful to examine potential trade-offs between robustness and efficiency for competing methods. I will look forward to details of his work. Dr. Fay also mentions similar trade-offs, in particular variance vs. bias, in the context of direct variance estimates vs. those derived from the method of generalized variance functions that relates the variance of a survey estimator to the expectation of the estimator by fitting a model to several survey estimates and corresponding variance estimates.

Resampling methods. Dr. Fay notes that replication methods such as the jackknife also provide estimates of conditional variance.I have found this to be true in the context of optimal regression estimator (4.7) which is conditionally valid. Similarly, linearizing the jackknife variance estimator of the ratio estimator $(\bar{y}/\bar{x})\bar{X}$ in simple random sampling, we get a variance estimator approximately equal to the model-based variance estimator of Royall and Eberhardt (1975). Thus, Fay is correct in saying that "practitioners may often not distinguish between unconditional and conditional inference" by using a resampling method. But the estimator itself should be conditionally valid under the conditional inference framework.

Dr. Fay mentions that hybrid strategies which use some replication methods for some sampling strata and others for other strata may be useful in some practical applications. It would be useful to study the asymptotic properties of resulting variance estimates. I also agree with him that resampling methods handle variance estimation with imputed survey data better than multiple imputation can under a design-based framework.

Dr. Eltinge notes that superpopulation inference can be viewed in the framework of two-phase sampling, where the first phase sampling generates a finite population through a stratified and clustered superpopulation model. Also, the stratification and clustering imposed by the sample design may cut across the first phase stratum and cluster boundaries. But, validating such superpopulation models can be difficult in practice, as noted in my paper. Point estimation of superpopulation parameters can be accomplished by assuming only correct mean specification and that the finite population can be regarded as a self-weighting sample from the superpopulation, not necessarily a simple random sample (Rao, Scott and Skinner, 1998). But standard error estimation and hypothesis testing can be problematic without actually specifying the superpopulation model structure.

Small area estimation. Dr. Lahiri's recent work on jackknife estimation of MSE of a linear empirical Bayes (LEB) estimator is useful because it leads to second-order valid estimators of MSE without making normality assumption on the random small area effects v_i in the area level model (6.4). Lahiri and Rao (1995) considered EBLUP estimator which does not require normality assumption on the v_i's and established the robustness of normality-based MSE estimator of Prasad-Rao. Their result does not depend on the posterior linearity assumption and EBLUP is optimal in the class of linear unbiased predictors. However, the result of Lahiri and Rao does not extend to nested error regression model (6.2) and other models, as noted by Dr. Lahiri. For the latter models, the jackknife method of Dr. Lahiri looks promising, provided posterior linearity assumption can be validated from the data.

Model diagnostics play an important role in small area estimation, but the literature on model diagnostics for random effect models is somewhat limited, unlike the extensive literature on fixed effect models. New methods in this area are therefore welcome and I am glad to read Dr. Lahiri's recent paper (Jiang *et al.*, 1998) which proposes a formal test for checking arbitrary distributional assumptions on the random effects.

Dr. Lahiri's recent paper (Jiang and Lahiri, 1998) on estimating the MSE of the empirical best predictor (EBP) under generalized linear mixed models, using the jackknife, is also useful and broadens the scope of empirical Bayes methods. He proposes a similar frequentist solution to handle models (6.1) and (6.6) which cannot be combined to produce a linear mixed model when $\theta_i g(\bar{Y}_i)$ is nonlinear. I find this proposal very useful.

The jackknife method of Jiang and Lahiri works for models with block diagonal covariance structure but may run into difficulties with more complex models such as the model proposed by Dr. Malay Ghosh.

Dr. Chaudhuri is asking if one should use a frequentist measure of error for the EBLUP. Such measures can be very unstable compared to the model-based MSE estimator of Prasad and Rao, as shown in Hwang and Rao (1987). Dr. Chaudhuri is also asking whether it makes sense to start with a GREG estimator and improve upon it by an EB estimator, following Fay and Herriot (1979). This can run into difficulty with MSE estimation if one ignores the variability of the estimated variance of GREG; note that Fay-Herriot assume that the sampling variance of the direct estimator is known or obtained through a generalized variance function approach.

Dr. Chaudhuri also asks if Kalman filtering methods can be used in small area estimation. Pfeffermann and Burck (1990) have used such methods for small area estimation combining time series and cross-sectional data. But these methods assume either independence of sampling errors over time or specific correlation structures. Also, it is difficult to produce MSE estimators that captures all sources of variability. Dr. Lahiri correctly notes that one should use both time series and cross-sectional data because the use of time series data only may not be feasible in the presence of a large number of model parameters.

References

Andreatta, G. and Kaufmann, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *J. Amer. Statist. Assoc.*, **81**, 657-666.

Basu, D. (1958). On sampling with and without replacement. *Sankhyā*, **20**, 287-294.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer. Statist. Assoc.*, **88**, 268-277.

Hwang, J. and Rao, J.N.K. (1987). Unpublished research.

Ljungquist, L. (1993). A unified approach to measures of privacy in randomized response models: a utilitarian perspective. *J. Amer. Statist. Assoc.*, **88**, 97-103.

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

Royall, R.M. and Eberhardt, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, *Ser. C*, **37**, 43-52.

Sirken, M.G. (1970). Household surveys with multiplicity. *J. Amer. Statist. Assoc.*, **65**, 257-266.

Statistics Canada (1998). Statistics Canada Quality Guidelines. *Statistics Canada*, Ottawa, Canada.