Statistical Science
2003, Vol. 18, No. 2, 199–210

On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation

P. Lahiri

Abstract. Development of valid bootstrap procedures has been a challenging problem for survey samplers for the last two decades. This is due to the fact that in surveys we constantly face various complex issues such as complex correlation structure induced by the survey design, weighting, imputation, small-area estimation, among others. In this paper, we critically review various bootstrap methods developed to deal with these challenging issues. We discuss two applications where the bootstrap has been found to be effective.

Key words and phrases: Imputation, resampling, small-area estimation, survey weights.

1. INTRODUCTION

Efron (1979) proposed his bootstrap method to study properties of various nonlinear smooth and nonsmooth statistics. The method involves generation of a large number of independent *resamples* or *bootstrap samples*, each drawn from the original sample *with replacement*. For each such resample, the nonlinear statistic of interest is calculated and the values of the statistic for these resamples form the basis of inference. The properties of the bootstrap method for both smooth and nonsmooth statistics have been extensively studied for the i.i.d. case. The bootstrap is a computer-intensive method. With the advent of modern computer technology, however, the method offers a convenient framework for analyzing data effectively, especially for complex problems where an analytical solution is either nonexistent or cumbersome to apply.

This paper provides a critical review of various modifications of Efron's original bootstrap to handle complex issues in survey sampling. We note that other resampling techniques such as the the jackknife and balanced repeated replication (BRR) have been used by survey samplers. The bootstrap, however, is probably the most flexible and efficient method of analyzing survey data since it can be used to solve a variety of challenging statistical problems (e.g., variance estimation, imputation, small-area estimation, etc.) for complex surveys involving both smooth and nonsmooth statistics. Readers interested in other resampling methods in survey sampling are referred to an excellent review paper by Rust and Rao (1996).

Various federal and private agencies routinely conduct large-scale sample surveys. In a typical sample survey, a suitable probability sampling scheme is employed to collect data from a finite survey population. Data collection usually involves stratification of the survey population and selection of ultimate sampling units in several stages. The sample selection process invariably introduces a complex correlation structure which makes the development of a theoretically valid bootstrap method challenging. A survey weight, representing a certain number of units in the finite population, is usually attached to each unit in the sample to account for various factors such as the unequal probability of selection, nonresponse, poststratification and calibration. The incorporation of these important survey weights in the bootstrap method has received considerable attention over the last decade. The survey-weighted statistic of interest yields unbiased or nearly unbiased estimators of the corresponding finite population parameter from the random-

P. Lahiri is Professor of Survey Methodology, Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, Maryland 20855 (e-mail: plahiri@survey.umd.edu).

ization theory perspective where the observations are treated as fixed and the random mechanism generating the survey data forms the basis for inference (see Cochran, 1977). For sample surveys, bootstrap methods have been traditionally validated under randomization theory.

In sample surveys, missing data are plentiful. When a sampling unit does not respond to any of the survey questions, the problem of unit nonresponse arises. It has been a common practice to handle such unit nonresponse by first forming suitable *weighting classes* using auxiliary variables observed on all sampling units and then suitably adjusting survey weights for all respondents. A nonresponse adjustment factor is the same within each weighting class but different for different classes (see Kalton and Kasprzyk, 1986). For a method of forming weighting classes using a logistic or probit regression model, see Little (1986). In case a sampling unit responds to some but not all survey questions, we encounter the problem of item nonresponse. Various deterministic (e.g., mean, ratio, regression, etc.) and random (e.g., hot deck) imputation methods exist to handle item nonresponse. See Shao and Sitter (1996). Hansen, Hurwitz and Madow (1953) recognized the danger of using standard survey sampling methods which treat imputed values as if they were true values. Imputation introduces errors which must be accounted for. Section 3 reviews the literature on bootstrap methods which account for imputation errors in analyzing complex survey data.

Large-scale sample surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas. However, for effective planning of health, social and other services, and for apportioning government funds, there is a growing demand to produce similar estimates for smaller geographic areas and subpopulations, called small areas, for which adequate samples are not available. The usual design-based small-area estimators are unreliable since they are based on a very few observations that are available from the area. In the absence of a reliable small-area design-based estimator, we may use either a frequentist predictor (e.g., empirical best linear unbiased predictor, or EBLUP) or a hierarchical Bayes estimator. These methods essentially use a suitable multi-level mixed model or hierarchical Bayes model which captures various salient features of the sampling design and combines information from censuses or administrative records in conjunction with the survey data. For a review of small-area estimation, see Ghosh and Rao (1994), Rao (1999), Lahiri and Meza (2002) and Pfeffermann (2002), among others. Section 4 reviews the parametric bootstrap methods which account for various sources of uncertainties for EBLUP [same as empirical Bayes (EB)]. The method emerges as a very flexible and general method in covering a wide variety of small-area models.

Finally, in Section 5, we briefly discuss two real-life applications of bootstrap methodology from the U.S. Department of Education and the U.S. Bureau of Labor Statistics.

2. BOOTSTRAP METHODS FOR COMPLEX SURVEYS

In this section, we provide a history of the evolution of various bootstrap methods for complex surveys (anything other than simple random sampling with replacement). In the process, we attempt to bring out similarities and differences among various bootstrap methods proposed in the literature. Bootstrap methods in survey sampling are usually justified from the randomization approach to survey sampling which may be described as follows.

Suppose we have a finite population of N $(< \infty)$ elements labeled as $\{1, \ldots, N\}$. Let y_i be the value of a characteristic (or possibly a vector of characteristics) for the ith unit of the finite population $(i = 1, \ldots, N)$. In general, we are interested in a variety of nonlinear functions of y_i $(i = 1, \ldots, N)$. Here are a couple of examples:

(i) A smooth function of finite population mean, say $\theta = g(\bar{Y})$, where $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$, the finite population mean.

(ii) A smooth function of a vector of finite population means, say $\theta = g(\bar{Y}_1, \ldots, \bar{Y}_p)^T$, where $\bar{Y}_j$ denotes the jth finite population mean $(j = 1, \ldots, p)$. This case covers finite population variance $(\sigma^2 = \bar{Y}_1 - N\bar{Y}_2^2$, where $\bar{Y}_1 = N^{-1} \sum_{i=1}^{N} y_i^2$ and $\bar{Y}_2 = \bar{Y})$, ratio of finite population means $(R = \bar{Y}_1/\bar{Y}_2$, where $\bar{Y}_1 = N^{-1} \sum_{i=1}^{N} y_i$ and $\bar{Y}_2 = N^{-1} \sum_{i=1}^{N} x_i)$ and correlation between two characteristics $[\rho = (\bar{Y}_5 - N\bar{Y}_1\bar{Y}_2)/\sqrt{(\bar{Y}_3 - N\bar{Y}_1^2)(\bar{Y}_4 - N\bar{Y}_2^2)}$, where $\bar{Y}_3 = N^{-1} \sum_{i=1}^{N} y_i^2$, $\bar{Y}_4 = N^{-1} \sum_{i=1}^{N} x_i^2$ and $\bar{Y}_5 = N^{-1} \sum_{i=1}^{N} x_i y_i]$.

(iii) Nonsmooth functions (e.g., a quantile) of a finite population.

A sample of size n is drawn using a probability sampling design. Let $p(s)$ denote the probability of drawing the particular sample s out of all possible samples S. Thus, $p(s) \geq 0$ and $\sum_{s \in S} p(s) = 1$. In the traditional randomization theory in survey sampling, y_i $(i = 1, \ldots, N)$ are treated as fixed and all the infer-

ences are made with respect to $p(s)$, that is, the random mechanism which generates data from the finite population. Design unbiasedness [i.e., unbiasedness with respect to the sampling design $p(s)$] has played an important role in survey sampling. For nonlinear cases as in (i) and (ii) above, a standard estimator of θ is $\hat{\theta} = g(\hat{\bar{Y}}_1, \ldots, \hat{\bar{Y}}_p)$, where $\hat{\bar{Y}}_j$ is an unbiased or nearly unbiased estimator of $\bar{Y}_j$ $(j = 1, \ldots, p)$ under $p(s)$. In the case of (iii), one can use the quantile of an unbiased or nearly unbiased estimator (under the sampling design) of the finite population distribution function.

As is apparent from the discussion in the previous paragraph, in survey sampling the evaluation of the bias of an estimator is important. Also important is the production of standard error estimates and the construction of confidence intervals. For the linear case, all these do not pose any problem and exact calculations are possible using the sampling design (see, e.g., Cochran, 1977). For the nonlinear case, however, exact calculation is either difficult or impossible and the linearization method (same as the Taylor series) has been used. However, this method is very cumbersome especially for complex sampling designs and different variance formulas are needed for different estimators.

As an alternative to the linearization method, various resampling methods have been developed for complex surveys. Compared to the linearization method, these methods are generally computer intensive but are easy to implement. The bootstrap has been found to be a very effective and versatile resampling method which works in all practical situations for smooth and nonsmooth functions involving complex surveys from finite populations. The bootstrap methods discussed in this section involve the following common steps:

1. Using a suitable probability sampling scheme, generate a resample (or bootstrap sample) from the original sample.
2. Calculate the nonlinear statistic, that is, $\hat{\theta}$, using the resample or a suitable rescaled version of the resample. Denote it by $\hat{\theta}^*$.
3. Calculate $\hat{\theta}$ for a large number (say, B) of independent resamples. Let $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ be estimates from the B independent resamples. These $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ form the basis of inference for θ. For example, the bias and the variance of $\hat{\theta}$ are estimated by

$$\text{bias}_b = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_\cdot^*),$$

$$v_b = (B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_\cdot^*)^2,$$

where $\hat{\theta}_\cdot^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$. For confidence intervals, we can use the percentile method or bootstrap-t confidence intervals (see Rao, Wu and Yue, 1992). Note that the bootstrap method, like the other resampling methods, requires just one standard formula which works for any statistics.

The finiteness of the survey population, the complexity of the survey design and the complex weighting scheme all contribute to the challenging task of finding a valid bootstrap procedure. There are two basic criteria for a good bootstrap procedure. First, the bootstrap bias estimate should be 0 for the customary unbiased estimator in the linear case. Also, in this case, the bootstrap method should match the customary unbiased variance estimator. In addition, a good bootstrap method should produce a consistent variance estimator for nonlinear statistics. We shall now discuss available bootstrap methods in the following sections.

2.1 Stratified Simple Random Sampling

Suppose we have a finite population partitioned into H strata where each stratum contains similar units and simple random sampling without replacement (SRSWOR) is used to select a number of ultimate units within each stratum. Sampling is carried out independently for different strata. For simplicity in notation, we shall consider bootstrapping for the case of a single stratum, that is, $H = 1$, since the same bootstrap method needs to be carried out independently for different strata.

It is well known that $\bar{y} = n^{-1} \sum_{i \in s} y_i$, the usual sample mean, is unbiased for the finite population mean $\bar{Y}$ under SRSWOR and the design variance of $\bar{y}$ is given by

$$V(\bar{y}) = (1-f)\frac{S^2}{n},$$

where $f = n/N$ is the finite population correction (f.p.c.) factor. The f.p.c. is an important factor in finite population sampling which makes $v(\bar{y}) = 0$ when $n = N$. When $f \approx 0$, simple random sampling with replacement (SRSWR) is a good approximation to SRSWOR. The customary unbiased estimator of $V(\bar{y})$ is given by

$$v(\bar{y}) = (1-f)\frac{s^2}{n},$$

where $s^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y})^2$ is the usual sample variance.

The usual bootstrap, as described in Efron (1979), entails taking a SRSWR sample of size n from the original sample $y_i, i \in s$. Let $\bar{y}^*$ denote the mean based on the bootstrap sample. Then the bootstrap variance estimator of $V(\bar{y})$ is given by

$$V_*(\bar{y}^\star) = \frac{n-1}{n}\frac{s^2}{n},$$

which does not yield $v(\bar{y})$. Define the average relative error (ARE) as

$$\text{ARE} = \frac{[E(V_*(\bar{y}) - V(\bar{y})]}{V(\bar{y})},$$

where E and V are with respect to SRSWOR. It is interesting to note that even for SRSWR from the finite population (the most favorable case for the bootstrap) the bootstrap method underestimates, the ARE being $-n^{-1}$ (e.g., 20% underestimation for $n = 5$). In a hypothetical population setup, Efron (1982) recognized the problem and suggested a simple solution of taking a bootstrap sample of size $n - 1$ instead of n. Needless to say, this works for finite population sampling as well when f is negligible. When f is not negligible, the bootstrap method, even with Efron's correction, overestimates since it fails to recover the f.p.c., resulting in a nonzero variance estimate even for the extreme case of the census when $n = N$.

The important problem of recovering the f.p.c. resulted in mainly three different approaches in the literature: (i) without-replacement bootstrap (BWO), (ii) with-replacement bootstrap (BWR) and (iii) a hybrid of the BWR and BWO bootstraps. The BWR approach attempts to adapt Efron's original bootstrap by carefully choosing the bootstrap sample size or by rescaling the generated bootstrap sample observations. On the other hand, BWO tries to mimic the original SRSWOR sampling design.

We shall now consider the following three cases which will bring out various issues in the application of the bootstrap for SRSWOR sampling.

2.1.1 *BWO Methods.* Obviously, drawing a bootstrap sample of size n without replacement does not make sense since it provides the original sample. One might think of a naive bootstrap without-replacement sample of size $n' = fn$ in order to capture the f.p.c. But this bootstrap sampling yields

$$V_*(\bar{y}^*) = (1-f)\frac{s^2}{n'},$$

resulting in an ARE of $f^{-1} - 1$ and thus overestimating the true variance. To reduce the amount of overestimation, one may consider the method proposed by Gross (1980). McCarthy and Snowden (1985) called this the BWO method. The method is as follows: if $N = kn$, create an artificial population of N units simply by copying each of the n elements in the original sample k times and then take a SRSWOR bootstrap sample of size n from this artificial population. The method produces the following variance estimator:

$$(1-f)\frac{s^2}{n}\frac{n-1}{n-f},$$

resulting in an ARE of $(n-1)/(n-f) - 1$. Thus, unlike the naive method, it suffers from an underestimation problem. But, asympototically, when both n and N increase with f fixed, the method provides a consistent variance estimator. However, for stratified sampling when the sample size within each stratum is bounded and the number of strata is large (a situation commonly encountered in small-area estimation; see Section 4), the method could yield an inconsistent variance estimator (see Bickel and Freedman, 1984). Furthermore, in many practical situations, k is not an integer. When $N = kn + r$ with $1 \le r \le n-1$, Bickel and Freedman (1984) proposed the following correction to Gross' method.

Construct two artificial populations of sizes kn and $(k+1)n$ by copying each sample element k and $k+1$ times, respectively. One of the two artificial populations is then selected using a randomization mechanism which assigns a probability α to population 1 and $1-\alpha$ to population 2. From the selected artificial population, draw a SRSWOR bootstrap sample of size n and then apply the bootstrap variance formula of Gross (1980) given above. The parameter α is selected so that the bootstrap variance estimator equals $v(\bar{y})$. McCarthy and Snowden (1985) discussed some examples when the bootstrap method of Bickel and Freedman (1984) is not feasible. In fact, it is not feasible when $n^3 < N^2$.

Sitter (1992b) proposed an alternative to randomization of the two artificial populations in order to obtain the customary variance estimator. He suggested choosing a bootstrap sample size n' and k so as to capture the f.p.c. and match the bootstrap variance estimator with the customary variance estimator. The method, however, produces noninteger n' and k and requires appropriate randomization between the bracketing integers.

2.1.2 *BWR Methods.* McCarthy and Snowden (1985) noted that it is not essential to mimic the original sampling design to capture the f.p.c. One can simply adapt Efron's original bootstrap by taking a larger bootstrap sample than is required if the original sampling plan were SRSWR. They suggested a sample of size $n' = (1-f)^{-1}(n-1)$ which yields the customary variance estimator $v(\bar{y})$. One problem with this method is that n' could be noninteger and some randomization is needed in most practical situations.

Rao and Wu (1988) proposed a BWR method which rescales the bootstrap sample so as to recover the f.p.c. in the usual SRSWOR variance formula. This procedure selects m BWR sample, say $y^\star = (y_1^\star, \ldots, y_m^\star)^T$, from the original sample and then rescales the bootstrap sample by

$$\tilde{y}_i = \bar{y} + m^{1/2}(n-1)^{-1/2}(1-f)^{1/2}(y_i^\star - \bar{y}).$$

The bootstrap variance estimator reduces to the customary variance estimator in the linear case for any choice of m. It is also possible to match the third moment by choosing m appropriately. This method avoids the problem associated with noninteger sample size but improper choice of m could lead to negative values of $\hat{\theta}$ even when θ is positive, as noted by Rao and Wu (1988). For stratified sampling with replacement, Rao and Wu (1988) proved the consistency of their rescaling bootstrap variance estimator for a smooth function of means. In this case, m can be chosen to match the third moment and it turns out the same choice ensures that the bootstrap histogram of a t statistic captures the second-order term of the Edgeworth expansion in the special case of known population strata variances. However, it is not yet known if a similar property will hold for more complex sampling designs and for nonsmooth functions. For simulation results on the bootstrap and other rival methods for stratified sampling, see Rao and Wu (1988), Kovar, Rao and Wu (1988), Sitter (1992a, b) and Rao, Wu and Yue (1992).

2.1.3 *A Hybrid of BWO and BWR Methods.* Sitter (1992a) proposed a mirror-match bootstrap method for a variety of complex survey designs. This method can be viewed as a combination of BWO and BWR. Sitter's method selects the bootstrap sample in two steps:

1. Draw a SRSWOR of size $n' < n$.
2. Generate the bootstrap sample of size $n^\star$ by drawing $k = (f^\star f)^{-1}(1-f^\star)$ independent SRSWR bootstrap samples, each of size $n^\star = n(1-f^\star)\cdot(1-f)^{-1}$, from the SRSWOR sample obtained in 1.

Note that the choice of $n' = 1$ in Sitter's method results in the BWR originally proposed by McCarthy and Snowden (1985). If k is not an integer, Sitter (1992a) suggested a randomization between the bracketing integers (note that $n^\star$ is an integer if k is). If $f \geq n^{-1}$, the choice of n' ensures the same f.p.c. as the original sampling design. For stratified sampling, Sitter (1992a) showed that, under suitable regularity conditions and a flexible asymptotic setup, his method yields a consistent variance estimator for smooth nonlinear statistics and for $\theta = \bar{y}$ the bootstrap histogram matches the second-order term of the Edgeworth expansion for an appropriate choice of $n^\star$, $f^\star = f$ within each stratum.

2.2 Stratified Multistage Sampling

We shall first consider two-stage sampling without stratification. In a two-stage sampling design, first a SRSWOR of n primary stage units (p.s.u.'s) is drawn from N primary stage units (p.s.u.'s) in the population, and then from each selected p.s.u. a SRSWOR of m_i secondary stage units (s.s.u.'s) is drawn from M_i s.s.u.'s in the ith p.s.u. ($i = 1, \ldots, n$). For this design, the bootstrap methods discussed in the previous section do not work. To see this, consider a special balanced case when $M_i = M$ ($i = 1, \ldots, N$) and $m_i = m$, $i \in s$. Consider the estimation of the population mean

$$\bar{Y} = (MN)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = N^{-1}\sum_{i=1}^{N}\bar{Y}_i,$$

where $\bar{Y}_i = M^{-1}\sum_{j=1}^{M} y_{ij}$ denotes the population mean for the ith p.s.u. ($i = 1, \ldots, N$). An unbiased estimator of the population mean is given by $\bar{y} = n^{-1}\sum_{i\in s}^{n}\bar{y}_i$, where $\bar{y}_i$ is the sample mean for the ith selected p.s.u. ($i = 1, \ldots, n$). Note that under two-stage sampling

$$V(\bar{y}) = \frac{(1-f_1)}{n}S_B^2 + \frac{1-f_2}{mn}S_W^2,$$

where $f_1 = n/N$, the f.p.c. for first-stage sampling; $f_2 = m/M$, the f.p.c. for second-stage sampling within the selected p.s.u.'s; $S_B^2 = (N-1)^{-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2$, the population variance of p.s.u. means; and $S_W^2 = \sum_{i=N}^{M}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2/N(M-1)$, the population variance among the elements within the p.s.u.'s. The customary unbiased estimator, say $v(\bar{y})$, of $V(\bar{y})$ is

given by

$$v(\bar{y}) = \frac{1-f_1}{n}s_B^2 + \frac{f_1(1-f_2)}{mn}s_W^2,$$

where $s_B^2 = (n-1)^{-1}\sum_{i\in s}(\bar{y}_i - \bar{y})^2$ and $s_W^2 = \sum_{i\in s}\sum_{j\in s_i}(y_{ij}-\bar{y}_i)^2/n(m-1)$. All these results are given in Cochran (1977, page 277). The bootstrap method (after the correction described in the previous section) is simply $v_b = (1-f_1f_2)s^2/mn$, where $s^2 = \sum_{(i,j)\in s}(y_{ij}-\bar{y})^2/(nm-1)$. It can be seen that

$$E(v_b) = (1-f_1f_2)\left[\frac{n-1}{nm-1}\frac{S_B^2}{n} + \frac{S_W^2}{mn}\right].$$

Thus, the bootstrap method does not produce the customary unbiased estimator $v(\bar{y})$. For the case when $f_1 = f_2 = 0$, we get

$$\text{ARE} = -\frac{n(m-1)}{nm-1}\frac{1}{1+R/m},$$

where $R = S_W^2/S_B^2$. This implies that the bootstrap methods discussed earlier underestimate the true variance and the amount of underestimation is a decreasing function of R and n. For the unbalanced case, the underestimation is expected to be more serious.

The bootstrap methods described in the previous section can be adapted to two-stage sampling. A common feature of all these methods is the drawing of bootstrap samples in two stages. In the case of the BWO method, one needs to find n' and k for bootstrap sampling of the p.s.u.'s and s.s.u.'s within the selected bootstrap p.s.u.'s with randomization to take care of noninteger sample sizes. See Sitter (1992b). In the case of the BWR rescaling method of Rao and Wu (1988), one needs to rescale at both stages of sampling. As Sitter (1992a) says, it is difficult to compare two methods in two-stage sampling either analytically or via simulation because both involve complex algorithms. The bootstrap methods for two-stage sampling meet the two basic properties. The methods can be readily extended to stratified two-stage sampling by simply drawing bootstrap samples independently for different strata.

Now, consider the general case of stratified multistage sampling. As before, since the same bootstrap sampling is performed for each stratum we consider the case of a single stratum. Suppose n p.s.u.'s are sampled and within the ith selected p.s.u. subsampling is performed in multiple stages to select n_i ultimate sampling units ($i = 1, \ldots, n$). Let y_{ij} and w_{ij} denote the observation and the sampling weight associated with the (i, j)th unit sampled ($i = 1, \ldots, n;\ j = 1, \ldots, n_i$).

Ignoring the complicated weighting adjustment factors in the construction of sampling weights and the f.p.c., Rao, Wu and Yue (1992) suggested a bootstrap method which resamples the p.s.u.'s with replacement. The method rescales the sampling weights instead of the bootstrap observations themselves in order to cover nonsmooth statistics. The weights are rescaled as follows:

$$w_{ij}^* = w_{ij}\big[(1-(m/(n-1))^{1/2}) + (m/(n-1))^{1/2}(n/m)r_i\big],$$

where m is the bootstrap sample size and r_i is the number of times the ith p.s.u. is selected. For the linear case with p.s.u.'s drawn with replacement, the method provides an unbiased variance estimator but not the usual with-replacement variance estimator because of the Monte Carlo error caused by the finite B. For the general case, the method overestimates the true variance to some extent but it has the attractive feature that it does not require knowledge of the sampling design beyond the first stage.

As Rust and Rao (1996) noted, there is considerable benefit and little loss, if any, in choosing $m = n-1$ in which case $w_{ij}^* = w_{ij}(n/(n-1))r_i$. For $n > 2$, as noted by Rust and Rao (1996), the bootstrap has a distinct advantage over BRR which is tedious to apply with $n > 2$. In this connection, see Nigam and Rao (1996) for the concept of balanced bootstrap.

Shao and Tu (1995, Section 6.4.4) discussed the consistency of different bootstrap variance estimators of a function of averages. They also discussed the consistency of bootstrap estimators of the distribution of an appropriate pivotal quantity involving a function of averages or a sample quantile. A general theory for higher order comparison of bootstrap confidence intervals with rival methods is not available due to the difficulties in establishing Edgeworth expansions for complex survey data. However, some simulation results show that bootstrap one-sided confidence intervals perform better than those constructed using normal approximations. See Shao and Tu (1995, Section 6.3) for further details.

3. IMPUTATION

A first application of the bootstrap method known as the approximate Bayesian bootstrap can be found in Rubin and Schenker (1986). To understand their method, consider the estimation of the finite population mean $\bar{Y}$ when a SRSWR of size n is drawn from a finite population. The method requires construction of M completed data sets using the following steps:

1. Draw r bootstrap donors by taking a SRSWR of size r from the r observed values (donors) in the original sample.
2. Impute $m = n - r$ missing values with SRSWR from the r bootstrap donors obtained in step 1.
3. Repeat steps 1 and 2 to obtain M completed data sets.

Let $\bar{y}_{Il}$ and s_{Il}^2 $(l = 1, \ldots, M)$ denote the sample mean and the sample variance for these M completed data sets. Then an approximate Bayesian bootstrap estimator of $\bar{Y}$ is given by $\bar{y}_I = M^{-1} \sum_{l=1}^{M} \bar{y}_{Il}$ with estimated variance

$$v(\bar{y}_I) = \frac{1}{M} \sum_{l=1}^{M} \frac{s_{Il}^2}{n} + \frac{M+1}{M} \left\{ \frac{1}{M-1} \sum_{l=1}^{M} (\bar{y}_{Il} - \bar{y}_I)^2 \right\}.$$

This approximate Bayesian bootstrap can be viewed as a modification of the multiple imputation which Rubin started in the early 1970's in the context of a compensatory database at the Educational Testing Service (Rubin, 1977, 1994). Such a modification was suggested in order to achieve good randomization properties of the multiple imputation for the popular hot-deck imputation which is improper in the sense of Rubin (1987, pages 118–119). For large m, the approximate Bayesian bootstrap method has been shown to be asymptotically valid under simple random sampling when estimating means or totals. For small m, certain adjustments are needed to achieve good randomization properties of the approximate Bayesian bootstrap. See Rubin and Schenker (1986) and Shao and Tu (1985, Section 6.5.3) for further details. It is, however, not clear how this approximate Bayesian bootstrap method can be extended so that it can provide valid randomization inferences for both smooth and nonsmooth statistics and for different imputation methods (proper or improper) in the context of complex surveys. See Fay (1993) and Rao (1996, 2000).

Following the suggestion of Efron (1994), Shao and Sitter (1996) considered a bootstrap method for analyzing imputed complex survey data. The method first replaces imputed values in the bootstrap sample by new imputed values obtained using the bootstrap donors and the same imputation method as in the original sample. It is interesting to compare the Shao–Sitter method with the Rubin–Schenker method for the SRSWR case with hot-deck imputation. The Shao–Sitter method involves the following steps:

1. Draw a SRSWR bootstrap sample of size $n-1$ from the original sample.
2. Replace each missing value in the bootstrap sample in step 1 by a randomly chosen observation from the donors in the bootstrap sample.
3. Repeat steps 1 and 2 to obtain M final bootstrap samples.

The bootstrap variance estimator can then be obtained using the M bootstrap samples using the standard formula, that is, v_b given earlier in this section.

For the general stratified multistage design given in Section 2.2, the Shao–Sitter method requires new imputations (using the same method as in the original sample) for missing values in the bootstrap sample but does not require any change in the bootstrap weights. Saigo, Shao and Sitter (2001) proposed a modification of the Shao–Sitter method that does not require rescaling and can be applied where random imputation is used and the first-stage stratum sample sizes are very small.

The role of modeling cannot be ignored in analyzing data that are not *missing completely at random* (MCAR). Rubin (1976) formalized models to explain the missing-data mechanism. Since then, a number of different classes of models have been proposed in the literature. Two such classes are the class of selection models (see Little and Rubin, 1987) and the class of pattern-mixture models (see Little, 1993). Likelihood-based methods, both classical and Bayesian, have been used to analyze missing data using these models. It is conceivable that the parametric bootstrap method will improve the classical likelihood-based methods in this context.

4. PARAMETRIC BOOTSTRAP IN SMALL-AREA ESTIMATION

To estimate per-capita income for small areas (population less than 1000), Fay and Herriot (1979) considered an aggregate level model and used an empirical Bayes method which combines survey data from the U.S. Current Population Survey with various administrative and census records. Their empirical Bayes estimator worked well when compared to the direct survey estimator and a synthetic estimator used earlier by the Census Bureau. To estimate areas planted with corn and soybeans for 12 counties (small areas) of north central Iowa, Battese, Harter and Fuller (1988) used a nested error regression model to combine information from a firm survey and satellite data which provide information on the number of pixels planted with corn and soybeans for each county. To cover

these two important small-area models, and other models in common use, Butar and Lahiri (2003) (see also Butar, 1997) proposed the following model for small-area data analysis.

Let Y_i be an $n_i \times 1$ vector of observations available from the ith small area and let U_i be a $k_i \times 1$ vector of small-area effects. Let X_i and Z_i be $n_i \times p$ and $n_i \times k_i$ matrices of known constants. Let $n = \sum_{i=1}^m n_i$ and $k = \sum_{i=1}^m k_i$. Consider the following Bayesian model:

MODEL 1.

1. $Y_i \mid U_i \overset{\text{ind}}{\sim} N_{n_i}(X_i\beta + Z_iU_i, R_i)$, $i = 1, \ldots, m$;
2. A priori, $U_i \overset{\text{ind}}{\sim} N_{k_i}(0, G_i)$, $i = 1, \ldots, m$,

where β is a $p \times 1$ column vector of unknown regression coefficients and $R_i = R_i(\psi)$ and $G_i = G_i(\psi)$ are, respectively, $n_i \times n_i$ and $k_i \times k_i$ matrices which possibly depend on ψ, an $s \times 1$ vector of unknown variance components.

EXAMPLE (The Fay–Herriot model).

1. $Y_i \mid \theta_i \overset{\text{ind}}{\sim} N(\theta_i, D_i)$, $i = 1, \ldots, m$;
2. A priori, $\theta_i \overset{\text{ind}}{\sim} N(X_i'\beta, A)$, $i = 1, \ldots, m$,

where the D_i's are known and the X_i's are $p \times 1$ vectors of known constants. In the notation of Model 1, $n_i = k_i = 1$, $Z_i = 1$, $U_i = \theta_i - X_i'\beta$, $\psi = A$, $R_i(\psi) = D_i$ and $G_i(\psi) = A$ $(i = 1, \ldots, m)$.

Generally, in small-area estimation, we consider the estimation of $\theta_i = l_i'\beta + \lambda_i'U_i$, where l_i and λ_i are $p \times 1$ and $k_i \times 1$ vectors of known constants, respectively. Under the above model and squared error loss function, the Bayes estimator of θ_i is given by

$$\begin{aligned}&\hat\theta_i(Y_i; \beta, \psi)\\ &\quad = l_i'\beta + \lambda_i'G_i(\psi)Z_i'V_i^{-1}(\psi)(Y_i - X_i\beta),\end{aligned}$$

where $V_i(\psi) = R_i + Z_iG_iZ_i'$ $(i = 1, \ldots, m)$.

When ψ is known but β is unknown, β is estimated by the maximum likelihood estimator $\hat\beta(\psi)$, where

$$\hat\beta(\psi) = \left[\sum_{i=1}^m X_i'V_i^{-1}(\psi)X_i\right]^{-1}\left[\sum_{i=1}^m X_i'V_i^{-1}(\psi)Y_i\right].$$

Plugging in $\hat\beta(\psi)$ for β in the Bayes estimator, we get the following empirical Bayes estimator of θ_i:

$$\begin{aligned}&\hat\theta_i\ (Y_i; \psi)\\ &\quad = l_i'\hat\beta(\psi) + \lambda_i'G_i(\psi)Z_i'V_i^{-1}(\psi)[Y_i - X_i\hat\beta(\psi)].\end{aligned}$$

Note that $\hat\theta_i(Y_i; \psi)$ is also the best linear unbiased predictor (BLUP) under the following mixed linear model:

$$Y_i = X_i\beta + Z_iU_i + e_i,$$

where the e_i's are independent of the U_i's and $e_i \overset{\text{ind}}{\sim} N_{n_i}(0, R_i)$, $i = 1, \ldots, m$. In practice, β and ψ are both unknown. In this case, an empirical Bayes estimator of θ_i is obtained as $\hat\theta_i(Y_i; \hat\psi)$, where $\hat\psi$ is a consistent estimator of ψ satisfying certain regularity conditions (see Butar and Lahiri, 2003). This is also an EBLUP under the above mixed linear model.

4.1 Parametric Bootstrap MSE Estimator

Butar and Lahiri (2003) developed a parametric bootstrap method to estimate the MSE of $\hat\theta_i(Y_i; \hat\psi)$, defined as $\text{MSE}[\hat\theta_i(Y_i; \hat\psi)] = E[\hat\theta_i(Y_i; \hat\psi) - \theta_i]^2$, where the expectation is taken with respect to Model 1. They first used the following well-known identity due to Kackar and Harville (1984):

$$\begin{aligned}\text{MSE}\big[\hat\theta_i(Y_i; \hat\psi)\big] &= g_{1i}(\psi) + g_{2i}(\psi)\\ &\quad + E\big[\hat\theta_i(Y_i; \hat\psi) - \hat\theta_i(Y_i; \psi)\big]^2,\end{aligned} \tag{1}$$

where $g_{1i}(\psi) = \text{MSE}[\hat\theta_i(Y_i; \beta, \psi)]$, and the second and third terms on the right-hand side of (1) measure additional uncertainties due to the estimation of β and ψ, respectively. Butar and Lahiri (2003) used their parametric bootstrap twice—once to estimate the first two terms of (1) by correcting the bias of $g_{1i}(\hat\psi) + g_{2i}(\hat\psi)$ and then to estimate the third term involving uncertainty due to the estimation of ψ. The parametric bootstrap MSE estimator is then given by

$$\begin{aligned}V_i^{\text{BOOT}} &= g_{1i}(\hat\psi) + g_{2i}(\hat\psi)\\ &\quad - E_\star\big[g_{1i}(\hat\psi^\star) + g_{2i}(\hat\psi^\star)\\ &\qquad - g_{1i}(\hat\psi) - g_{2i}(\hat\psi)\big]\\ &\quad + E_*\big[\hat\theta_i(Y_i; \hat\psi^*) - \hat\theta_i(Y_i; \hat\psi)\big]^2,\end{aligned}$$

where $E_\star$ is the expectation with respect to the following bootstrap model (i.e., Model 2) which mimics Model 1 and the calculation of $\hat\psi^\star$ is the same as that of $\hat\psi$ except that it is based on $Y_i^\star$'s instead of Y_i's.

MODEL 2.

1. $Y_i^\star \mid U_i^\star \overset{\text{ind}}{\sim} N_{n_i}(X_i\hat\beta + Z_iU_i^\star, \hat R_i)$, $i = 1, \ldots, m$;
2. A priori, $U_i^\star \overset{\text{ind}}{\sim} N_{k_i}(0, \hat G_i)$, $i = 1, \ldots, m$,

where $\hat R_i = R_i(\hat\psi)$ and $\hat G_i = G_i(\hat\psi)$.

Under certain regularity conditions, Butar and Lahiri (2003) showed that

$$E\big[V_i^{\text{BOOT}}\big] = \text{MSE}(\hat\theta_i^{\text{EB}}) + o(m^{-1}).$$

For a special balanced case of the Fay–Herriot model, the Butar–Lahiri parametric bootstrap is identical [up to order $O(m^{-1})$] to the measure of uncertainty proposed by Morris (1983) who approximated the posterior variance under flat priors on the hyperparameters. For an application of the Butar–Lahiri parametric bootstrap for nonnormal mixed models, see Lahiri and Maiti (2002). Alternative methods have been proposed to estimate the MSE of EBLUP or EB. See Lahiri (1995), Jiang, Lahiri and Wan (2002) and Chen and Lahiri (2002) for jackknife methods and Prasad and Rao (1990) and Datta and Lahiri (2000) for Taylor series methods.

Laird and Louis (1987) proposed a measure of uncertainty of an empirical Bayes estimator for a very special case of the Fay–Herriot model with $X_i'\beta = \mu$ and $D_i = D$ $(i = 1, \ldots, m)$. Instead of estimating the MSE, they attempted to approximate the posterior variance under an unspecified prior on the hyperparameters. Butar and Lahiri (2003) argued that in certain cases the Laird–Louis parametric bootstrap method may provide a measure smaller than the naive measure $g_{1i}(\hat{\psi}) + g_{2i}(\hat{\psi})$ which does not account for additional uncertainties due to the estimation of various hyperparameters and hence may severely underestimate the true uncertainty of the empirical Bayes estimator. The Laird–Louis parametric bootstrap method has been extended to solve a variety of problems. See Arora, Lahiri and Mukherjee (1997), Booth and Hobert (1998), Butar and Lahiri (2003), Gail, Pfeiffer, van Houwelingen and Carroll (2000), among others.

Recently, Pfeffermann and Tiller (2002) considered a parametric bootstrap approximation to the prediction mean square error (PMSE) for state-space models with estimated parameters. Their paper and Butar and Lahiri (2003) complement each other and one cannot be considered a special case of the other. This is because of the differences in the asymptotic settings. As noted earlier, Butar and Lahiri (2003) developed asymptotic properties of their parametric bootstrap when m is large with bounded sample size within each area. In contrast, asymptotic validity of the parametric bootstrap of Pfeffermann and Tiller (2002) is with respect to the sample size within the single area (e.g., n_1 in our notation).

The parametric bootstrap method is quite general and can conceivably be applied to any model whatsoever for which an original estimation problem is defined. However, the study of the asymptotic properties of such a parametric bootstrap method could be very challenging.

4.2 Interval Estimation

The naive confidence interval of the small-area mean θ_i based on the normal posterior distribution with estimated hyperparameters is usually too narrow to meet the target empirical Bayes coverage probability in the sense of Morris (1983). One possible reason for this undercoverage is that the naive method does not account for the additional uncertainties incurred due to the estimation of the hyperparameters. Even after the corrections described in Section 4.1, the confidence interval continues to have an undercoverage problem. This may be due to the fact that the distribution of the pivotal random variable used in the construction of the confidence interval is not adequately approximated by the normal distribution.

Chatterjee and Lahiri (2002) suggested a parametric bootstrap confidence interval for θ_i in the Fay–Herriot model. To explain their method, first note that an empirical Bayes estimator of θ_i is given by

$$\hat{\theta}_i^{\mathrm{EB}} = (1 - \hat{B}_i)Y_i + \hat{B}_i X_i'\hat{\beta},$$

where $\hat{B}_i = D_i/(D_i + \hat{A})$. The parametric bootstrap empirical Bayes estimator of θ_i^* is given by

$$\hat{\theta}_i^{*\mathrm{EB}} = (1 - \hat{B}_i^*)Y_i^* + \hat{B}_i^* X_i'\hat{\beta}^*,$$

where $(Y_1^*, \ldots Y_m^*)$ is the bootstrap sample generated using Model 2 and the quantities $\hat{B}_i^*$ and $\hat{\beta}^*$ are computed from the bootstrap sample using identical formulas as with the original data. Now obtain t_0 such that

$$\mathbb{P}^*\Big[\theta_i^* \in \Big\{\hat{\theta}_i^{\mathrm{EB}*} \pm t_0\sqrt{D_i[1 - \hat{B}_i(\hat{A}^*)]}\Big\}\Big] = 1 - \alpha.$$

Chatterjee and Lahiri (2002) showed that

$$\mathbb{P}\left[\theta_i \in \left\{\hat{\theta}_i^{\mathrm{EB}} \pm t_0\sqrt{D_i(1 - \hat{B}_i)^{1/2}}\right\}\right] = 1 - \alpha + O(n^{-3/2}).$$

Extensions of the above parametric bootstrap method to other complex small-area models are currently under investigation. Note that Carlin and Gelfand (1991) proposed certain parametric bootstrap confidence intervals for a wide variety of models following a suggestion of Efron (1987). However, the orders of accuracy of their intervals are not yet known.

5. APPLICATIONS

In this section, we briefly discuss two applications of the bootstrap in government agencies. For other real-life applications, see Kaufman (1996), Butar and Lahiri (2003) and Roberts, Kovacevic, Mantel and Phillips (2001), among others.

EXAMPLE 1 (National Center for Education Statistics). Zhang, Brick, Kaufman and Walter (1998) used a variant of the Shao–Sitter bootstrap method to assess the effect of imputation error on variance estimation using data from the 1993–1994 Schools and Staffing Survey (SASS). To estimate standard errors using the bootstrap method, 48 replicate weights were carefully created. The survey contains many quantitative and categorical variables on the 47,105 public school teachers who responded to the survey. However, the authors considered six categorical and seven continuous variables for their study.

For an arbitrary estimate, say $\hat{\theta}$, inflation of the standard error was measured by $\text{ste}_I(\hat{\theta})/\text{ste}(\hat{\theta})$, where ste_I and ste denote, respectively, the bootstrap standard error that incorporates the extra imputation error and the traditional standard error that does not. The study shows that this inflation factor could be high especially for the categorical variables with high imputation rates.

The biggest impact of the bootstrap was observed for T0040, a categorical variable generated by the following question with six possible response categories:

"Which of these categories best describes your other assignment at this school?"

(1) Administrator (e.g., principal, assistant principal, director, school head),
(2) counselor,
(3) library media specialist/librarian,
(4) coach,
(5) other professional staff (e.g., department head, curriculum coordinator),
(6) support staff (e.g., secretary, aide).

The imputation rate for this variable is 24%. We chose this variable to demonstrate the impact of the bootstrap method. Four types of imputation methods were used in the SASS. However, for their study the authors chose a directional nearest neighbor procedure which is explained below. The imputation was carried out in the following steps:

1. Different imputation classes were formed by the following matching variables: (i) groups of states with similar schools (STGROUP), (ii) state, (iii) instructional level of teacher (TEALEVEL), (iv) type of community where the school is located (URB) and (v) number of students enrolled in the school (ENR). The following steps are then carried out within each imputation class.
2. The records are sorted by STGROUP, state, TEALEVEL, grade level taught this year (GRADELEVEL), URB, teaching assignment field (TEAFIELD) and ENR.
3. If the first record in the file is a nonrespondent, it is replaced by a donor which is the first nonmissing value as one goes down the data file sequentially. Otherwise, the first record is stored as a donor for the first missing value in the file.
4. If the second record is missing, it is imputed by the value stored in step 3. Otherwise, the nonmissing value for the second record replaces the donor stored in step 3 for imputations of the subsequent records in the data file.

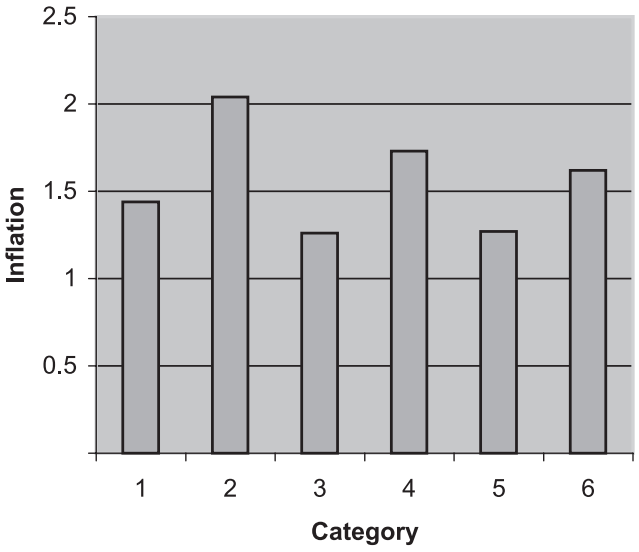

FIG. 1.

Figure 1 is obtained from the data given in Table 4 of Zhang, Brick, Kaufman and Walter (1998). It displays the inflation of standard error estimates. For all the categories, inflation is more than unity, implying that the traditional standard error estimates are generally underestimated. The amount of inflation could be as high as 2.1.

EXAMPLE 2 (U.S. Bureau of Labor Statistics). Pfeffermann and Tiller (2002), recently applied their method (discussed in Section 4) to a model fitted to 7-year (1992–1998) monthly data for *Employment to Population Ratio in the District of Columbia.* This series represents the number of employed persons as a percentage of the total population over 15 years of age and is one of the key economic series published

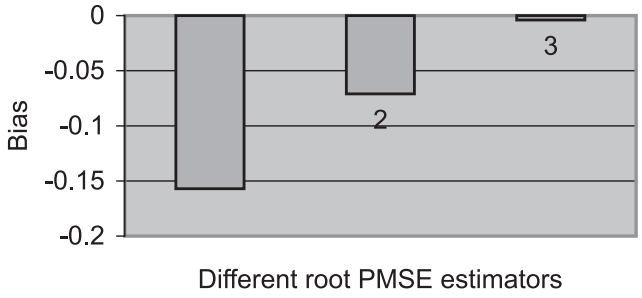

FIG. 2.

monthly by the Bureau of Labor Statistics for all the 50 states and the District of Columbia.

Due to the rotation pattern of the sampling design of the Current Population Survey data, sampling autocorrelations cannot be ignored and so in their simulation study Pfeffermann and Tiller (2002) considered an AR(15) model to describe the observational equation that corresponds to the sampling errors. The state equation part of their model (which corresponds to the unknown population ratio) involves trend, slope and seasonal components.

The bar charts given in Figures 2 and 3 are obtained using the data given in Table 1 of Pfeffermann and Tiller (2002). They display the biases and root PMSE of different root PMSE estimators of the trend predictor. In the bar charts, we use labels 1, 2 and 3 for the naive estimator, the bias corrected naive estimator and the parametric bootstrap estimator, respectively. The parametric bootstrap method is clearly very effective in eliminating the relatively large and very significant biases of the naive prediction MSE even for their complex model involving 18 hyperparameters estimated by a three-step estimation procedure. It also performs well in terms of the root PMSE. For other relevant results, see Pfeffermann and Tiller (2002).

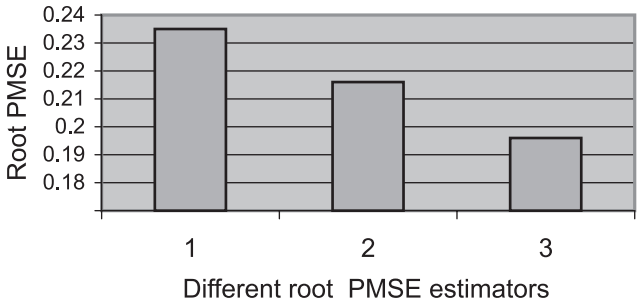

FIG. 3.

6. CONCLUDING REMARKS

In this article, we have reviewed a variety of problems encountered in survey sampling and discussed how bootstrap methods have been used to deal with such problems in an effective manner. Survey sampling is a fascinating field and is constantly offering practical problems that are theoretically challenging. The flexibility of the bootstrap and its straightforward implementation in a complex environment will certainly make the method promising to handle new problems in the field of survey sampling. We hope that this paper will serve as a bridge between mathematical statisticians developing bootstrap methods and practitioners working in various survey organizations.

ACKNOWLEDGMENTS

I thank the Editor, Steve Kaufman, Danny Pfeffermann and J. N. K. Rao for a number of constructive suggestions on an earlier draft of this paper. Supported in part by NSF Grant SES 99-78145 and a grant from the Gallup Organization.

REFERENCES

ARORA, V., LAHIRI, P. and MUKHERJEE, K. (1997). Empirical Bayes estimation of finite population means from complex surveys. *J. Amer. Statist. Assoc.* **92** 1555–1562.

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.* **83** 28–36.

BICKEL, P. J. and FREEDMAN, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12** 470–482.

BOOTH, J. G. and HOBERT, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 262–272.

BUTAR, F. B. (1997). Empirical Bayes methods in survey sampling. Ph.D. dissertation, Univ. Nebraska, Lincoln.

BUTAR, F. B. and LAHIRI, P. (2003). On measures of uncertainty of empirical Bayes small area estimators. *J. Statist. Plann. Inference* **112** 63–76.

CARLIN, B. and GELFAND, A. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc. Ser B* **53** 189–200.

CHATTERJEE, S. and LAHIRI, P. (2002). Parametric bootstrap confidence interval in small area estimation problems. Unpublished manuscript.

CHEN, S. and LAHIRI, P. (2002). A weighted jackknife method in small-area estimation. In *ASA Proc. Joint Statistical Meetings* 473–477. Amer. Statist. Assoc. Alexandria, VA.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* **10** 613–627.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM. Philadelphia.

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–200.

EFRON, B. (1994). Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* **89** 463–479.

FAY, R. E. (1993). Valid inferences from imputed survey data. In *Proc. Section on Survey Research Methods* 41–48. Amer. Statist. Assoc., Alexandria, VA.

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.

GAIL, M. H., PFEIFFER, R., VAN HOUWELINGEN, H. C. and CARROLL, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1** 231–246.

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statist. Sci.* **9** 55–93.

GROSS, S. (1980). Median estimation in sample surveys. In *Proc. Section on Survey Research Methods* 181–184. Amer. Statist. Assoc., Alexandria, VA.

HANSEN, M., HURWITZ, W. and MADOW, W. (1953). *Sample Survey Methods and Theory* **2**. Wiley, New York.

JIANG, J., LAHIRI, P. and WAN, S.-M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Ann. Statist.* **30** 1782–1810.

KACKAR, R. N. and HARVILLE, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.* **79** 853–862.

KALTON, G. and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology* **12** 1–16.

KAUFMAN, S. (1996). Estimating the variance in the presence of imputation using a residual. In *Proc. Section on Survey Research Methods* 423–428. Amer. Statist. Assoc., Alexandria, VA.

KOVAR, J. G., RAO, J. N. K. and WU, C.-F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canad. J. Statist.* **16** 25–45.

LAHIRI, P. (1995). A jackknife measure of uncertainty of linear empirical Bayes estimators. Unpublished manuscript.

LAHIRI, P. and MAITI, T. (2002). Empirical Bayes estimation of relative risks in disease mapping. *Calcutta Statist. Assoc. Bull.* **53** 213–223.

LAHIRI, P. and MEZA, J. (2002). Small-area estimation. In *Encyclopedia of Environmetrics* (A. H. El-Shaarawi and W. W. Piegorsch, eds.) **4** 2010–2014. Wiley, New York.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.* **82** 739–757.

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Internat. Statist. Rev.* **54** 139–157.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

MCCARTHY, P. J. and SNOWDEN, C.B. (1985). The bootstrap and finite population sampling. In *Vital and Health Statistics* 2–95. Public Health Service Publication 85-1369, U. S. Government Printing Office, Washington.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78** 47–65.

NIGAM, A. K. and RAO, J. N. K. (1996). On balanced bootstrap for stratified multistage samples. *Statist. Sinica* **6** 199–214.

PFEFFERMANN, D. (2002). Small area estimation—new developments and directions. *Internat. Statist. Rev.* **70** 125–143.

PFEFFERMANN, D. and TILLER, R. (2002). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. Technical report, U.S. Bureau of Labor Statistics.

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.

RAO, J. N. K. (1996). On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.* **91** 499–506.

RAO, J. N. K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology* **25** 175–186.

RAO, J. N. K. (2000). Variance estimation in the presence of imputation for missing data. In *Proc. Second International Conference on Establishment Surveys*. Amer. Statist. Assoc., Alexandria, VA.

RAO, J. N. K. and WU, C.-F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241.

RAO, J. N. K., WU, C.-F. J. and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18** 209–217.

ROBERTS, G., KOVACEVIC, M., MANTEL, H. and PHILLIPS, O. (2001). Cross-sectional inference based on longitudinal surveys: Some experiences with Statistics Canada surveys. Available at http://www.fcsm.gov/01papers/Mantel.pdf.

RUBIN, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63** 581–592.

RUBIN, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Statist. Assoc.* **72** 538–543.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

RUBIN, D. B. (1994). Discussion of "Missing data, imputation, and the bootstrap," by B. Efron. *J. Amer. Statist. Assoc.* **89** 475–478.

RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374.

RUST, K. F. and RAO, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5** 283–310.

SAIGO, H., SHAO, J. and SITTER, R. R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology* **27** 189–196.

SHAO, J. and SITTER, R. R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.* **91** 1278–1288.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

SITTER, R. R. (1992a). A resampling procedure for complex survey data. *J. Amer. Statist. Assoc.* **87** 755–765.

SITTER, R. R. (1992b). Comparing three bootstrap methods for survey data. *Canad. J. Statist.* **20** 135–154.

ZHANG, F., BRICK, M., KAUFMAN, S. and WALTER, E. (1998). Variance estimation of imputed survey data. Working Paper 98-14, U.S. Department of Education.