Statistical Science
2003, Vol. 18, No. 2, 191–198
© Institute of Mathematical Statistics, 2003

Impact of the Bootstrap on Sample Surveys

Jun Shao

Abstract. This article discusses the impact of the bootstrap on sample surveys and introduces some of the main developments of the bootstrap methodology for sample surveys in the last twenty five years.

Key words and phrases: Variance estimation, easy implementation, robustness, without replacement sampling, stratification, imputation.

1. INTRODUCTION

A crucial part of statistical inference is the assessment of variability in estimators of unknown parameters, which has a direct impact on measuring the uncertainty in parameter estimation, comparing the efficiencies of different estimators and/or sampling designs and constructing inference procedures such as confidence sets. In sample surveys, for example, it is a common practice to report parameter estimates in a tabular form along with their variance estimates (or estimates of coefficient of variation). In statistical literature, there exist two general approaches in assessing variability. One is the traditional analytic approach and the other one is data resampling, which requires a large amount of computation and has seen a steady growth along with the fast developments in computing facilities.

Since the publication of the first research article about the bootstrap (Efron, 1979), the bootstrap has become the most important and popular data resampling method. There is no doubt that the success of the bootstrap relies on the research developments over the last 25 years that not only made the bootstrap a sophisticated tool applicable to nearly all areas of statistical analysis, but also greatly advanced the theory of data resampling methods. This article discusses the reasons why the bootstrap is an important tool in sample surveys (Section 2) and introduces the main developments of the bootstrap methodology for survey applications (Sections 3–5).

2. THE BOOTSTRAP IN SAMPLE SURVEYS

Let $\mathbf{X}$ be the observed dataset, θ be an unknown parameter of interest and $\hat{\theta} = \hat{\theta}(\mathbf{X})$ be a chosen estimator of θ. In the analytic approach, the variability of $\hat{\theta}$ is assessed by first deriving an explicit theoretical formula that approximates the distribution of $\hat{\theta}$ or its characteristics such as the variance. If the derived theoretical formula contains some unknown quantities, then they are substituted by some estimates based on $\mathbf{X}$. For example, one may show that $\hat{\theta}$ is approximately normal with mean θ and an unknown asymptotic variance v that can be estimated using $\mathbf{X}$. The bootstrap method, however, works in a different manner. Suppose that we can estimate the statistical model that produces $\mathbf{X}$ and generate a number of bootstrap datasets, $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$, from the estimated model. The bootstrap method assesses the variability of $\hat{\theta}$ by the variability among the bootstrap estimates $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{X}^{*b})$, $b = 1, \dots, B$. Thus, the bootstrap method replaces theoretical derivations in the analytic approach by the generation of $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ and repeated computations of $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{X}^{*b})$, $b = 1, \dots, B$. In many situations, researchers have shown that the analytic approach and the bootstrap method produce approximately the same results and, therefore, the choice between these two approaches depends on the feasibility of their implementation.

Although in most cases the validity of the bootstrap is established by showing that the conditional distribution of $\hat{\theta}^{*b}$ given $\mathbf{X}$ is approximately the same as the distribution of $\hat{\theta}$, the explicit theoretical formula required by the analytic approach (such as the asymptotic variance v) is not needed in applying the bootstrap. When the derivation of the explicit theoretical formula is much more complicated than showing the existence

Jun Shao is Professor of Statistics, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1685.

of such a formula (which is usually enough for the validity of the bootstrap), the bootstrap is preferred since the required computations are typically routine and can be handled by a powerful computer. To illustrate this point and the attractiveness of the bootstrap in survey problems, we consider the following two examples.

EXAMPLE 1. Consider the estimation of the poverty line (low-income cutoff) of the population in the study of income shares or wealth distribution. For the ith sampled family, let z_i be the expenditure on necessities, y_i be the total income and x_{it}, $t = 1, \dots, T$, be variables such as urbanization category and family size. Then

$$(1) \qquad \log z_i = \gamma_1 + \gamma_2 \log y_i + \sum_{t=1}^{T} \beta_t x_{it} + \text{error},$$

where the γ's and β's are unknown parameters. Let γ_0 be the overall proportion of income spent on necessities. Then the poverty line θ can be defined as the solution of

$$(2) \quad \log[(\gamma_0 + 0.2)\theta] = \gamma_1 + \gamma_2 \log\theta + \sum_{t=1}^{T} \beta_t x_{0t}$$

for a particular set of $x_{01}, \dots, x_{0T}$ (Mantel and Singh, 1991). Suppose that the regression parameters in (1) are estimated by the least squares estimators $\hat{\gamma}_j$ and $\hat{\beta}_t$ and γ_0 is estimated by $\bar{z}/\bar{y}$, where $\bar{z}$ and $\bar{y}$ are the sample means of the z_i's and y_i's, respectively. Then a natural estimator $\hat{\theta}$ of θ is the solution of (2) with γ_j and β_t replaced by $\hat{\gamma}_j$ and $\hat{\beta}_t$, respectively. Let $X_i = (z_i, y_i, \log z_i, \log y_i, x_{i1}, \dots, x_{iT})$ and let $\bar{X}$ be the sample mean of X_i's. Then $\hat{\theta} = g(\bar{X})$ for an implicit differentiable function g. Applying Taylor's expansion and the central limit theorem, one can show that $\hat{\theta}$ is approximately normal with mean θ and variance $[\nabla g(\mu)]' V(\bar{X}) \nabla g(\mu)$, where ∇g is the vector of partial derivatives of g, $\mu = E(\bar{X})$ and $V(\bar{X})$ is the variance–covariance matrix of $\bar{X}$. To apply the analytic approach, one has to derive a formula for ∇g, which is complicated. This is not only because g is implicitly defined, but also because the $\hat{\gamma}_t$'s and $\hat{\beta}_t$'s are functions (of sample means of $\log z_i$, $\log y_i, x_{i1}, \dots, x_{iT}$) whose derivatives are messy when T is large. To apply the bootstrap, one just needs to repeatedly compute $g(\bar{X}^{*b})$, $b = 1, \dots, B$, where $\bar{X}^{*b}$ is the same as $\bar{X}$ but based on the bth bootstrap dataset. Note that any algorithm used to compute the original $g(\bar{X})$ can be used to compute $g(\bar{X}^{*b})$. The empirical distribution of $g(\bar{X}^{*b})$, $b = 1, \dots, B$, is a valid approximation to the distribution of $\hat{\theta}$ as long as g is continuously differentiable, since its limit is the same as that of $\hat{\theta}$ (Bickel and Freedman, 1984). But the explicit form of ∇g is not needed in applying the bootstrap.

EXAMPLE 2. A unique feature of survey data is the existence of a large proportion of nonrespondents. Imputation is commonly applied to compensate for nonresponse. We consider the Current Employment Survey (CES) conducted monthly by the U.S. Bureau of Labor Statistics. The main variables are the number of employees, the number of nonsupervisory workers and the hours and earnings of workers on nonagricultural establishment payrolls. Population employment counts are obtained once a year (month 0) from unemployment insurance administrative records. In any particular month, imputation for nonresponse using reported data from previous months generally provides more efficient survey estimators than ignoring nonrespondents and adjusting survey weights. Starting from month 1, nonrespondents are imputed using the following imputation method proposed by Butani, Harter and Wolter (1997). Let $y^{\mathrm{E}}_{t,i}$ be the number of employees of the ith sampled unit at month t. If $y^{\mathrm{E}}_{t,i}$ is a nonrespondent, then it is imputed by

$$\tilde{y}^{\mathrm{E}}_{t,i} = \tilde{y}^{\mathrm{E}}_{t-1,i} \frac{\sum_{j \in R_t} w_j y^{\mathrm{E}}_{t,j}}{\sum_{j \in R_t} w_j y^{\mathrm{E}}_{t-1,j}},$$

where $\tilde{y}^{\mathrm{E}}_{t-1,i} = y^{\mathrm{E}}_{t-1,i}$ if $y^{\mathrm{E}}_{t-1,i}$ is a respondent and is an imputed value otherwise, the w_j's are the survey weights (see Section 4) and R_t is the set of all reporting units for months t and $t-1$. If $y^{\mathrm{W}}_{t,i}$ (the number of nonsupervisory workers) is a nonrespondent, then it is imputed by

$$\tilde{y}^{\mathrm{W}}_{t,i} = \tilde{y}^{\mathrm{W}}_{t-1,i} \tilde{y}^{\mathrm{E}}_{t,i} / \tilde{y}^{\mathrm{E}}_{t-1,i},$$

where $\tilde{y}^{\mathrm{W}}_{t-1,i}$ is defined similarly to $\tilde{y}^{\mathrm{E}}_{t-1,i}$. If $y^{\mathrm{H}}_{t,i}$ (the number of hours worked) is a nonrespondent, then it is imputed by

$$\tilde{y}^{\mathrm{H}}_{t,i} = \frac{\tilde{y}^{\mathrm{H}}_{t-1,i} \tilde{y}^{\mathrm{W}}_{t,i}}{\tilde{y}^{\mathrm{W}}_{t-1,i}} \frac{\sum_{j \in R_t} w_j y^{\mathrm{H}}_{t,j} / \sum_{j \in R_t} w_j y^{\mathrm{W}}_{t,j}}{\sum_{j \in R_t} w_j y^{\mathrm{H}}_{t-1,j} / \sum_{j \in R_t} w_j y^{\mathrm{W}}_{t-1,j}},$$

where $\tilde{y}^{\mathrm{H}}_{t-1,i}$ is defined similarly to $\tilde{y}^{\mathrm{E}}_{t-1,i}$. Finally, if $y^{\mathrm{P}}_{t,i}$ (the weekly gross pay) is a nonrespondent, then it is imputed by

$$\tilde{y}^{\mathrm{P}}_{t,i} = \frac{\tilde{y}^{\mathrm{P}}_{t-1,i} \tilde{y}^{\mathrm{H}}_{t,i}}{\tilde{y}^{\mathrm{H}}_{t-1,i}} \frac{\sum_{j \in R_t} w_j y^{\mathrm{P}}_{t,j} / \sum_{j \in R_t} w_j y^{\mathrm{H}}_{t,j}}{\sum_{j \in R_t} w_j y^{\mathrm{P}}_{t-1,j} / \sum_{j \in R_t} w_j y^{\mathrm{H}}_{t-1,j}},$$

where $\tilde{y}^{\mathrm{P}}_{t-1,i}$ is defined similarly to $\tilde{y}^{\mathrm{E}}_{t-1,i}$. After imputation, weighted averages of imputed data are used to estimate monthly totals of the four variables. It can be seen that the estimated totals at month t are differentiable functions of various weighted averages of data from months $0, 1, \dots, t-1, t$, but the forms of these functions are very messy, because imputed values in months $1, \dots, t-1$ are carried over to impute nonrespondents in month t. Thus, deriving an explicit variance formula (using Taylor's expansion) for an estimated total is very complicated. The application of the bootstrap method in this case is straightforward, except that each bootstrap dataset $\mathbf{X}^{*b}$ should be reimputed using the imputation method for the original dataset (see Section 5). Hence, applying the bootstrap requires many computations, but the complicated derivations for explicit variance formulas can be avoided. Another example of this kind can be found in Section 5.

In the CES, the nonresponse rate is about 20–40% and about 60% of the nonrespondents in a given month may become available one or several months later. At month t, reported data for month $s < t$ (which were not reported at month s) are used in the imputation procedure for nonrespondents at month t. This adds another complication to the analytic approach, but has no effect on applying the bootstrap as long as each bootstrap dataset is imputed using exactly the same imputation algorithm as that for the original dataset. We may view this as a robustness property of the bootstrap, against changes in the imputation procedure.

Easy implementation and the robustness property described in Example 2 are probably the reasons that the bootstrap method has become popular in many survey agencies such as the U.S. Census Bureau, the U.S. Bureau of Labor Statistics and Statistics Canada. On the other hand, some special features of survey problems have had an impact on the development of the bootstrap methodology. In the 1980's, research on the bootstrap for sampling without replacement and for stratified sampling with many small-size strata was very active and resulted in many different ways of generating bootstrap datasets. In the 1990's, attention was given to the application of the bootstrap to data with imputed nonrespondents. These developments together with other stimulating research on the bootstrap (e.g., the research on bootstrap confidence intervals and bootstrapping dependent data) made the bootstrap a more complete methodology for statistical inference.

There exist other data resampling methods that assess the variability of $\hat{\theta}$ by $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{X}^{*b})$, $b = 1, \dots, B$, but differ from the bootstrap in the construction of $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$. Two data resampling methods having a long history of application in sample surveys are the jackknife and balanced repeated replication (BRR). The jackknife constructs $\mathbf{X}^{*b}$ by deleting some units in $\mathbf{X}$, whereas the BRR forms $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ as proper subsets of $\mathbf{X}$ that have some balancedness property. Unlike the bootstrap data, $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ for the jackknife or the BRR are not constructed randomly. As a result, the jackknife and the BRR can be applied to approximate the first and second moments (bias and variance) of $\hat{\theta}$, but not the distribution of $\hat{\theta}$. The jackknife is known to have problems in approximating variances of not very smooth estimators such as the sample quantiles, whereas the construction of balanced subsets $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ with a reasonable size B may be difficult to implement in the BRR. The bootstrap is more flexible in implementation and can be applied to different estimators (smooth or nonsmooth) or different problems (variance estimation or distribution estimation) so that a single resampling method can be used for various problems. Furthermore, special features in the sampling process of $\mathbf{X}$ can often be built into the bootstrap sampling process of $\mathbf{X}^{*b}$ (e.g., data with imputed values or data sampled without replacement) and, therefore, the bootstrap is more natural and much easier to understand than the other resampling methods.

3. WITHOUT-REPLACEMENT BOOTSTRAP

In sample surveys, the original dataset $\mathbf{X}$ is often sampled without replacement from a finite population. Let n be the sample size (the number of sampled units in $\mathbf{X}$) and let N be the population size (the number of units in the finite population). The ratio n/N is called the sampling fraction. Intuitively, if the sampling fraction is negligible, then the distribution of $\hat{\theta} = \hat{\theta}(\mathbf{X})$ under sampling without replacement is almost the same as that under sampling with replacement. Technically, we need to specify an asymptotic framework so that we can describe the negligibility of n/N in a limiting sense. We assume that the finite population under study is a member of a sequence of finite populations indexed by $\nu = 1, 2, \dots$. Thus, our sample $\mathbf{X} = \mathbf{X}_\nu$ is a sample of size n_ν from the νth finite population of size N_ν, but the index ν in $\mathbf{X}$, n, and N will be suppressed for simplicity of notation. As $\nu \to \infty$, both n and N increase to ∞ and the sampling fraction is negligible if and only if $n/N \to 0$.

Consider the case where $\mathbf{X} = \{X_1, \dots, X_n\}$ is a simple random sample without replacement from a finite population with N units, where X_i is a d-vector of observations. Without loss of generality, we assume that the finite population is $\{X_1, \dots, X_n, X_{n+1}, \dots, X_N\}$. Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ be the sample mean and let $\mu = N^{-1}\sum_{i=1}^{N} X_i$ be the unknown population mean. From the sampling theory (Bickel and Freedman, 1984), $\bar{X}$ is asymptotically (as $\nu \to \infty$) normal with mean μ and covariance matrix

$$\Sigma_{wo} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)'.$$

If sampling is with replacement, then $\bar{X}$ is asymptotically normal with mean μ and covariance matrix

$$\Sigma_w = \frac{1}{nN}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)'.$$

Note that Σ_{wo} and Σ_w are asymptotically the same if and only if $n/N \to 0$.

Let $\mathbf{X}^*$ be a bootstrap sample generated from $\mathbf{X}$; that is, $\mathbf{X}^*$ is a simple random sample from $\mathbf{X}$. Suppose that $\mathbf{X}^* = \{X_1^*, \dots, X_n^*\}$ is of size n and sampled with replacement. Then, conditional on $\mathbf{X}$, the sample mean $\bar{X}^* = n^{-1}\sum_{i=1}^{n} X_i^*$ is asymptotically normal with mean $\bar{X}$ and covariance matrix

$$\hat{\Sigma}_w = \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

Note that $\bar{X}$ is asymptotically the same as μ and $\hat{\Sigma}_w$ is asymptotically the same as Σ_w. Hence, if $\mathbf{X}$ is sampled with replacement and a large number of bootstrap estimates $\bar{X}^{*b}$, $b = 1, \dots, B$, are obtained, then the empirical distribution of $\bar{X}^{*b}$, $b = 1, \dots, B$, provides an asymptotically valid approximation to the distribution of $\bar{X}$. On the other hand, if $\mathbf{X}$ is sampled without replacement and bootstrap sampling is with replacement, then the empirical distribution of $\bar{X}^{*b}$, $b = 1, \dots, B$, does not provide an asymptotically valid approximation to the distribution of $\bar{X}$ when $n/N \not\to 0$, since Σ_{wo} and Σ_w are not asymptotically the same.

Ideally, one should generate a bootstrap sample by taking a simple random sample without replacement from $X_1, \dots, X_n$, where the bootstrap sampling fraction is the same as the original sampling fraction n/N. Such a bootstrap procedure results in a bootstrap sample size much smaller than n, which may not be desirable. Gross (1980) and Chao and Lo (1985) proposed the following without-replacement bootstrap method. Assume for simplicity that $N = kn$ with an integer k. We first create a pseudo-population of size N by replicating $X_1, \dots, X_n$ exactly k times and then generate a bootstrap sample $\mathbf{X}^*$ by taking a simple random sample of size n without replacement from the pseudo-population. Note that the bootstrap sampling size and sampling fraction are the same as the original sampling size and sampling fraction, respectively. If $\bar{X}^*$ is the bootstrap sample mean based on $\mathbf{X}^*$ generated based on this without-replacement bootstrap, then, conditional on $\mathbf{X}$, $\bar{X}^*$ is asymptotically normal with mean $\bar{X}$ and covariance matrix

$$\hat{\Sigma}_{wo} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{N}{n(N-1)}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})',$$

which is asymptotically the same as Σ_{wo}. Thus, the empirical distribution of $\bar{X}^{*b}$, $b = 1, \dots, B$, provides an asymptotically valid approximation to the distribution of $\bar{X}$.

More sophisticated without-replacement bootstraps can be found in Bickel and Freedman (1984) and Sitter (1992a, b).

4. BOOTSTRAPPING STRATIFIED SAMPLES WITH SMALL STRATUM SIZES

Stratified sampling with small stratum sizes is a popular sampling design in modern sample surveys. Suppose that the finite population under study has been stratified into H strata with N_h population units in the hth stratum. For each h, $n_h \geq 2$ units are sampled from stratum h using some probability sampling, independently across the strata. Let $\mathcal{S}$ denote the set of indices of the sampled units and let $\mathcal{P}$ denote the set of indices of the population units. We assume that survey weights w_i, $i \in \mathcal{S}$, are constructed according to the sampling design so that, for any set of values $\{y_i : i \in \mathcal{P}\}$,

$$E\left(\sum_{i\in\mathcal{S}} w_i y_i\right) = \sum_{i\in\mathcal{P}} y_i,$$

where E is the expectation with respect to $\mathcal{S}$. Note that $\hat{Y} = \sum_{i\in\mathcal{S}} w_i y_i$ is the so-called Horvitz–Thompson estimator of the population total $Y = \sum_{i\in\mathcal{P}} y_i$.

We consider the case where all n_h are small, that is, the n_h's are bounded by a fixed positive integer, but H is large. (Technically, we assume that $H \to \infty$ as $\nu \to \infty$, where ν is the index for the finite population; see Section 3.) Let $n = \sum_h n_h$ and $N = \sum_h N_h$. Then n/N is the overall sampling fraction. Since the n_h's are bounded, n/N is usually negligible ($n/N \to 0$) and, therefore, we may ignore the sampling fractions even if the sampling within each stratum is without replacement.

Within each stratum, it is difficult to generate a bootstrap sample according to the original probability sampling design, especially when n_h is small. Making use of the survey weights, one may simplify the bootstrap procedure by generating a simple random sample from the observed values. Let $\mathcal{S}_h$ be the set of indices of the sampled units in stratum h ($\bigcup_h \mathcal{S}_h = \mathcal{S}$). A straightforward application of the standard bootstrap is to generate a simple random sample $\mathcal{S}_h^*$ of size n_h with replacement from $\mathcal{S}_h$ and define the bootstrap sample to be $\mathcal{S}^* = \bigcup_h \mathcal{S}_h^*$, where $\mathcal{S}_h^*$, $h = 1, \ldots, H$, are independently generated. The bootstrap Horvitz–Thompson estimator of Y is

$$\hat{Y}^* = \sum_{i\in\mathcal{S}^*} w_i y_i. \tag{3}$$

Let E_* and V_* be the bootstrap expectation and variance, respectively, with respect to the bootstrap sample (conditional on $\mathcal{S}$). It is easy to show that $E_*(\hat{Y}^*) = \hat{Y}$. By the central limit theorem, the conditional distribution of $\hat{Y}^*$ approximates that of $\hat{Y}$ if $V_*(\hat{Y}^*)$ is approximately the same as the sampling variance of $\hat{Y}$.

In the mid-1980's, however, researchers found that this standard bootstrap method produces invalid bootstrap estimates when the n_h's are bounded. This can be explained as follows. From the theory of sample surveys, an approximately valid estimator of the variance of $\hat{Y}$ is

$$v = \sum_h n_h s_h^2,$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i\in\mathcal{S}_h} \left(w_i y_i - \frac{1}{n_h} \sum_{j\in\mathcal{S}_h} w_j y_j \right)^2.$$

Since $\mathcal{S}_h^*$ is a simple random sample of size n_h with replacement from $\mathcal{S}_h$,

$$V_*(\hat{Y}^*) = \sum_h n_h V_*(w_{i^*} y_{i^*}) = \sum_h (n_h - 1) s_h^2,$$

where i^* denotes a unit in $\mathcal{S}_h^*$ and the last equality follows from

$$\begin{aligned} V_*(w_{i^*} y_{i^*}) &= \frac{1}{n_h} \sum_{i\in\mathcal{S}_h} \left(w_i y_i - \frac{1}{n_h} \sum_{j\in\mathcal{S}_h} w_j y_j \right)^2 \\ &= \frac{(n_h - 1) s_h^2}{n_h}. \end{aligned} \tag{4}$$

From these formulas, we conclude that the variability of the bootstrap estimator $\hat{Y}^*$ is much too small compared with the variance of $\hat{Y}$. If $n_h = 2$ for all h, for example, $v = 2V_{\mathcal{S}^*}(\hat{Y}^*)$. This problem does not exist if all the n_h's are large ($\min_h n_h \to \infty$).

Several modified bootstrap procedures have been proposed to circumvent this problem since the mid-1980's. They can be grouped into the following three types.

The Use of Different Bootstrap Sample Sizes

McCarthy and Snowden (1985) proposed using $n_h - 1$ as the bootstrap sample size in the hth stratum. If $\mathcal{S}_h^*$ is a simple random sample of size $n_h - 1$ with replacement from $\mathcal{S}_h$ and $\hat{Y}^*$ in (3) is replaced by

$$\tilde{Y}^* = \sum_h \frac{n_h}{n_h - 1} \sum_{i\in\mathcal{S}_h^*} w_i y_i$$

[note that the factor $n_h/(n_h - 1)$ is added to reflect the fact that the bootstrap sample size in stratum h is $n_h - 1$, not n_h], then $E_*(\tilde{Y}^*) = \hat{Y}$ and

$$\begin{aligned} V_*(\tilde{Y}^*) &= \sum_h \frac{n_h^2}{(n_h - 1)^2} (n_h - 1) V_*(w_{i^*} y_{i^*}) \\ &= \sum_h n_h s_h^2 = v \end{aligned}$$

[by (4)]. Hence, this bootstrap procedure is an asymptotically valid method in assessing the variability of $\hat{Y}$. Sitter (1992a) proposed a mirror-match bootstrap procedure that extends McCarthy and Snowden's.

The Rescaling Bootstrap

Rao and Wu (1988) proposed the following rescaling of the original bootstrap. Let $\mathcal{S}_h^*$ be a simple random sample of size m_h with replacement from $\mathcal{S}_h$. Replace $\hat{Y}^*$ in (3) by the rescaling bootstrap estimator

$$\begin{aligned} \tilde{Y}^* = \sum_h \Bigg[& \sqrt{\frac{m_h}{n_h - 1}} \frac{n_h}{m_h} \sum_{i\in\mathcal{S}_h^*} w_i y_i \\ & + \left(1 - \sqrt{\frac{m_h}{n_h - 1}} \right) \sum_{i\in\mathcal{S}_h} w_i y_i \Bigg]. \end{aligned}$$

Then, for any m_h, $E_*(\hat{Y}^*) = \hat{Y}$ and

$$\begin{aligned} V_*(\tilde{Y}^*) &= \sum_h \frac{m_h}{n_h - 1} \frac{n_h^2}{m_h^2} m_h V_*(w_{i^*} y_{i^*}) \\ &= \sum_h n_h s_h^2 = v \end{aligned}$$

[by (4)]. Note that the bootstrap sample size can be the same as the original sample size, that is, $m_h = n_h$.

When $m_h = n_h - 1$, the rescaling bootstrap reduces to McCarthy and Snowden's. When $n_h > 2$, Rao and Wu (1988) suggested $m_h \approx (n_h - 2)^2/(n_h - 1)$ by matching the third-order moments of $\hat{Y}$ and $\tilde{Y}^*$.

The Repeated Half-Sample Bootstrap

The previous two types of modified bootstrap require either rescaling or a bootstrap sample size different from n_h, which causes a problem when applying the bootstrap to survey data with imputed nonrespondents. Saigo, Shao and Sitter (2001) proposed the following repeated half-sample bootstrap that uses the bootstrap sample size n_h and does not require rescaling.

First, consider an even n_h. Let $\tilde{\mathcal{S}}_h^*$ be a simple random sample of size $n_h/2$ without replacement from $\mathcal{S}_h$. Let the bootstrap sample in stratum h be $\mathcal{S}_h^* = \tilde{\mathcal{S}}_h^* \cup \tilde{\mathcal{S}}_h^*$, which indicates why this bootstrap method is called the repeated half-sample bootstrap. Then

$$E_*\left(\sum_{i \in \mathcal{S}_h^*} w_i y_i\right) = E_*\left(2 \sum_{i \in \tilde{\mathcal{S}}_h^*} w_i y_i\right) = \sum_{i \in \mathcal{S}_h} w_i y_i$$

and

$$\begin{aligned} V_*\left(\sum_{i \in \mathcal{S}_h^*} w_i y_i\right) &= V_*\left(2 \sum_{i \in \tilde{\mathcal{S}}_h^*} w_i y_i\right) \\ &= n_h^2 V_*\left(\frac{1}{n_h/2} \sum_{i \in \tilde{\mathcal{S}}_h^*} w_i y_i\right) \\ &= 2n_h\left(1 - \frac{1}{2}\right) s_h^2 = n_h s_h^2, \end{aligned}$$

where the third equality follows from the theory of sampling without replacement.

Next, consider an odd n_h.

(i) If $\tilde{\mathcal{S}}_h^*$ is a simple random sample of size $(n_h - 1)/2$ without replacement from $\mathcal{S}_h$ and we define the bootstrap sample in stratum h to be $\mathcal{S}_h^* = \tilde{\mathcal{S}}_h^* \cup \tilde{\mathcal{S}}_h^* \cup \{i^*\}$, where i^* is a single unit selected at random from $\tilde{\mathcal{S}}_h^*$, then the size of $\mathcal{S}_h^*$ is n_h, $E_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = \sum_{i \in \mathcal{S}_h} w_i y_i$ and $V_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = (n_h + 3)s_h^2$.

(ii) If $\tilde{\mathcal{S}}_h^*$ is a simple random sample of size $(n_h - 1)/2 + 1$ without replacement from $\mathcal{S}_h$ and we define the bootstrap sample in stratum h to be $\mathcal{S}_h^* = \tilde{\mathcal{S}}_h^* \cup \tilde{\mathcal{S}}_h^* - \{i^*\}$, where i^* is a single unit selected at random from $\tilde{\mathcal{S}}_h^*$, then the size of $\mathcal{S}_h^*$ is n_h, $E_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = \sum_{i \in \mathcal{S}_h} w_i y_i$ and $V_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = (n_h - 1)s_h^2$.

Thus, if we use method (i) with probability $1/4$ and method (ii) with probability $3/4$ in forming the bootstrap sample $\mathcal{S}_h^*$, the resulting bootstrap estimator has the desired property that $E_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = \sum_{i \in \mathcal{S}_h} w_i y_i$ and $V_*(\sum_{i \in \mathcal{S}_h^*} w_i y_i) = n_h s_h^2$.

If bootstrap samples are independently generated across strata according to the previously described method (for even and odd n_h's), then the bootstrap estimator $\hat{Y}^*$ defined by (3) satisfies $E_*(\hat{Y}^*) = \hat{Y}$ and

$$V_*(\hat{Y}^*) = \sum_h n_h s_h^2 = v.$$

Therefore, the repeated half-sample bootstrap is asymptotically valid in assessing the variability of $\hat{Y}$. Some simulation results can be found in Saigo, Shao and Sitter (2001).

5. BOOTSTRAPPING IMPUTED DATA

As we described in Example 2, nonresponse often occurs in surveys and imputation is commonly applied to compensate for nonresponse. Let $\mathbf{X}_I$ be the dataset with imputed nonrespondents. If $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is an approximately unbiased estimator of a parameter θ in the case of no nonresponse, then typically the imputation method is designed so that $\hat{\theta}_I = \hat{\theta}(\mathbf{X}_I)$ has approximately the same mean as $\hat{\theta}$. The variability of $\hat{\theta}_I$, however, is typically larger than that of $\hat{\theta}$, because of nonresponse and imputation. If we treat $\mathbf{X}_I$ as the observed dataset and apply any of the bootstrap procedures described in the previous sections to obtain a bootstrap dataset $\mathbf{X}_I^*$, then the variability of the bootstrap estimator $\hat{\theta}(\mathbf{X}_I^*)$ is smaller than that of $\hat{\theta}_I$, since the imputation process is ignored. This was noted by Efron (1994) and Shao and Sitter (1996) and they proposed reimputing the bootstrap dataset $\mathbf{X}_I^*$ in the same way as the original dataset was imputed (assuming that a response indicator is attached to each unit so that nonrespondents in $\mathbf{X}_I^*$ can be identified) and assessing the variability of $\hat{\theta}_I$ based on the re-imputed bootstrap dataset $\mathbf{X}_{II}^* = (\mathbf{X}_I^*)_I$. Does the conditional distribution of $\hat{\theta}(\mathbf{X}_{II}^*)$ provide a valid approximation to the distribution of $\hat{\theta}_I$? The answer depends on the method used to impute nonrespondents, the sampling design (whether the n_h's are large) and the type of bootstrap procedure.

We now summarize the results that have been established since the mid-1990's. For simple random sampling, Efron (1994) showed that the standard bootstrap procedure together with reimputing bootstrap datasets produces a valid approximation to the distribution of $\hat{\theta}_I$. The same result was established in Shao and Sitter (1996) for stratified sampling with large n_h's ($\min_h n_h \to \infty$).

The situation of stratified sampling with some small n_h's (which is considered in Section 4), however, is more complicated. As we discussed in Section 4, the standard bootstrap does not produce valid approximations when some n_h's are small. One of the three types of modified bootstrap described in Section 4 has to be used. Shao and Sitter (1996) pointed out that the second type of modified bootstrap in Section 4, the rescaling bootstrap proposed by Rao and Wu (1988), does not work for data with imputed values. The success of the first type of modified bootstrap (i.e., McCarthy and Snowden's bootstrap that uses $n_h - 1$ as the bootstrap sample size in stratum h) depends on the type of imputation procedure. In general, imputation procedures can be classified into two types. The first type is deterministic imputation; that is, given the observed data (and auxiliary data), imputed values are nonrandom. Mean imputation, ratio imputation and regression imputation are examples of deterministic imputation. The second type is random imputation, which imputes nonrespondents by random values generated from some conditional distributions given the observed data (and auxiliary data). The simplest example of random imputation is imputing nonrespondents by a random sample drawn from the respondents. Another example, random regression imputation, is given in Example 3.

Shao and Sitter (1996) showed that McCarthy and Snowden's bootstrap procedure together with reimputing bootstrap datasets produces a valid approximation of the distribution of $\hat{\theta}_I$ when imputation is deterministic. For random imputation, however, Saigo, Shao and Sitter (2001) showed that McCarthy and Snowden's bootstrap procedure together with reimputing bootstrap datasets overestimates the variability of $\hat{\theta}$, because the bootstrap sample size in stratum h is $n_h - 1$, not the original sample size n_h. The overestimation is serious only when some n_h's are very small, according to the simulation results in Saigo, Shao and Sitter (2001). The case $n_h = 2$ is, however, an important special case in stratified sampling. This motivates the development of the third type of modified bootstrap in Section 4, the repeated half-sample bootstrap, which together with reimputing bootstrap datasets produces a valid approximation of the distribution of $\hat{\theta}_I$, regardless of whether the imputation is random or not and whether n_h is small or not (Saigo, Shao and Sitter, 2001).

To appreciate the importance of using random imputation and, again, the attractiveness of using the bootstrap to replace theoretical derivations required by the analytic approach, we consider the following example.

EXAMPLE 3. Let $\mathcal{S}$ be a stratified sample from a population $\mathcal{P}$ as described in Section 4 and let X_i, $i \in \mathcal{P}$, be a q-dimensional vector of variables of interest. In addition to the population totals of components of X_i, suppose that we are also interested in estimating the population correlation coefficients between any pairs of components of X_i. Let x_{ij} and x_{ik} be, respectively, the jth and kth components of X_i. When there is no nonresponse, a standard estimator of the correlation coefficient between the jth and kth components is the sample correlation coefficient

$$(5)\quad \frac{\sum_{i\in\mathcal{S}} w_i(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{[\sum_{i\in\mathcal{S}} w_i(x_{ij}-\bar{x}_j)^2 \sum_{i\in\mathcal{S}} w_i(x_{ik}-\bar{x}_k)^2]^{1/2}},$$

where $\bar{x}_j = \sum_{i\in\mathcal{S}} w_i x_{ij} / \sum_{i\in\mathcal{S}} w_i$ is the estimated population mean of the jth component. A Taylor expansion variance estimator for the estimator in (5) can be derived, although the derivation is complicated (for the simple random sample case, this derivation is an exercise given in Serfling, 1980, Problem 6 on page 136).

When there are nonrespondents, we consider regression imputation based on the following model:

$$(6)\quad \begin{aligned} &X_i = B'Z_i r + V_i^{1/2}E_i \quad \text{and} \\ &P(A_i = A|X_i, Z_i) = P(A_i = A|Z_i), \quad i \in \mathcal{P}, \end{aligned}$$

where Z_i is a vector of auxiliary variables (covariates), B is a matrix of unknown parameters, V_i is a diagonal matrix whose elements are known functions of Z_i, E_i is a random vector independent of Z_i with mean 0 and unknown covariance matrix Σ and A_i is a vector of response indicators of X_i. The first condition in (6) is a typical assumption of a multivariate regression model between X_i and Z_i and the second condition in (6) means that the response indicators for X_i are conditionally independent of X_i, given Z_i. In practice, model (6) may not hold for all units in $\mathcal{P}$, but may hold for units in $\mathcal{P}_t \subset \mathcal{P}$ with $\bigcup_t \mathcal{P}_t = \mathcal{P}$, in which case B and Σ may depend on t and imputation is carried out within each $\mathcal{P}_t$. For simplicity, we assume model (6) holds for all units in $\mathcal{P}$. Note that model (6) can still be used even if there is no covariate Z_i.

Let $\hat{B}$ be the weighted least squares estimator of B based on the respondents. Let $\nu \subset \{1, \dots, q\}$ denote the set of indices corresponding to missing components. For any vector c, let c_ν be the subvector containing components of c indexed by the integers in ν. A deterministic regression imputation method imputes $X_{i\nu}$ by $\hat{B}'_\nu Z_i$, where $\hat{B}_\nu$ is the submatrix containing columns of $\hat{B}$ indexed by the integers in ν. After imputation, parameter estimators are obtained by using standard estimation formulas for the case of no nonresponse and treating imputed values as observed data. For example, the correlation coefficient between the jth and kth components is estimated by using formula (5) with nonrespondents replaced by imputed values. Although this deterministic regression imputation produces approximately unbiased estimators for the population means of components of X_i, it does not provide an approximately unbiased estimator for the correlation coefficient (Shao and Wang, 2002).

A random regression imputation method that produces approximately unbiased estimators for both population means and correlation coefficients is proposed in Shao and Wang (2002). For any $q \times q$ matrix C and two subsets ν_1 and ν_2 of $\{1, \dots, q\}$, let $C^{\nu_1 \nu_2}$ be the submatrix containing rows of C indexed by the integers in ν_1 and columns of C indexed by the integers in ν_2. The random regression imputation method imputes $X_{i\nu}$ by

$$\hat{B}'_\nu Z_i + (V_i^{\nu\nu})^{1/2}[\hat{\Sigma}^{\nu\nu^c}(\hat{\Sigma}^{\nu^c\nu^c})^{-1} \cdot (V_i^{\nu^c\nu^c})^{-1/2}(X_{i\nu^c} - \hat{B}'_{\nu^c} Z_i) + \tilde{E}_i],$$

where $\nu^c = \{1, \dots, q\} - \nu$,

$$\hat{\Sigma} = \frac{\sum_{i \in R} w_i V_i^{-1/2}(X_i - \hat{B}' Z_i)(X_i - \hat{B}' Z_i)' V_i^{-1/2}}{\sum_{i \in R} w_i},$$

R is the set of indices for which X_i has no missing components and, given the observed data, the $\tilde{E}_i$'s are independent random vectors with mean 0 and covariance matrix $\hat{\Sigma} - \hat{\Sigma}^{\nu\nu^c}(\hat{\Sigma}^{\nu^c\nu^c})^{-1}\hat{\Sigma}^{\nu\nu^c}$.

It can be seen that the estimator given by (5) with nonrespondents imputed by random regression imputation is a function of various weighted averages, although the form of this function is very complicated. Thus, the repeated half-sample bootstrap with reimputation can be applied in assessing the variability of this estimator. On the other hand, deriving an explicit theoretical formula for the variance of this estimator is extremely complicated.

ACKNOWLEDGMENT

Supported in part by NSF Grant DMS-01-02223.

REFERENCES

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

BUTANI, S., HARTER, R. and WOLTER, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. In *Proc. Section on Survey Research Methods* 523–528. Amer. Statist. Assoc., Alexandria, VA.

CHAO, M. T. and LO, S.-H. (1985). A bootstrap method for finite populations. *Sankhyā Ser. A* **47** 399–405.

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

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

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

MANTEL, H. J. and SINGH, A. C. (1991). Standard errors of estimates of low proportions: A proposed methodology. Technical report, Statistics Canada.

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, DC.

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

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

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

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

SHAO, J. and WANG, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *J. Amer. Statist. Assoc.* **97** 544–552.

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.