Statistica Sinica **8**(1998), 1071-1085

BOOTSTRAPPING SAMPLE QUANTILES BASED ON COMPLEX SURVEY DATA UNDER HOT DECK IMPUTATION

Jun Shao and Yinzhong Chen

University of Wisconsin-Madison

Abstract: The bootstrap method works for both smooth and nonsmooth statistics, and replaces theoretical derivations by routine computations. With survey data sampled using a stratified multistage sampling design, the consistency of the bootstrap variance estimators and bootstrap confidence intervals was established for smooth statistics such as functions of sample means (Rao and Wu (1988)). However, similar results are not available for nonsmooth statistics such as the sample quantiles and the sample low income proportion. We consider a more complicated situation where the data set contains nonrespondents imputed using a random hot deck method. We establish the consistency of the bootstrap procedures for the sample quantiles and the sample low income proportion. Some empirical results are also presented.

Key words and phrases: Imputation classes, low income proportion, stratified multistage sampling.

1. Introduction

Variance estimation and confidence intervals are the main research focuses in survey problems. Popular methods include the linearization-substitution, the jackknife, the balanced repeated replication, and the bootstrap, which can be applied to problems in which the parameter of interest is a smooth function of population totals. When the parameter of interest is a population quantile or other nonsmooth parameters such as the low income proportion (to be described later), the jackknife is not applicable; a counterpart of the linearization-substitution method is Woodruff's method (Woodruff (1952)), but Kovar, Rao and Wu (1988) and Sitter (1992) found that Woodruff's method had poor empirical performance when the population was stratified using a concomitant variable highly correlated with the variable of interest. The balanced repeated replication method works well for quantiles (e.g., Shao and Wu (1992), Shao and Rao (1994), Shao (1996)), but it requires the construction of balanced subsamples, which can be very difficult if we have a stratified sample with many strata and very different stratum sizes. The bootstrap method works for both smooth and nonsmooth statistics so that one can use a single method for estimating variances and setting confidence intervals based on various statistics. The bootstrap requires a large amount of routine computation, which is usually not a serious problem with the fast computers we have nowadays.

Asymptotic validity of the bootstrap was shown by Rao and Wu (1988) for the case of smooth functions of population totals. Similar results for quantile problems are expected, but have not been rigorously established. While studying properties of the bootstrap for quantile problems, we go a step further; that is, we consider the situation where the data set contains missing values imputed by using a random hot deck method. Most surveys involve nonrespondents and random hot deck imputation is commonly used to compensate for missing data. If we apply the bootstrap by treating the imputed values as if they are true values, then the bootstrap variance estimator has a serious negative bias when the proportion of nonrespondents is appreciable. A correct bootstrap is obtained by imputing the bootstrap samples in the same way as imputing the original data set (Efron (1994), Shao and Sitter (1996)).

After describing (in Section 2) the sampling design, the imputation procedure, the survey estimators, and the bootstrap procedure, we establish the asymptotic validity of the bootstrap estimators in Section 3. Some empirical results are presented in Section 4.

2. Sampling Design, Imputation, and the Bootstrap

The following commonly used stratified multistage sampling design is considered. The population $\mathcal{P}$ under consideration has been stratified into L strata with N_h clusters in the hth stratum. From the hth stratum, $n_h \geq 2$ clusters are selected with replacement (or so treated) using some probability sampling plan, independently across the strata. Within the (h,i)th selected cluster, n_{hi} ultimate units are sampled according to some sampling methods, $i = 1,\ldots,n_h$, $h = 1,\ldots,L$. We do not need to specify the number of stages and the sampling methods used after the first-stage sampling. Associated with the kth ultimate unit in the ith sampled cluster of stratum h is a characteristic y_{hik} and a survey weight w_{hik}, $k = 1,\ldots,n_{hi}, i = 1,\ldots,n_h, h = 1,\ldots,L$. The survey weights are constructed so that

$$G(x) = \frac{1}{M}\sum_{S} w_{hik} I_{\{y_{hik}\}}(x) \tag{1}$$

is unbiased for the population distribution

$$F(x) = \frac{1}{M}\sum_{y_{hik}\in\mathcal{P}} I_{\{y_{hik}\}}(x), \tag{2}$$

where M is the total number of ultimate units in population $\mathcal{P}$, $\sum_S$ is the summation over S, the set of indices (h,i,k) in the sample, and $I_C(x)$ is the indicator function of the set C. Note that M may be unknown in many survey problems and G in (1) is not necessarily a distribution function. Hence, in the case of no missing data, a customary estimate of F in (2) is the empirical distribution

$$\hat{F} = G/G(\infty) = \sum_S w_{hij} I_{\{y_{hik}\}} / \sum_S w_{hik}.$$

Suppose that some values y_{hik} in the sample are missing and that there are V disjoint subsets S_v of S such that $S = S_1 \cup \cdots \cup S_V$ and units in S_v respond independently with the same probability, $v = 1, \ldots, V$. That is, within S_v, we have a uniform response mechanism. The S_v are called imputation classes and are constructed using auxiliary variables observed for every sampled unit; e.g., a variable x_{hik} taking the same value when $(h, i, k) \in S_v$ (Schenker and Welsh (1988), Section 4). Imputation is carried out within the imputation classes; that is, a missing value with $(h, i, k) \in S_v$ is imputed based on the respondents with indices in S_v. For a concise presentation we assume $V = 1$ throughout the paper. The extensions of the results to the case of any fixed $V \geq 2$ are straightforward.

There are two types of commonly used imputation methods: deterministic imputation (e.g., the mean imputation, the ratio imputation, and the regression imputation) and random (hot deck) imputation (e.g., Schenker and Welsh (1988), Rao and Shao (1992)). Deterministic imputation, however, does not directly produce valid estimates of population quantiles, since it does not preserve the distribution of item values. We shall, therefore, concentrate on the following random hot deck imputation. Let S_R and S_M be subsets of S containing the indices of respondents and nonrespondents, respectively. The random hot deck method imputes missing values by a random sample selecting (with replacement) y_{hik} with probability $w_{hik}/\sum_{S_R} w_{hik}$ for each $(h, i, k) \in S_R$. Based on the imputed data set, the empirical distribution is

$$\hat{F}^I = \Big(\sum_{S_R} w_{hik} I_{\{y_{hik}\}} + \sum_{S_M} w_{hik} I_{\{y^I_{hik}\}} \Big) / \sum_S w_{hik}, \tag{3}$$

where y^I_{hik} denotes the imputed value for y_{hik}, $(h, i, k) \in S_M$.

In studies of income shares or wealth distributions, an important class of population characteristics is $\theta = F^{-1}(p) = \inf\{x : F(x) \geq p\}$, the pth quantile of F, $p \in (0, 1)$. Another important parameter is the proportion of low income economic families. Let $\mu = F^{-1}(\frac{1}{2})$ be the population median family income. Then, the population low income proportion can be defined as $\rho = F(\frac{1}{2}\mu)$, ($\frac{1}{2}\mu$ is called the poverty line; see Wolfson and Evans (1990)).

Customary survey estimates of θ (with a fixed p) and ρ are the sample pth quantile and the sample low income proportion defined by

$$\hat{\theta}^I = (\hat{F}^I)^{-1}(p) \quad \text{and} \quad \hat{\rho}^I = \hat{F}^I(\tfrac{1}{2}\hat{\mu}^I), \tag{4}$$

respectively, where $\hat{\mu}^I = \hat{\theta}^I$ with $p = \frac{1}{2}$, the sample median. Under some conditions, $\hat{\theta}^I$ and $\hat{\rho}^I$ are shown to be asymptotically normal (Chen and Shao (1996)).

We now consider the bootstrap with no missing data. Since some of the n_h may not be large, the original bootstrap (Efron (1979)) has to be modified. We focus on McCarthy and Snowden's (1985) "with replacement bootstrap" method, which is a special case of Rao and Wu's (1988) rescaling bootstrap. Let $\{\boldsymbol{y}^*_{hi} : i = 1, \ldots, n_h - 1\}$ be a simple random sample with replacement from $\{\boldsymbol{y}_{hi} : i = 1, \ldots, n_h\}$, $h = 1, \ldots, L$, independently across the strata, where $\boldsymbol{y}_{hi} = \{y_{hik} : k = 1, \ldots, n_{hi}\}$. Define $\hat{F}^* = \sum_{S^*} w^*_{hik} I_{\{y^*_{hik}\}} / \sum_{S^*} w^*_{hik}$, where y^*_{hik} is the kth component of $\boldsymbol{y}^*_{hi}$, w^*_{hik} is $n_h/(n_h - 1)$ times the survey weight associated with y^*_{hik}, and S^* is the index set for the bootstrap sample. The bootstrap variance estimator for $\hat{\theta} = \hat{F}^{-1}(p)$ is $\text{Var}_*(\hat{\theta}^*)$, where $\hat{\theta}^* = (\hat{F}^*)^{-1}(p)$ and Var_* is the variance taken under the bootstrap distribution conditional on y_{hik}, $(h, i, k) \in S$. A level $1 - 2\alpha$ bootstrap confidence interval for θ is $C^* = [\hat{\theta} + d^*_{1-\alpha}, \hat{\theta} + d^*_{\alpha}]$, where d^*_a is the $(1 - a)$th quantile of the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$, conditional on y_{hik}, $(h, i, k) \in S$. Monte Carlo approximations are used if $\text{Var}_*(\hat{\theta}^*)$ or C^* has no closed form.

When there are imputed data, however, treating y^I_{hik} as if they are true values and applying the previously described bootstrap do not lead to consistent bootstrap variance estimators and correct bootstrap confidence intervals. The bootstrap procedure has to be modified to take into account the effect of missing data and imputation. Shao and Sitter (1996) proposed a bootstrap method which can be described as follows. Assume that the data set carries identification flags a_{hik} indicating whether a unit is a respondent; that is, $a_{hik} = 1$ if $(h, i, k) \in S_R$ and $a_{hik} = 0$ if $(h, i, k) \in S_M$.

1. Draw a simple random sample $\{\boldsymbol{y}^*_{hi} : i = 1, \ldots, n_h - 1\}$ with replacement from $\{\boldsymbol{y}_{hi} : i = 1, \ldots, n_h\}$, $h = 1, \ldots, L$, independently across the strata, where missing values in $\boldsymbol{y}_{hi}$ are imputed by y^I_{hik}.
2. Let a^*_{hik} be the response indicator associated with y^*_{hik}, $S^*_M = \{(h, i, k) : a^*_{hik} = 0\}$, and $S^*_R = \{(h, i, k) : a^*_{hik} = 1\}$. Then impute the "missing" value y^*_{hik}, $(h, i, k) \in S^*_M$, by y^{*I}_{hik} selected with replacement from $\{y^*_{hik} : (h, i, k) \in S^*_R\}$ with probability $w^*_{hik} / \sum_{S^*_R} w^*_{hik}$, independently for $(h, i, k) \in S^*_M$.
3. Define

$$\hat{F}^{*I} = \Big(\sum_{S^*_R} w^*_{hik} I_{\{y^*_{hik}\}} + \sum_{S^*_M} w^*_{hik} I_{\{y^{*I}_{hik}\}} \Big) \Big/ \sum_{S^*} w^*_{hik}, \tag{5}$$

$$\hat{\theta}^{*I} = (\hat{F}^{*I})^{-1}(p), \quad \text{and} \quad \hat{\rho}^{*I} = \hat{F}^{*I}(\tfrac{1}{2}\hat{\mu}^{*I}),$$

which are bootstrap analogues of $\hat{F}^I$, $\hat{\theta}^I$, and $\hat{\rho}^I$, respectively. The bootstrap variance estimators for $\hat{\theta}^I$ and $\hat{\rho}^I$ are $\text{Var}_*(\hat{\theta}^{*I})$ and $\text{Var}_*(\hat{\rho}^{*I})$, respectively, and bootstrap percentile confidence intervals for θ and ρ are

$$C^{*I} = [\hat{\theta}^I + d^{*I}_{1-\alpha}, \hat{\theta}^I + d^{*I}_{\alpha}] \quad \text{and} \quad \tilde{C}^{*I} = [\hat{\rho}^I + \tilde{d}^{*I}_{1-\alpha}, \hat{\rho}^I + \tilde{d}^{*I}_{\alpha}],$$

respectively, where d^{*I}_a and $\tilde{d}^{*I}_a$ are $(1-a)$th quantiles of the bootstrap distributions of $\hat{\theta}^{*I} - \hat{\theta}^I$ and $\hat{\rho}^{*I} - \hat{\rho}^I$, respectively.

The key feature in this bootstrap method is step 2; that is, the bootstrap data y^*_{hik} are imputed in the same way as the original data set. If there is no missing data, then step 2 can be omitted and this bootstrap reduces to that in McCarthy and Snowden (1985).

3. Consistency of the Bootstrap

We now show the consistency of the bootstrap estimators. For this purpose, assume that the finite population $\mathcal{P}$ is a member of a sequence of finite populations $\{\mathcal{P}_\nu : \nu = 1, 2, \ldots\}$. Therefore, the values L, M, N_h, n_h, n_{hi}, w_{hik}, y_{hik}, and y^I_{hik} depend on ν, but, for simplicity of notation, the population index ν for these values will be suppressed. Also, F, θ, μ, ρ, G, $\hat{F}$, $\hat{F}^I$, $\hat{\theta}^I$, $\hat{\mu}^I$, and $\hat{\rho}^I$ depend on ν and will be denoted by F_ν, θ_ν, μ_ν, ρ_ν, G_ν, $\hat{F}_\nu$, $\hat{F}^I_\nu$, $\hat{\theta}^I_\nu$, $\hat{\mu}^I_\nu$, and $\hat{\rho}^I_\nu$, respectively. All limiting processes will be understood to be as $\nu \to \infty$. We always assume that the sequence $\{\theta_\nu : \nu = 1, 2, \ldots\}$ is bounded and that the total number of selected first-stage clusters $n = \sum_{h=1}^L n_h$ is large; that is, $n \to \infty$ as $\nu \to \infty$.

We first state a lemma whose proof is given in the Appendix.

Lemma 1. *Assume that*

(A1) $\max_{h,i,k} n_{hi} w_{hik}/M = O(n^{-1})$;

(A2) *There is a sequence of functions* $\{f_\nu : \nu = 1, 2, \ldots\}$ *such that* $0 < \inf_\nu f_\nu(\theta_\nu) \leq \sup_\nu f_\nu(\theta_\nu) < \infty$ *and for any* $\delta_\nu = O(n^{-1/2})$,

$$\lim_{\nu \to \infty} \Big[\frac{F_\nu(\theta_\nu + \delta_\nu) - F_\nu(\theta_\nu)}{\delta_\nu} - f_\nu(\theta_\nu) \Big] = 0.$$

Then

$$\sup_{|x - \theta_\nu| \leq cn^{-1/2}} |H^*_\nu(x)| = o_p(n^{-1/2})$$

for any constant $c > 0$, *where*

$$H^*_\nu(x) = \hat{F}^{*I}_\nu(x) - \hat{F}^{*I}_\nu(\theta_\nu) - \hat{F}^I_\nu(x) + \hat{F}^I_\nu(\theta_\nu). \tag{6}$$

The following result provides a Bahadur representation for the bootstrap sample quantiles.

Theorem 1. *Assume* (A1) *and* (A2). *Then*

$$\hat{\theta}^{*I}_\nu = \hat{\theta}^I_\nu + \frac{\hat{F}^I_\nu(\theta_\nu) - \hat{F}^{*I}_\nu(\theta_\nu)}{f_\nu(\theta_\nu)} + o_p(n^{-1/2}) \tag{7}$$

and

$$\sup_x |K_\nu^*(x) - K_\nu(x)| = o_p(1), \tag{8}$$

where K_ν *is the distribution of* $\sqrt{n}(\hat{\theta}_\nu^I - \theta_\nu)$ *and* K_ν^* *is the bootstrap distribution of* $\sqrt{n}(\hat{\theta}_\nu^{*I} - \hat{\theta}_\nu^I)$, *conditional on* S_R *and* y_{hik}, $(h,i,k) \in S_R$.

Proof. Define $\zeta_\nu^*(t) = \sqrt{n}[F_\nu(\theta_\nu + tn^{-1/2}) - \hat{F}_\nu^{*I}(\theta_\nu + tn^{-1/2})]/f_\nu(\theta_\nu)$ and $\eta_\nu^*(t) = \sqrt{n}[F_\nu(\theta_\nu + tn^{-1/2}) - \hat{F}_\nu^{*I}(\hat{\theta}_\nu^{*I})]/f_\nu(\theta_\nu)$. By Lemma 1 and Lemma 1 in Chen and Shao (1996),

$$\zeta_\nu^*(0) - \zeta_\nu^*(t) = \sqrt{n}[H_\nu(\theta_\nu + tn^{-1/2}) + H_\nu^*(\theta_\nu + tn^{-1/2})]/f_\nu(\theta_\nu) = o_p(1),$$

where $H_\nu(x) = \hat{F}_\nu^I(x) - \hat{F}_\nu^I(\theta_\nu) - F_\nu(x) + F_\nu(\theta_\nu)$. From (A2),

$$\sqrt{n}\frac{F_\nu(\theta_\nu + tn^{-1/2}) - F_\nu(\theta_\nu)}{f_\nu(\theta_\nu)} \to t.$$

Also,

$$\Big|\sqrt{n}\frac{F_\nu(\theta_\nu) - \hat{F}_\nu^{*I}(\hat{\theta}_\nu^{*I})}{f_\nu(\theta_\nu)}\Big| \le \frac{\sqrt{n}}{f_\nu(\theta_\nu)}\Big[\frac{1}{M} + \frac{1}{G_\nu^*(\infty)}\max_{h,i,k}\frac{w_{hik}^*}{M}\Big] = O_p(n^{-1/2}).$$

Hence $\eta_\nu^*(t) - t = o_p(1)$ and

$$\sqrt{n}\Big[\hat{\theta}_\nu^{*I} - \theta_\nu - \frac{F_\nu(\theta_\nu) - \hat{F}_\nu^{*I}(\theta_\nu)}{f_\nu(\theta_\nu)}\Big] = \sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) - \zeta_\nu^*(0) = o_p(1). \tag{9}$$

Then result (7) follows from (9) and the Bahadur representation for $\hat{\theta}_\nu^I$ (Theorem 2 in Chen and Shao (1996)). Result (8) can be shown using result (7) and the bootstrap central limit theorem (Bickel and Freedman (1984)).

The next result is for bootstrapping the sample low income proportion.

Theorem 2. *Assume that* (A1) *holds and that* (A2) *holds when* $\theta_\nu = \mu_\nu$ *and* $\theta_\nu = \frac{1}{2}\mu_\nu$. *Then*

$$\hat{\rho}_\nu^{*I} - \hat{\rho}_\nu^I = \hat{F}_\nu^{*I}(\tfrac{1}{2}\mu_\nu) - \hat{F}_\nu^I(\tfrac{1}{2}\mu_\nu) + \tfrac{f_\nu(\frac{1}{2}\mu_\nu)}{2f_\nu(\mu_\nu)}[\hat{F}_\nu^I(\mu_\nu) - \hat{F}_\nu^{*I}(\mu_\nu)] + o_p(n^{-1/2}) \tag{10}$$

and

$$\sup_x |\tilde{K}_\nu^*(x) - \tilde{K}_\nu(x)| = o_p(1), \tag{11}$$

where $\tilde{K}_\nu$ *is the distribution of* $\sqrt{n}(\hat{\rho}_\nu^I - \rho_\nu)$ *and* $\tilde{K}_\nu^*$ *is the bootstrap distribution of* $\sqrt{n}(\hat{\rho}_\nu^{*I} - \hat{\rho}_\nu^I)$.

Proof. Result (11) can be obtained from result (10). Hence we only need to show (10). From Lemma 1,

$$U_\nu^{*I} = \hat{F}_\nu^{*I}(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \hat{F}_\nu^{*I}(\tfrac{1}{2}\mu_\nu) - \hat{F}_\nu^I(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) + \hat{F}_\nu^I(\tfrac{1}{2}\mu_\nu) = o_p(n^{-1/2}).$$

Using (A2), Lemma 1, Theorem 1, and Theorem 3 in Chen and Shao (1996), we have

$$\begin{aligned}\hat{\rho}_\nu^{*I} - \hat{\rho}_\nu^I &= \hat{F}_\nu^{*I}(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \hat{F}_\nu^I(\tfrac{1}{2}\hat{\mu}_\nu^I)\\ &= \hat{F}_\nu^{*I}(\tfrac{1}{2}\mu_\nu) - \hat{F}_\nu^I(\tfrac{1}{2}\mu_\nu) + \hat{F}_\nu^I(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \hat{F}_\nu^I(\tfrac{1}{2}\hat{\mu}_\nu^I) + U_\nu^{*I}\\ &= \hat{F}_\nu^{*I}(\tfrac{1}{2}\mu_\nu) - \hat{F}_\nu^I(\tfrac{1}{2}\mu_\nu) + F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^I) + o_p(n^{-1/2})\\ &= \hat{F}_\nu^{*I}(\tfrac{1}{2}\mu_\nu) - \hat{F}_\nu^I(\tfrac{1}{2}\mu_\nu) + \tfrac{f_\nu(\frac{1}{2}\mu_\nu)}{2f_\nu(\mu_\nu)}[\hat{F}_\nu^I(\mu_\nu) - \hat{F}_\nu^{*I}(\mu_\nu)] + o_p(n^{-1/2}).\end{aligned}$$

From Theorems 1 and 2, the bootstrap confidence intervals C^{*I} and $\tilde{C}^{*I}$ are asymptotically correct, i.e., $P\{\theta_\nu \in C^{*I}\} \to 1-2\alpha$ and $P\{\rho_\nu \in \tilde{C}^{*I}\} \to 1-2\alpha$.

The consistency of the bootstrap variance estimators is established in the next two theorems. Their proofs are given in the Appendix.

Theorem 3. *Assume* (A1) *and* (A2′) (A2) *holds with any* $\delta_\nu = O(n^{-1/2}\sqrt{\log n})$. *Assume further that*

$$\frac{1}{n^{1+\epsilon}}\sum_{h=1}^{L}\sum_{i=1}^{n_h} E\Big(\max_{k=1,\ldots,n_{hi}} |y_{hik}|^\epsilon\Big) \to 0 \tag{12}$$

for some $\epsilon > 0$ *and that* $\liminf_\nu n\sigma_\nu^2(\theta_\nu) > 0$, *where* $\sigma_\nu^2(x)$ *is the asymptotic variance of* $\hat{F}_\nu^I(x)$. *Then*

$$n[\mathrm{Var}_*(\hat{\theta}_\nu^{*I}) - \sigma_\nu^2(\theta_\nu)/f_\nu^2(\theta_\nu)] = o_p(1). \tag{13}$$

Remark. (1) Condition (A2′) is slightly stronger than (A2). Using (A2′) and the same arguments in the proof of Lemma 1 as in Chen and Shao (1996), we can show that

$$\sup_{|x-\theta_\nu|\le cn^{-1/2}\sqrt{\log n}} |\hat{F}_\nu^R(x) - \hat{F}_\nu^R(\theta_\nu) - F_\nu(x) + F_\nu(\theta_\nu)| = o_p(n^{-1/2}) \tag{14}$$

for any $c > 0$.
(2) Condition (12) is a very weak moment condition since ϵ can be any positive number. Without this condition, the bootstrap variance estimator for $\hat{\theta}_\nu^I$ may be inconsistent (Ghosh, Parr, Singh and Babu (1984)).

Theorem 4. *Assume that* (A1) *holds and that* (A2′) *holds with* $\theta_\nu = \frac{1}{2}\mu_\nu$ *and* $\theta_\nu = \mu_\nu$. *Then*

$$n[\mathrm{Var}_*(\hat{\rho}_\nu^{*I}) - \gamma_\nu^2] = o_p(1), \tag{15}$$

where γ_ν^2 *is the asymptotic variance of* $\hat{\rho}_\nu^I$ *given by*

$$\gamma_\nu^2 = \sigma_\nu^2(\tfrac{1}{2}\mu_\nu) + \sigma_\nu^2(\mu_\nu)\Big[\tfrac{f_\nu(\frac{1}{2}\mu_\nu)}{2f_\nu(\mu_\nu)}\Big]^2 - \sigma_\nu(\tfrac{1}{2}\mu_\nu,\, \mu_\nu)\tfrac{f_\nu(\frac{1}{2}\mu_\nu)}{f_\nu(\mu_\nu)}, \tag{16}$$

and $\sigma_\nu(x,y)$ is the asymptotic covariance between $\hat{F}^I_\nu(x)$ and $\hat{F}^I_\nu(y)$.

4. A Simulation Study

We present the results from a simulation study examining the performance of the bootstrap percentile confidence interval and the bootstrap variance estimator in a problem where $\hat{\theta}^I = (\hat{F}^I)^{-1}(\frac{1}{2})$, the sample median. For comparison, we also included Woodruff's method in the simulation study.

We considered a population with $L = 32$ strata. In the hth stratum, the y-values of the population were generated according to $y_{hi} \overset{\text{i.i.d.}}{\sim} N(\bar{Y}_h, \sigma_h^2)$, $i = 1,\ldots,N_h$, where the population parameters N_h, $\bar{Y}_h$, and σ_h are listed in Table 1.

Table 1. Population parameters and n_h

h	N_h	$\bar{Y}_h$	σ_h	n_h	h	N_h	$\bar{Y}_h$	σ_h	n_h
1	38	13.7	6.7	3	17	34	9.2	4.5	2
2	38	13.0	6.5	3	18	34	9.0	4.6	2
3	38	12.5	6.4	3	19	34	9.8	4.4	2
4	38	12.0	6.6	3	20	34	8.6	4.1	2
5	38	12.3	6.1	3	21	34	8.3	4.9	2
6	38	11.7	6.8	3	22	22	8.2	4.6	2
7	38	11.4	6.3	3	23	22	8.0	4.3	2
8	38	11.2	6.4	3	24	22	7.9	4.7	2
9	38	11.0	5.5	3	25	22	7.8	3.1	2
10	38	10.8	5.6	3	26	22	7.5	3.9	2
11	38	10.6	5.9	3	27	22	7.2	3.7	2
12	34	10.3	5.3	2	28	22	7.0	3.6	2
13	34	10.1	5.4	2	29	22	6.7	3.4	2
14	34	9.7	5.8	2	30	22	6.4	3.2	2
15	34	9.5	4.8	2	31	22	6.1	3.5	2
16	34	9.4	4.7	2	32	22	6.0	3.7	2

After the population was generated, a simple random sample of size n_h was drawn from stratum h, independently across the 32 strata. Thus, the sampling design is a stratified one stage simple random sampling. The sample sizes n_h are also listed in Table 1.

The respondents $\{y_{hi}, (h,i) \in S_R\}$ were obtained by assuming that the sampled units responded with equal probability p. The missing values $\{y_{hi}, (h,i) \in S_M\}$ were imputed by taking an i.i.d. sample from $\{y_{hi}, (h,i) \in S_R\}$, with selection probability $w_{hi}/\sum_{A_r} w_{hi}$ for y_{hi}, $(h,i) \in S_R$, where the survey weight $w_{hi} = w_h = N_h/n_h$ in this special case.

The bootstrap percentile confidence interval (for the population median) and the bootstrap variance estimator (for the sample median $\hat{\theta}^I$) were computed according to the procedure described in Section 2. Monte Carlo approximations of size 1,000 were used in computing quantities such as d_a^{*I} and $\text{Var}_*(\hat{\theta}^{*I})$. The Woodruff's confidence interval is

$$[(\hat{F}^I)^{-1}(\tfrac{1}{2} - 1.96\hat{\sigma}(\hat{\theta}^I)),\ (\hat{F}^I)^{-1}(\tfrac{1}{2} + 1.96\hat{\sigma}(\hat{\theta}^I))],$$

where, for any fixed x, $\hat{\sigma}^2(x)$ is the adjusted jackknife variance estimator (Rao and Shao (1992)) for the "statistic" $\hat{F}^I(x)$. Woodruff's variance estimator is given by

$$\Big[\frac{(\hat{F}^I)^{-1}(\frac{1}{2} + 1.96\hat{\sigma}_\nu^*(\hat{\theta}^I)) - (\hat{F}^I)^{-1}(\frac{1}{2} - 1.96\hat{\sigma}_\nu^*(\hat{\theta}^I))}{2 \times 1.96}\Big]^2.$$

The simulation size was 10,000. All the computations were done in UNIX at the Department of Statistics, University of Wisconsin-Madison, using IMSL subroutines GENNOR, IGNUIN and GENUNF for random number generations.

Table 2. Results for the sample median

		RB (%)		MSE		Prob (%)		Length	
p (%)	True Var	BM	WM	BM	WM	BM	WM	BM	WM
100	0.69	1.4	6.3	0.32	0.37	92.6	93.6	3.19	3.10
90	0.86	0.4	12.1	0.41	0.48	92.9	96.0	3.54	3.57
80	1.03	2.3	19.2	0.53	0.63	93.1	96.8	3.91	4.03
70	1.20	5.9	25.9	0.67	0.80	93.3	97.4	4.28	4.47
60	1.47	4.7	25.9	0.84	0.98	93.1	97.9	4.70	4.94
50	1.77	8.0	28.8	1.10	1.27	93.8	97.7	5.21	5.45
40	2.19	11.4	32.0	1.44	1.67	94.6	97.7	5.88	6.10

BM: the bootstrap method
WM: Woodruff's method

Table 2 lists, for some values of the response probability p, the true variance of $\hat{\theta}^I$ (approximated by the sample variance of the 10,000 simulated values of $\hat{\theta}^I$), the empirical relative bias (RB) of the bootstrap or Woodruff's variance estimator, which is defined as (the average of simulated variance estimates − the true variance)/ the true variance, the mean square error (MSE) of the bootstrap or Woodruff's variance estimator, and the coverage probability and length of 95% bootstrap or Woodruff's confidence interval.

The following is a summary of the simulation results.

1. Variance estimation. The bootstrap variance estimator performs well, but Woodruff's variance estimator has a large relative bias (except in the case of no missing data) which results in a large MSE. In the case of no missing data, the poor performance of Woodruff's variance estimator was noticed and explained by Kovar, Rao and Wu (1988) and Sitter (1992). In our simulation

study, Woodruff's variance estimator behaves well when there is no missing data, but performs poorly when there are imputed data; the relative bias of Woodruff's variance estimator increases as the response rate decreases.

2. Confidence interval. In the case of no missing data, both confidence intervals have coverage probability about 93%; Woodruff's interval is slightly better and has slightly shorter length. When there are imputed data, Woodruff's interval is slightly longer than the bootstrap percentile confidence interval (which is related to the large positive bias of Woodruff's variance estimator); the coverage probability of the bootstrap percentile confidence interval is around 93-94%, whereas the coverage probability of Woodruff's interval is around 96-98%. The coverage error of Woodruff's interval is not serious in this simulation study, but the problem could be much worse if the coverage probability of Woodruff's interval were 95% in the case of no missing data or if the bias of Woodruff's variance estimator were negative instead of positive.

Acknowledgement

The research was partially supported by National Sciences Foundation Grant DMS-9504425 and National Security Agency Grant MDA904-96-1-0066.

Appendix

Proof of Lemma 1. The proof is very similar to that of Lemma 1 in Chen and Shao (1996). Let E_{*I} be the expectation under the bootstrap imputation. Then

$$E_{*I}(\hat{F}_\nu^{*I}) = G_\nu^{*R}/q_\nu^{*R} = \hat{F}_\nu^{*R},$$

where

$$G_\nu^{*R} = \frac{1}{M}\sum_{S_R^*} w_{hik}^* I_{\{y_{hik}^*\}},$$

$q_\nu^{*R} = G_\nu^{*R}(\infty)$, and $\hat{F}_\nu^{*R} = G_\nu^{*R}/q_\nu^{*R}$. Define

$$H_\nu^{*I}(x) = \hat{F}_\nu^{*I}(x) - \hat{F}_\nu^{*I}(\theta_\nu) - \hat{F}_\nu^{*R}(x) + \hat{F}_\nu^{*R}(\theta_\nu)$$

and

$$H_\nu^{*R}(x) = G_\nu^{*R}(x) - G_\nu^{*R}(\theta_\nu) - q_\nu^{*R}[\hat{F}_\nu^{R}(x) - \hat{F}_\nu^{R}(\theta_\nu)].$$

Then $E_*[H_\nu^{*R}(x)] = 0$. Following the proof of Lemma 1 in Chen and Shao (1996), we can show that

$$\sup_{|x-\theta_\nu|\le cn^{-1/2}} |H_\nu^{*I}(x) + H_\nu^{*R}(x)/q_\nu^{*R}| = o_p(n^{-1/2})$$

and

$$\sup_{|x-\theta_\nu|\le cn^{-1/2}} |H_\nu^I(x)| = o_p(n^{-1/2}).$$

Hence, the result follows from $H_\nu^*(x) = H_\nu^{*I}(x) + H_\nu^{*R}(x)/q_\nu^{*R} - H_\nu^I(x)$.

Proof of Theorem 3. The proof parallels that of Theorem 1 in Ghosh et al. (1984), but some modifications are needed to take into account the stratified multistage sampling and the imputation. From (8) and $\hat{\theta}_\nu^I - \theta_\nu = O_p(n^{-1/2})$, it suffices to show that

$$E_*|\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu)|^{2+\delta} = (2+\delta)\int_0^\infty t^{1+\delta} P_*\{\sqrt{n}|\hat{\theta}_\nu^{*I} - \theta_\nu| > t\}dt = O_p(1) \quad (17)$$

for a $\delta > 0$, where E_* and P_* are the expectation and probability with respect to bootstrap sampling. Note that

$$\frac{|\hat{\theta}_\nu^{*I}|^\epsilon}{n^{1+\epsilon}} \le \frac{1}{n^{1+\epsilon}} \max_{h,i,k} |y_{hik}|^\epsilon \le \frac{1}{n^{1+\epsilon}} \sum_{h=1}^L \sum_{i=1}^{n_h} \max_{k=1,\ldots,n_{hi}} |y_{hik}|^\epsilon = o_p(1)$$

under condition (12). Hence $P\{P_*\{\sqrt{n}|\hat{\theta}_\nu^{*I} - \theta_\nu| \le b_n\} = 1\} \to 1$, where $b_n = n^{1/2+(1+\epsilon)/\epsilon}$. Therefore, (17) follows from

$$\int_1^{b_n} t^{1+\delta} P_*\{\sqrt{n}|\hat{\theta}_\nu^{*I} - \theta_\nu| > t\}dt = O_p(1). \quad (18)$$

From (A1), there is a constant $c_1 > 0$ such that

$$P\Big\{n \sup_x \mathrm{Var}_*[G_\nu^{*R}(x) - q_\nu^{*R}\hat{F}_\nu^R(x)] \le c_1\Big\} \to 1.$$

Also, (A1) implies the existence of a constant $c_2 > 0$ such that $\sup_\nu \max_{h,i,k} w_{hik}/M \le c_2^2$. Let $c_\epsilon = \frac{1}{2}[\frac{1}{2} + (1+\epsilon)/\epsilon](2+\delta)\max\{c_1, c_2\}$ and $a_n = c_0\sqrt{c_\epsilon \log n}$, where c_0 is a constant satisfying $\inf_\nu f_\nu(\theta_\nu) > 4p_R^{-1}c_0^{-1}$. By (14),

$$a_n^{-1}n^{1/2}[\hat{F}_\nu^R(\theta_\nu + a_n n^{-1/2}) - \hat{F}_\nu^R(\theta_\nu) - F_\nu(\theta_\nu + a_n n^{-1/2}) + F_\nu(\theta_\nu)] = o_p(1);$$

by (A2′), $\liminf_\nu a_n^{-1}n^{1/2}[F_\nu(\theta_\nu + a_n n^{-1/2}) - F_\nu(\theta_\nu)] > 4p_R^{-1}c_0^{-1}$ and $a_n^{-1}n^{1/2}[\hat{F}_\nu^R(\theta_\nu) - p] = o_p(1)$. Hence $P\{\hat{F}_\nu^R(\theta_\nu + a_n n^{-1/2}) - p \ge 4p_R^{-1}c_0^{-1}a_n n^{-1/2}\} \to 1$. Thus, in the rest of the proof we may assume

$$n \sup_x \mathrm{Var}_*[G_\nu^{*R}(x) - q_\nu^{*R}\hat{F}_\nu^R(x)] \le c_1 \quad (19)$$

and

$$\hat{F}_\nu^R(\theta_\nu + a_n n^{-1/2}) - p \ge 4p_R^{-1}c_0^{-1}a_n n^{-1/2}. \quad (20)$$

Define

$$B_\nu^* = \{q_\nu^{*R}|(\hat{F}_\nu^{*R} - \hat{F}_\nu^R)(\theta_\nu + a_n n^{-1/2})| \geq c_0^{-1} a_n n^{-1/2}\} \cup \{q_\nu^* \leq p_R/2\}.$$

Using Bernstein's and Hoeffding's inequalities, together with (19), (A1), (A2′) and the choice of c_ϵ,

$$P_*(B_\nu^*) = O_p(b_n^{-(2+\delta)}).$$

By (20), on the complement of the set B_ν^*, $q_\nu^{*R}[\hat{F}_\nu^{*I}(\theta_\nu + a_n n^{-1/2}) - p] \geq c_0^{-1} a_n n^{-1/2}$. Hence, by Hoeffding's inequality again,

$$\begin{aligned} P_{*I}\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) \geq a_n\} &= P_{*I}\{p \geq \hat{F}_\nu^{*I}(\theta_\nu + a_n n^{-1/2})\} \\ &\leq \exp\{-2n[q_\nu^{*R}(\hat{F}_\nu^{*R}(\theta_\nu + a_n n^{-1/2}) - p)]^2/c_2\} + I_{B_\nu^*} \\ &\leq \exp\{-2c_\epsilon \log n/c_2\} + I_{B_\nu^*}, \end{aligned}$$

where P_{*I} is the probability with respect to the bootstrap imputation. Therefore,

$$\begin{aligned} \int_{a_n}^{b_n} t^{1+\delta} P_*\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) > t\}dt &= E_* \int_{a_n}^{b_n} t^{1+\delta} P_{*I}\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) > t\}dt \\ &\leq E_*[b_n^{2+\delta} P_{*I}\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) > a_n\}] \\ &\leq b_n^{2+\delta}[\exp\{-2c_\epsilon \log n/c_2\} + P_*(B_\nu^*)] \\ &= O_p(1). \end{aligned} \tag{21}$$

Similarly, it can be shown that

$$\int_{a_n}^{b_n} t^{1+\delta} P_*\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) < -t\}dt = O_p(1). \tag{22}$$

Let $t_\nu = \theta_\nu + tn^{-1/2}$. Then

$$\begin{aligned} \int_1^{a_n} t^{1+\delta} P_*\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) > t\}dt &= \int_1^{a_n} t^{1+\delta} P_*\{p \geq \hat{F}_\nu^{*I}(t_\nu)\}dt \\ &\leq \int_1^{a_n} t^{1+\delta} P_*\{|F_\nu(t_\nu) - \hat{F}_\nu^{*I}(t_\nu)| \geq F_\nu(t_\nu) - p\}dt \\ &\leq \int_1^{a_n} \frac{t^{1+\delta}}{[F_\nu(t_\nu) - p]^4} E_*|F_\nu(t_\nu) - \hat{F}_\nu^{*I}(t_\nu)|^4 dt \\ &\leq O_p(n^{-2}) \int_1^{a_n} \frac{t^{1+\delta}}{t^4 n^{-2}} dt, \end{aligned}$$

where the last inequality follows from (A2′) and the fact that (A1) and (A2′) imply

$$\sup_{1 \leq t \leq a_n} E_*|F_\nu(t_\nu) - \hat{F}_\nu^{*I}(t_\nu)|^4 = O_p(n^{-2}). \tag{23}$$

By choosing $\delta \in (0,1)$, we obtain

$$\int_1^{a_n} t^{1+\delta} P_*\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) > t\} dt = O_p(1). \tag{24}$$

Similarly,

$$\int_1^{a_n} t^{1+\delta} P_*\{\sqrt{n}(\hat{\theta}_\nu^{*I} - \theta_\nu) < -t\} dt = O_p(1). \tag{25}$$

Hence (18) follows from (21)-(25). This completes the proof.

Proof of Theorem 4. From (11) and $\hat{\rho}_\nu^I - \rho_\nu = O_p(n^{-1/2})$, we only need to show

$$E_*|\sqrt{n}(\hat{\rho}_\nu^{*I} - \rho_\nu)|^{2+\delta} = O_p(1) \tag{26}$$

for a $\delta > 0$. From (A2$'$), there is a constant $c > 0$ such that $\sqrt{n}|F_\nu(\frac{1}{2}\hat{\mu}_\nu^{*I}) - \rho_\nu| \geq t$ implies $\sqrt{n}|\hat{\mu}_\nu^{*I} - \mu_\nu| \geq ct$ for all ν and $0 < t \leq \sqrt{\log n}$. Noting that $0 \leq F_\nu \leq 1$, we obtain

$$\begin{aligned} E_*|\sqrt{n}[F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \rho_\nu]|^{2+\delta} &= (2+\delta)\int_0^{\sqrt{2n}} t^{1+\delta} P_*\{\sqrt{n}|F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \rho_\nu)| \geq t\} dt \\ &\leq (2n)^{1+\delta/2} P_*\{\sqrt{n}|F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \rho_\nu)| \geq \sqrt{\log n}\} \\ &\quad + (2+\delta)\int_0^{\sqrt{\log n}} P_*\{\sqrt{n}|F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I}) - \rho_\nu)| \geq t\} dt \\ &\leq (2n)^{1+\delta/2} P_*\{\sqrt{n}|\hat{\mu}_\nu^{*I} - \mu_\nu| \geq c\sqrt{\log n}\} \\ &\quad + (2+\delta)\int_0^{\sqrt{\log n}} P_*\{\sqrt{n}|\hat{\mu}_\nu^{*I} - \mu_\nu| \geq ct\} dt \\ &= O_p(1) \end{aligned} \tag{27}$$

by the proof of Theorem 3. Similarly,

$$\begin{aligned} E_*|\sqrt{n}[\hat{\rho}_\nu^{*I} - F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I})]|^{2+\delta} &\leq n^{1+\delta/2} P_*\{\sqrt{n}|\hat{\mu}_\nu^{*I} - \mu_\nu| \geq \sqrt{\log n}\} \\ &\quad + E_*|\sqrt{n}[\hat{\rho}_\nu^{*I} - F_\nu(\tfrac{1}{2}\hat{\mu}_\nu^{*I})]|^{2+\delta} I_{\{\sqrt{n}|\hat{\mu}_\nu^{*I} - \mu_\nu| \geq \sqrt{\log n}\}} \\ &= O_p(1) + E_* \sup_{x \in D_\nu} |\sqrt{n}[\hat{F}_\nu^{*I}(x) - F_\nu(x)]|^{2+\delta}, \end{aligned} \tag{28}$$

where $D_\nu = \{x : |x - \frac{1}{2}\mu_\nu| \leq n^{-1/2}\sqrt{\log n}\}$. Hence (26) follows from (27), (28) and

$$E_* \sup_{x \in D_\nu} |\sqrt{n}[\hat{F}_\nu^{*I}(x) - F_\nu(x)]|^{2+\delta} = O_p(1). \tag{29}$$

Since $E_*|\sqrt{n}[\hat{F}_\nu^{*I}(\frac{1}{2}\mu_\nu) - \rho_\nu]|^{2+\delta} = O_p(1)$, (29) follows from (14) and

$$E_* \sup_{x \in D_\nu} |\sqrt{n}[H_\nu^*(x)]|^{2+\delta} = O_p(1), \tag{30}$$

where $H_\nu^*(x)$ is defined by (6) with $\theta_\nu = \frac{1}{2}\mu_\nu$. From the proof of Lemma 1 in the Appendix, (30) follows from

$$E_* \max_{-n\le\ell\le n} |\sqrt{n}H_\nu^{*R}(\eta_{\nu\ell})|^{2+\delta} = O_p(1) \tag{31}$$

and

$$E_* \max_{-n\le\ell\le n} |\sqrt{n}H_\nu^{*I}(\eta_{\nu\ell})|^{2+\delta} = O_p(1), \tag{32}$$

where $\eta_{\nu\ell} = \frac{1}{2}\mu_\nu + n^{-3/2}\sqrt{\log n}\ell$, $\ell = -n,\ldots,n$, $H_\nu^{*R}(x)$ and $H_\nu^{*I}(x)$ are defined in the proof of Lemma 1. Using Bernstein's inequality, we obtain that

$$\begin{aligned}
&E_* \max_{-n\le\ell\le n} |\sqrt{n}H_\nu^{*R}(\eta_{\nu\ell})|^{2+\delta}\\
&= (2+\delta)\int_0^\infty t^{1+\delta} P_*\Big\{\max_{-n\le\ell\le n} |\sqrt{n}H_\nu^{*R}(\eta_{\nu\ell})| \ge t\Big\}dt\\
&\le 4(2+\delta)n\int_0^\infty t^{1+\delta}\exp\Big\{-\frac{t^2}{O_p(n^{-1/2})(\sqrt{\log n}+t)}\Big\}dt\\
&= o_p(1).
\end{aligned}$$

Therefore, (31) holds. (32) can be similarly shown. The proof is completed.

References

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* **12**, 470-482.

Chen, Y. and Shao, J. (1996). Inference with complex survey data imputed by hot deck when nonrespondents are nonidentifiable. Technical Report, Department of Statistics, University of Wisconsin-Madison.

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.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42**, 1957-1961.

Ghosh, M., Parr, W. C., Singh, K. and Babu, G. J. (1984). A note on bootstrapping the sample median. *Ann. Statist.* **12**, 1130-1135.

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**, Supplement, 25-45.

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

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

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

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550-1566.

Shao, J. (1996). Resampling methods in sample surveys (with discussions). *Statist.* **27**, 203-254.

Shao, J. and Rao, J. N. K. (1994). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya Ser. B*, Special Volume **55**, 393-414.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.* **91**, 1278-1288.

Shao J. and Wu, C. F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Ann. Statist.* **20**, 1571-1593.

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

Wolfson, M. C. and Evans, J. M. (1990). Statistics Canada low income cut-offs, methodological concerns and possibilities. Discussion Paper, Statistics Canada.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47**, 635-646.

Department of Statistics, University of Wisconsin-Madison, 1210 West Dayton Street, Madison WI 53706-1685, U.S.A.

E-mail: shao@stat.wisc.edu

(Received June 1996; accepted June 1997)