Survey Methodology, December 1992
Vol. 18, No. 2, pp. 209–217
Statistics Canada

Some Recent Work on Resampling Methods for Complex Surveys

J.N.K. RAO, C.F.J. WU and K. YUE[1]

ABSTRACT

Resampling methods for inference with complex survey data include the jackknife, balanced repeated replication (BRR) and the bootstrap. We review some recent work on these methods for standard error and confidence interval estimation. Some empirical results for non-smooth statistics are also given.

KEY WORDS: Balanced repeated replication; Bootstrap; Jackknife; Stratified multistage designs; Variance estimation.

1. INTRODUCTION

Standard sampling theory is largely devoted to estimation of mean square error (MSE) of unbiased or approximately unbiased estimators $\hat{Y}$ of a population total Y. An estimator of MSE, or a variance estimator, provides us with a measure of uncertainty in the estimator $\hat{Y}$. It is a common practice to assume that the estimator $\hat{Y}$ is approximately normally distributed and then use a two-sided confidence interval $\hat{Y} \pm z_{\alpha/2} s(\hat{Y})$ or a one-sided confidence interval $(\hat{Y} - z_\alpha s(\hat{Y}), \infty)$ or $(-\infty, \hat{Y} + z_\alpha s(\hat{Y}))$, where $s(\hat{Y})$ is the standard error of $\hat{Y}$ (*i.e.*, square root of estimated MSE) and z_α is the upper α-point of a $N(0, 1)$ variable. These intervals cover the true total Y with a probability of approximately $1 - \alpha$ in large samples, but the actual coverage probability could be significantly lower than $1 - \alpha$ in small samples or in highly clustered samples. For nonlinear statistics, such as ratios, regression or correlation coefficients, the well-known linearization (or Taylor expansion) method is often used (see Rao 1988 for detailed applications). Resampling methods, such as the jackknife, balanced repeated replication (BRR) and the bootstrap, are also being used, and in fact several agencies in the U.S.A and Canada have adopted the jackknife method of variance estimation for stratified multistage surveys. An advantage of the linearization method is that it is applicable to general sampling designs, but involves the derivation of a separate standard error formula, $s(\hat{\theta})$, for each nonlinear statistic, $\hat{\theta}$. On the other hand, resampling methods employ a single standard error formula for all statistics $\hat{\theta}$. However, the jackknife and the BRR methods are strictly applicable only to those stratified multistage designs in which clusters within strata are sampled with replacement or the first-stage sampling fraction is negligible. The bootstrap method of Rao and Wu (1987) works for more general designs, but it is computationally cumbersome and its properties for complex designs have not been fully investigated.

This paper provides an account of some recent work on resampling methods for complex surveys. Some empirical results on jackknife and bootstrap variance estimation for non-smooth statistics, such as the median, under stratified cluster sampling and stratified simple random sampling are also given.

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6. C.F.J. Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1. Kim Yue, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

2. STRATIFIED MULTISTAGE SAMPLING

Large-scale surveys often employ stratified multistage designs with large numbers of strata, L, and relatively few primary sampling units (clusters), $n_h(\geq 2)$, sampled within each stratum h. In fact, it is quite common to select $n_h = 2$ clusters within each stratum to permit maximum degree of stratification of clusters consistent with the provision of a valid variance estimator. We assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals Y_{hi}, $i = 1, \ldots, n_h$; $h = 1, \ldots, L$.

Let $w_{hik}(> 0)$ be the survey weight attached to the k-th sample element (ultimate unit) in the i-th sample cluster belonging to h-th stratum. Often, the basic weights w_{hik} are subjected to post-stratification adjustment to ensure consistency with known totals of post-stratification variables. For example, the Canadian Labour Force Survey uses a generalized regression estimator to ensure consistency. We shall, however, ignore this complication in the present paper. An estimator of the population total Y is of the form

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}, \tag{2.1}$$

where s denotes the sample of elements and y_{hik} is the value of a characteristic of interest, y, associated with the sample element $(hik)\in s$. We assume complete response on all items.

It is a common practice to sample the clusters with probabilities proportional to sizes (pps) and without replacement to increase the efficiency of the estimators compared to pps sampling with replacement and to avoid the possibility of selecting the same cluster more than once in the sample. However, at the stage of variance estimation the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement and subsampling done independently each time a cluster is selected. This approximation leads to overestimation of variance of $\hat{Y}$, but the relative bias is likely to be small if the first stage sampling fraction is small in each stratum.

Writing $\hat{Y}$ as

$$\hat{Y} = \sum_{h=1}^{L} \bar{r}_h, \tag{2.2}$$

with

$$r_{hi} = \sum_k (n_h w_{hik}) y_{hik}, \quad \bar{r}_h = \sum_i r_{hi}/n_h,$$

we note that the r_{hi} are independent and identically distributed (iid) random variables with the same mean, Y_h, and the same variance in each stratum h, under with replacement sampling of clusters. It therefore follows that an unbiased estimator of variance of $\hat{Y}$ is given by

$$s^2(\hat{Y}) = \sum_h s_{rh}^2/n_h, \tag{2.3}$$

with

$$(n_h - 1)s_{rh}^2 = \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2.$$

Under without-replacement sampling of clusters, $s^2(\hat{Y})$ will overestimate the true variance of $\hat{Y}$.

We are also often interested in estimating the population distribution function, $F(t)$, and the p-th quantile, $\theta = F^{-1}(p)$, $0 < p < 1$; in particular, the population median $\theta = F^{-1}(1/2)$. The survey estimator of $F(t)$ is given by

$$\hat{F}(t) = \sum_{(hik)\epsilon s} \tilde{w}_{hik} a_{hik}, \tag{2.4}$$

where $\tilde{w}_{hik} = w_{hik}/\sum_s w_{hik}$ are the normalized weights ($\sum_s \tilde{w}_{hik} = 1$) and $a_{hik} = 1$ if $y_{hik} \le t$, $a_{hik} = 0$ otherwise. The sample p-th quantile is obtained as

$$\hat{\theta} = \hat{F}^{-1}(p). \tag{2.5}$$

In practice, $\hat{\theta}$ is computed by first arranging the sampled values y_{hik} in an ascending order, say $\{y_{(hik)}\}$, and then cumulating the associated weights $\tilde{w}_{hik}$ until p is first crossed. The first $y_{(hik)}$ encountered after crossing p is taken as the sample p-th quantile, $\hat{\theta}$. Woodruff (1952) obtained confidence intervals for a quantile, and Rao and Wu (1987) obtained a simple variance estimator using Woodruff's interval (see also Kovar, Rao and Wu 1988, Francisco and Fuller 1991). Shao (1991) considered general L-statistics, including the sample Lorenz curve and the Gini coefficient, which are examples of smooth L-statistics, and the sample quantiles which are examples of non-smooth L-statistics.

Many nonlinear parameters of interest, such as population means, ratios, regression and correlation coefficients, can be expressed as smooth functions, $\theta = g(Y)$, of a vector of totals, $Y = (Y_1, \ldots, Y_q)'$, of suitably defined variates. An estimator of θ is given by $\hat{\theta} = g(\hat{Y})$. The linearization method may be used to estimate the variance of $g(\hat{Y})$, under any complex design (see Binder 1983 and Rao 1988).

3. RESAMPLING METHODS

Resampling methods, such as the jackknife and the bootstrap, are widely used in the iid case. Suitable modification/extensions of these methods have also been developed to handle survey data involving stratification and clustering. We now give a brief account of some recent work on three such methods: jackknife, balanced repeated replication and bootstrap, in the context of stratified multistage sampling.

3.1 Jackknife

For simplicity, assume $\hat{\theta} = g(\hat{Y})$, a smooth function of the estimated total $\hat{Y}$. Let $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$ be the estimator of θ obtained from the sample after omitting the data from the j-th sampled cluster in g-th stratum ($j = 1, \ldots, n_g$; $g = 1, \ldots, L$), where

$$\hat{Y}_{(gj)} = \sum_{\substack{(hik)\epsilon s \\ h \ne g}} w_{hik} y_{hik} + \sum_{\substack{(gik)\epsilon s \\ i \ne j}} \left\{\frac{n_g}{n_g - 1} w_{gik}\right\} y_{gik}. \tag{3.1}$$

Note that $\hat{Y}_{(gj)}$ is obtained by changing the weight of (gik)-th element to $n_g w_{gik}/(n_g - 1)$, $i \ne j$, but retaining the original weights, w_{hik}, for $h \ne g$. A customary delete-1 cluster jackknife variance estimator of $\hat{\theta}$ is given by

$$s_J^2(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2. \tag{3.2}$$

Two variations of $s_J^2(\hat{\theta})$ are obtained by changing $\hat{\theta}$ in (3.2) to $\hat{\theta}_{(g.)} = \sum_j \hat{\theta}_{(gj)}/n_g$ and $\hat{\theta}_{(..)} = \sum_g \sum \hat{\theta}_{(gj)}/n$, where $n = \sum_g n_g$. In the linear case, $\hat{\theta} = \hat{Y}$, all the jackknife variance estimators reduce to the "correct" variance estimator, $s^2(\hat{Y})$, given by (2.3). Rao and Wu (1987) made a second order analysis of the resampling variance estimators when $\hat{\theta}$ is expressed as a smooth function of totals, $\hat{Y}$. Their main results on the jackknife are: (1) Different jackknife variance estimators are asymptotically equal to higher order terms, as the number of strata, L, increases. (2) In the important case of $n_h = 2$ for all h, the linearization variance estimator, $s_L^2(\hat{\theta})$, and any jackknife variance estimator are asymptotically equal to higher order terms, indicating that the choice between the two methods should depend more on operational considerations than on statistical criteria.

A drawback of the customary delete-1 jackknife method in the case of independent and identically distributed (i.i.d.) observations is that, unlike the bootstrap, it fails to provide a consistent variance estimator for non-smooth statistics, such as the median. Shao and Wu (1989), however, have shown that this deficiency of the delete-1 jackknife can be rectified by using a more general jackknife, called the delete-d jackknife, with the number of observations deleted, d, depending on a smoothness measure of the statistic. In particular, for the sample quantiles, the delete-d jackknife with d satisfying $n^{1/2}/d \rightarrow 0$ and $n - d \rightarrow \infty$ as $n \rightarrow \infty$ leads to consistent variance estimators in the case of i.i.d. observations. This result suggests that a similar effect might hold in the case of delete-1 cluster jackknife for stratified multistage sampling since all the sampled elements in a sampled cluster (gj) are deleted in computing $s_J^2(\hat{\theta})$ given by (3.2). At present we are studying this problem theoretically, but we performed a limited simulation study which suggests that the delete-1 cluster jackknife variance estimator $s_J^2(\hat{\theta})$ might perform quite well. We now report the results of the simulation study for the median, $\hat{\theta} = \hat{F}^{-1}(1/2)$.

For the simulation study, we generated stratified cluster samples $\{y_{hik}, k = 1, \ldots, M; i = 1, \ldots, n_h; h = 1, \ldots, L\}$ employing the nested error model $y_{hik} = \mu_h + a_{hi} + e_{hik}$ with $a_{hi} \overset{iid}{\sim} N(0,\sigma_{ah}^2)$ and $e_{hik} \overset{iid}{\sim} N(0,\sigma_{eh}^2)$, where the cluster size, M is assumed to be equal for all clusters (hi), and the intra-cluster correlations, $\sigma_{ah}^2/(\sigma_{ah}^2 + \sigma_{eh}^2) = \rho_h$, are assumed to be equal for all strata h (*i.e.*, $\rho_h = \rho$). The normalized survey weights are given by $\tilde{w}_{hik}$ with $w_{hik} = W_h/(n_h M)$ and W_h denotes the relative size of stratum h. The number of strata $L (= 32)$, strata means, μ_h, variances $\sigma_h^2 = \sigma_{ah}^2 + \sigma_{eh}^2$ and sizes W_h were chosen to correspond to real populations encountered in the US National Assessment of Educational Progress Study (Hansen and Tepping 1985). We generated 1,000 independent stratified cluster samples with $n_h = 2$ for each selected combination (ρ,M) and then computed the bias and relative bias of the jackknife variance estimator, $s_J^2(\hat{\theta})$, for the median: $\text{Bias}[s_J^2(\hat{\theta})] = \sum_t s_{Jt}^2(\hat{\theta})/1{,}000 - \text{MSE}(\hat{\theta})$, where $s_{Jt}^2(\hat{\theta})$ is the value of $s_J^2(\hat{\theta})$ for the t-th simulated sample $(t = 1, \ldots, 1{,}000)$ and Rel. $\text{Bias}[s_J^2(\hat{\theta})] = \text{Bias}[s_J^2(\hat{\theta})]/\text{MSE}(\hat{\theta})$. We calculated $\text{MSE}(\hat{\theta})$ from an independent set of 10,000 stratified cluster samples for each (ρ,M): $\text{MSE}(\hat{\theta}) = \sum_t (\hat{\theta}_t - \hat{\theta}.)^2/10{,}000$, where $\hat{\theta}_t$ is the value of $\hat{\theta}$ for the t-th simulated sample, $\hat{\theta}. = \sum \hat{\theta}_t/10{,}000$ and $t = 1, \ldots, 10{,}000$.

Table 1 reports the simulated values of bias and relative bias (in brackets) of the jackknife variance estimator for selected combinations of ρ and M. First, we note that for the special case of stratified simple random sampling $(\rho = 0, M = 1)$, the relative bias is very large (116%) thus confirming the inconsistency of $s_J^2(\hat{\theta})$ in this case. Second, we observe that both the bias and relative bias decrease as M increases for a given ρ. Moreover, for a given cluster

Table 1

Bias and % Relative Bias (in Brackets) of Jackknife Variance Estimator for the Median Under Stratified Cluster Sampling ($n_h = 2, L = 32$) and Selected Values of Equal Intra-Cluster Correlation, ρ, and Equal Cluster Size, M

ρ	M				
	1	10	20	30	50
0	7.5(116)	.28(41)	.09(29)	.04(15)	.01(15)
0.05	-	.22(27)	.09(18)	.05(12)	.03 (8)
0.10	-	.28(28)	.10(14)	.06 (9)	.02 (3)
0.20	-	.31(22)	.11(10)	.08 (8)	.03 (3)
0.30	-	.32(18)	.11 (7)	.07 (5)	.01 (1)
0.50	-	.44(17)	.15 (6)	.11 (5)	.04 (2)

size M, the bias generally increases with ρ, but the relative bias in fact decreases because MSE $(\hat{\theta})$ is increasing faster than the bias as ρ increases. It is indeed gratifying that the relative bias is no more than 10% for $M \geq 30$ and $\rho \geq 0.10$ or $M \geq 20$ and $\rho \geq 0.20$.

3.2 Balanced Repeated Replication (BRR)

Balanced repeated replication (BRR) was proposed by McCarthy (1969) for the important special case of $n_h = 2$ clusters per stratum. A set of R balanced half-samples (replications) is formed by deleting one cluster from the sample in each stratum. This set may be defined by a $R \times L$ design matrix (δ_h^r), $1 \leq r \leq R$, $1 \leq h \leq L$ with $\delta_h^r = +1$ or -1 according as whether the first or second sample cluster in the h-th stratum is in the r-th half-sample, and $\sum_r \delta_h^r \delta_{h'}^r = 0$ for all $h \neq h'$, *i.e.* the columns of the matrix are orthogonal. A minimal set of R balanced half-samples may be constructed from Hadamard matrices ($L + 1 \leq R \leq L + 4$) by choosing any L columns, excluding the column of + 1's.

Let $\hat{\theta}^{(r)}$ be the estimator of θ obtained from the r-th half-sample. Note that $\hat{\theta}^{(r)}$ is obtained from $\hat{\theta}$ by changing the weight of (hik)-th element to $2w_{hik}$ or 0 according as the (hi)-th cluster is selected or not selected in the half-sample. A BRR variance estimator of $\hat{\theta}$ is given by

$$s_{\mathrm{BRR}}^2(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2. \tag{3.3}$$

Several variations of $s_{\mathrm{BRR}}^2(\hat{\theta})$ are also available; for example, $\hat{\theta}$ may be changed to $\hat{\theta}(\cdot) = \sum_r \hat{\theta}^{(r)}/R$. In the linear case, $\hat{\theta} = \hat{Y}$, all the BRR variance estimators reduce to the "correct" variance estimator, $s^2(\hat{Y})$, as in the case of the jackknife.

Krewski and Rao (1981) established the consistency of $s_J^2(\hat{\theta})$ and $s_{\mathrm{BRR}}^2(\hat{\theta})$ for smooth statistics $\hat{\theta} = g(\hat{Y})$, as L increases. Rao and Wu (1985) made a second order analysis and showed that $s_{\mathrm{BRR}}^2(\hat{\theta})$ and $s_L^2(\hat{\theta})$ are not asymptotically equivalent to second order terms, unlike $s_J^2(\hat{\theta})$ and $s_L^2(\hat{\theta})$. Shao and Wu (1992) established the consistency of $s_{\mathrm{BRR}}^2(\hat{\theta})$ for the quantiles, $\hat{\theta} = \hat{F}^{-1}(p)$.

The BRR method has been extended to the case of $n_h = p > 2$ clusters per stratum for p prime or power of prime (Gurney and Jewett 1975), but the number of replications, R, needed is much larger than in the case of $n_h = 2$. In many survey designs n_h's are not equal. To accommodate the general case of unequal n_h, Gupta and Nigam (1987) and Wu (1991) advocated the use of mixed-level orthogonal arrays of strength two for drawing balanced replicates, where n_h is the number of symbols in the h-th column of the array. Orthogonality of the array guarantees that the replicates drawn are balanced. Unlike the case of equal n_h, the adjustment of survey weights is more complicated. A correct method was given by Wu (1991). From his formula (6), two separate adjustments should be applied to the sampled and unsampled units in each replicate. Simple algebra on Wu's equation (6) shows that w_{hik} is changed to $w'_{hik} = [1 + (n_h - 1)^{½}]w_{hik}$ or $w''_{hik} = [1 - (n_h - 1)^{½}]w_{hik}$ according as the (hik)-th element is selected or not selected in the replicate. (Note that $w'_{hik} = 2$ and $w''_{hik} = 0$ for $n_h = 2$). The remaining calculation of $\hat{\theta}^{(r)}$ and $s^2_{\mathrm{BRR}}(\hat{\theta})$ are the same as in (3.3). Furthermore, these modified survey weights can be applied to $\hat{\theta} = \hat{F}^{-1}(p)$ and more general $\hat{\theta} = T(\hat{F})$, where T is a functional of $\hat{F}$. All we need to do is to change w_{hik} in (2.4) to w'_{hik} or w''_{hik} according as the (hik)-th element is selected or not selected in the r-th replicate to get $\hat{F}^{(r)}$ of F for the r-th replicate, and $\hat{\theta}^{(r)} = T(\hat{F}^{(r)})$. The calculation of the BRR variance estimator is the same as in (3.3).

There are two problems with the use of mixed orthogonal arrays. First, the array size can be large for general n_h. Second, orthogonal arrays do not exist for any combination of n_h's. A practical solution is to group the n_h sample psu's in stratum h into two to four groups of psu's and then apply the method to the groups by treating the groups as units in the BRR method. This extension is called the grouped BRR method. As shown by Wu (1991), its efficiency loss can be relatively small, compared to the full BRR, if the groupings are done judiciously. For example, more groups are needed if n_h is large and the units within the stratum are more heterogeneous. For $n_h = 2$, 3 or 4, many mixed orthogonal arrays have been constructed (see, for example, Dey 1985 and Wang and Wu 1991). If n_h can only take 2 or 4, saturated orthogonal arrays for any combination can be easily constructed as in Wu (1989). That is, the number of replications can be as small as possible. It is therefore possible to compile a large collection of mixed orthogonal arrays for practical use if n_h is restricted to 2, 3 or 4.

The BRR method and extensions considered thus far only take one unit (psu) per stratum for each replicate. If n_h is large, say more than 3, Sitter (1992) proposed the use of orthogonal multi-arrays to allow the number of resampled units per stratum to be greater than one. It may require fewer replicates and it can cover cases where orthogonal arrays of strength two are not available; for example, $n_h = 6$.

3.3 Bootstrap

The bootstrap method for the iid case has been extensively studied (Efron 1982). Rao and Wu (1987) provided an extension to stratified multistage designs, but covering only smooth statistics $\hat{\theta} = g(\hat{Y})$. They required that, in order to have valid variance estimation in the case of small n_h, some scale adjustment, similar to those in Section 3.2, is necessary. What they did not realize is that the scale adjustment should be made on the survey weights w_{hik} rather on the y_{hik} values directly, which is what they proposed. As a result, their method cannot be extended to cover the quantile $\theta = F^{-1}(p)$. We now present a general method that covers smooth as well as non-smooth statistics for arbitrary sizes, n_h. It works as follows: (i) Draw a simple random sample of m_h clusters with replacement from the n_h sample clusters, independently for each h. Let m^*_{hi} be the number of times (hi)-th sample cluster is selected $(\sum_i m^*_{hi} = m_h)$. Define the bootstrap weights

$$w^*_{hik} = \left[\{1 - (m_h/(n_h - 1))^{½}\} + (m_h/(n_h - 1))^{½}(n_h/m_h)m^*_{hi}\right]w_{hik}. \tag{3.4}$$

If the (hi)-th cluster is not selected in the bootstrap sample, $m^*_{hi} = 0$ and the second term of (3.4) vanishes. If m_h is chosen to be less than or equal to $n_h - 1$, then the bootstrap weights w^*_{hik} are all positive if $w_{hik} > 0$ for all $(hik) \epsilon s$ Calculate θ^*, the bootstrap estimator of θ, using the weights w^*_{hik} in the formula for $\hat{\theta}$. The bootstrap median, for example, is calculated as before using the normalized bootstrap weights $\tilde{w}^*_{hik} = w^*_{hik} / \sum_s w^*_{hik}$, provided all $w^*_{hik} > 0$.
(ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimates $\theta^*_{(1)}, \ldots, \theta^*_{(B)}$.

The bootstrap variance estimator $s^2_{\mathrm{BOOT}}(\hat{\theta}) = E_*(\theta^* - E_*\theta^*)^2$, is approximated by

$$\tilde{s}^2_{\mathrm{BOOT}}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} [\theta^*_{(b)} - \hat{\theta}]^2. \tag{3.5}$$

A variation of (3.5) is obtained by changing $\hat{\theta}$ to $\theta^*_{(.)} = \sum_b \theta^*_{(b)}/B$. In the linear case, $\tilde{s}^2_{\mathrm{BOOT}}(\hat{\theta})$ reduces to the "correct" variance estimator $s^2(\hat{Y})$.

Rao and Wu (1987) obtained bootstrap-t confidence intervals for smooth functions, $\theta = g(Y)$, by approximating the distribution of $t = (\hat{\theta} - \theta)/s_J(\hat{\theta})$ by its bootstrap counterpart $t^* = (\theta^* - \hat{\theta})/s_J(\theta^*)$, where $s^2_J(\theta^*)$ is obtained from(3.2) with w_{hik} changed to w^*_{hik}. A two-sided $(1 - \alpha)$-level confidence interval for θ is then given by $\{\hat{\theta} - t^*_U s_J(\hat{\theta}), \hat{\theta} - t^*_L s_J(\hat{\theta})\}$, where t^*_L and t^*_U are the lower and upper $\alpha/2$-points of t^* obtained from the bootstrap histogram of $t^*_{(1)}, \ldots, t^*_{(B)}$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Empirical work by Kovar, Rao and Wu (1988) for smooth functions indicates that the bootstrap-t interval with $m_h = n_h - 1$ tracks the error rates in both the lower and upper tails better than the jackknife interval $\{\hat{\theta} - z_{\alpha/2} s_J(\hat{\theta}), \hat{\theta} + z_{\alpha/2} s_J(\hat{\theta})\}$, but the total error rate is not distinguishable from the latter, *i.e.*, for two-sided intervals, they exhibit similar performance in terms of actual coverage probability. If a variance stabilizing transformation can be found, such as the $\tanh^{-1}$ transformation on the estimated correlation coefficient, then the problem of uneven error rates in the two tails for the jackknife interval seems to be corrected. This suggests that the jackknife interval, or any other normal-theory interval, based on such transformations can be useful when the transformations are known, while the bootstrap provides an alternative when such transformations do not exist or are unknown.

We now present the results of a limited simulation study on the performance of the proposed bootstrap method in the case of the median. Employing the Hansen-Tepping basic population 1 with $L = 32$ strata (see Kovar *et al.* 1988, Sections 3 and 6 for details), we generated 500 independent stratified simple random samples with $n_h = 5$ and then computed the relative bias and coefficient of variation (relative stability) of the Woodruff-based variance estimator with $\alpha = 0.1$ (see Kovar *et al.* 1988, eq. (2.8)), the BRR variance estimator (3.3) and the bootstrap variance estimator (3.5) and its variation obtained by changing $\hat{\theta}$ to $\theta^*_{(.)}$. We used $m_h = n_h - 1$ and $n_h - 3$ and $B = 500$ bootstrap replicates for each sample, while the BRR replicates were obtained from an orthogonal array with 250 runs. The true MSE of $\hat{\theta}$ was approximated by selecting 10,000 independent stratified random samples. We also calculated the error rates in each tail (nominal rate of 5% in each tail) and standardized lengths of the normality-based confidence interval using the BRR variance estimator, the Woodruff interval and the bootstrap interval obtained from the percentile method using the bootstrap histogram of $\theta^*_{(1)}, \ldots, \theta^*_{(B)}$ for each sample.

Table 2 reports the simulated values of the relative bias, coefficient of variation, lower (L) and upper (U) error rates, and standardized lengths. First, we note that the bootstrap variance estimator (3.5) has a larger relative bias and a slightly larger coefficient of variation (CV) than

Table 2

% Relative Bias and % CV of Variance Estimator and Error Rates and Standardized Lengths of Confidence Intervals (Nominal Level of 5% in Each Tail) for the Median Under Stratified Simple Random Sampling $L = 32$, $n_h = 5$)

Method	% Rel. Bias	% CV	Error Rate		St. Length
			L	U	
Woodruff	4.2	47	4.2	5.6	0.997
BRR	3.1	31	5.0	5.0	1.004
Bootstrap*:					
$m_h = 4$	12.6 (7.5)	52 (48)	5.0	5.2	0.987
$m_h = 2$	13.0 (7.8)	54 (49)	5.0	4.8	0.988

* Results for the variation of the bootstrap variance estimator are given in the brackets.

its variation obtained by changing $\hat{\theta}$ to $\theta^*_{(\cdot)}$: Relative bias of 12.6% *vs.* 7.5% and CV of 52% *vs.* 48% for $m_h = n_h - 1 = 4$. On the other hand, the BRR variance estimator has the smallest relative bias (3.1%) and the smallest CV (31%), while the Woodruff-based variance estimator has a smaller relative bias (4.2%) and a comparable CV (47%). Secondly, the lower and upper error rates are close to the nominal level (5%) for the bootstrap and the BRR intervals, while the error rates are slightly uneven for the Woodruff interval ($L = 4.2\%$ and $U = 5.6\%$). Finally, we note that the standardized lengths are roughly equal for all the methods. Overall, the bootstrap variance estimator and the bootstrap intervals based on the percentile method did not exhibit better performance relative to either the BRR variance estimator and the associated normality-based interval or the Woodruff-based variance estimator and the Woodruff interval.

ACKNOWLEDGEMENT

J.N.K. Rao's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

DEY, A. (1985). *Orthogonal Fractional Factorial Designs.* New Delhi: Wiley Eastern.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

GURNEY, M., and JEWETT, R.S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*, 70, 819-821.

HANSEN, M., and TEPPING, B.J. (1985). Estimation for variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.

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

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics, Vol. 6*, (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: Elsevier Science, 427-447.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., and WU, C.F.J. (1987). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data. *Bulletin of the International Statistical Institute*.

SHAO, J. (1991). *L*-statistics in complex survey problems. Technical Report, University of Ottawa, Ottawa.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

SHAO, J., and WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics*, 20 (to appear).

SITTER, R.R. (1992). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, (to appear).

WANG, J.C., and WU, C.F.J. (1991). An approach to the construction of asymmetrical orthogonal arrays. *Journal of the American Statistical Association*, 86, 450-456.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other positional measures. *Journal of the American Statistical Association*, 47, 635-646.

WU, C.F.J. (1989). Construction of 2^m4^n designs via a grouping scheme. *Annals of Statistics*, 17, 1880-1885.

WU, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78, 181-188.