The Canadian Journal of Statistics
Vol. 20, No. 2, 1992, Pages 135–154
La Revue Canadienne de Statistique

Comparing three bootstrap methods for survey data

R.R. SITTER*

Carleton University

Key words and phrases: Bootstrap, jackknife, two-stage cluster sampling, probability sampling, Rao-Hartley-Cochran method, Edgeworth expansion.
AMS 1985 subject classifications: 62D05, 62G05.

ABSTRACT

Various bootstrap methods for variance estimation and confidence intervals in complex survey data, where sampling is done without replacement, have been proposed in the literature. The oldest, and perhaps the most intuitively appealing, is the without-replacement bootstrap (BWO) method proposed by Gross (1980). Unfortunately, the BWO method is only applicable to very simple sampling situations. We first introduce extensions of the BWO method to more complex sampling designs. The performance of the BWO and two other bootstrap methods, the rescaling bootstrap (Rao and Wu 1988) and the mirror-match bootstrap (Sitter 1992), are then compared through a simulation study. Together these three methods encompass the various bootstrap proposals.

RÉSUMÉ

Différentes variantes de la méthode du bootstrap ont été proposées afin d'estimer la variance et construire des intervalles de confiance dans le contexte de sondages complexes où l'échantillonnage se fait sans remise. La plus ancienne et probablement la plus naturelle est le bootstrap sans remise (BWO) proposé par Gross (1980). Malheureusement cette méthode n'est applicable qu'à des plans d'échantillonnage très simples. Nous proposons une généralisation de la méthode BWO à des plans d'échantillonnage plus complexes. Cette nouvelle méthode et deux autres variantes du bootstrap, proposées respectivement par Rao et Wu (1988) et Sitter (1992), sont comparées à l'aide de simulations. Ces trois méthodes englobent plusieurs des différentes variantes proposées.

1. INTRODUCTION

For most complex survey designs, unbiased variance estimators of statistics that are expressible as linear functions of the observations can be derived. For nonlinear statistics and functionals this is generally not the case. Various methods for obtaining variance estimates have been proposed in the literature. Three commonly used methods are the linearization (or Taylor) method, the jackknife method, and balanced repeated replications (BRR). Krewski and Rao (1981) show the asymptotic consistency of variance estimates for nonlinear functions of means based on these methods applied to multistage designs in which the primary sampling units are selected with replacement.

However, these methods have some limitations. The linearization method requires theoretical calculation and programming of derivatives. This can make it cumbersome

*R.R. Sitter is Assistant Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6. This research composed a portion of the author's Ph.D. thesis at the University of Waterloo, Waterloo, Ontario, under the supervision of C.F.J. Wu.

to implement. The jackknife, when the primary sampling units are selected without replacement, has only been developed for stratified sampling.

The BRR method is only applicable to stratified sampling with restrictions on the within stratum sample sizes (see Gurney and Jewett 1975, Gupta and Nigam 1987, and Wu 1992). In recognition of these limitations, various bootstrap procedures for variance estimation and confidence intervals in sample survey data have been proposed in the literature (Gross 1980, Bickel and Freedman 1984, McCarthy and Snowden 1985, Rao and Wu 1988, and Sitter 1992). Bootstrap methods reutilize the existing estimation system repeatedly, using computing power to avoid theoretical work. This article attempts to compare the performance of three bootstrap methods: the rescaling bootstrap (Rao and Wu 1988), the mirror-match bootstrap (Sitter 1992), and the without-replacement bootstrap (BWO) (Gross 1980, Bickel and Freedman 1984), in terms of variance estimation and confidence intervals in complex survey data where the sample is taken without replacement. Together these three methods encompass the various bootstrap proposals.

The three methods are outlined in Section 2, with the algorithms for stratified sampling without replacement explicitly stated. The BWO, though it has intuitive appeal and seems the most natural of the three nethods, is unfortunately only applicable to simple random sampling without replacement. To consider it as a competitor of the other two bootstraps, extensions to stratified sampling, two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) method of unequal probability sampling are proposed in Section 3. In Section 4, three existing nonbootstrapping methods are discussed: linearization, the jackknife, and a method based on Woodruff's (1952) confidence intervals for quantiles. The .results of a simulation study using various finite populations are given in Section 5. The methods are compared in terms of variance estimation and confidence intervals for the following nonlinear statistics: the ratio r, the regression coefficient b, the correlation coefficient c, and the median m. Stratified random sampling without replacement with different numbers of strata, stratum population sizes, and within-stratum sampling fractions is considered. The linearization and jackknife methods are included in the study for r, b, and c, and the Woodruff-based method is included for the median. The article concludes with Section 6, which summarizes the results and puts forth some suggestions.

2. THE THREE BOOTSTRAPS

To introduce necessary notation, stratified sampling without replacement is outlined at this point. In stratified random sampling, the finite population, consisting of N units, is partitioned into L nonoverlapping *strata* of $N_1 N_2, \ldots, N_L$ units, respectively; thus, $\sum_h N_h = N$. A simple random sample without replacement (SRSWOR) is taken independently from each stratum. The sample sizes within each stratum are denoted by $n_1, n_2, \ldots, n_L$, and the total sample size is $n = \sum_h n_h$. A vector of measurements of some unit characteristics is represented as $\mathbf{y}_{hi} = (y_{1hi}, y_{2hi}, \ldots, y_{\tau hi})^{\mathrm{T}}$, where the subscript h refers to the stratum label and the subscript i refers to the ith unit within the hth stratum. The population parameter of interest $\theta = \theta(\mathsf{S})$, where $\mathsf{S} = \{\mathbf{y}_{hi} : h = 1, 2, \ldots, L; i = 1, 2, \ldots, N_h\}$, is usually estimated by $\hat{\theta} = \hat{\theta}(\mathsf{S})$, where $\mathsf{S} = \{\mathbf{y}_{hi} : h = 1, 2, \ldots, L; i = 1, 2, \ldots, n_h\}$. The population mean vector is denoted $\bar{\mathbf{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_\tau)^{\mathrm{T}}$, and its unbiased estimate is $\bar{\mathbf{y}} = \sum_{h=1}^{L} W_h \bar{\mathbf{y}}_h = (\bar{y}_1, \ldots, \bar{y}_\tau)^{\mathrm{T}}$, where $\bar{\mathbf{y}}_h = \sum_{i=1}^{n_h} \mathbf{y}_{hi}/n_h = (\bar{y}_{1h}, \bar{y}_{2h}, \ldots, \bar{y}_{\tau h})^{\mathrm{T}}$ and $W_h = N_h/N$.

For $\tau = 1$, an unbiased estimate of $\mathcal{V}ar\ \bar{y}$, given in Cochran (1977), is

$$\operatorname{var} \bar{y} = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} s_h^2, \tag{2.1}$$

where $f_h = n_h/N_h$, and $s_h^2 = \sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2/(n_h - 1)$.

2.1. The Rescaling Method.

Many commonly used statistics can be written as $\hat{\theta} = g(\bar{\mathbf{y}})$, a function of means (e.g., the ratio, the regression coefficient, and the correlation coefficient). For estimating the variance of $\hat{\theta} = g(\bar{\mathbf{y}})$, Rao and Wu (1988) proposed the following procedure. A resample vector is drawn with replacement from the original sample, each resampled unit is rescaled, and the original estimator is applied to the rescaled vector. The rescaling factors are chosen so that the variance under resampling matches the usual variance estimator in the linear case. For stratified random sampling, the rescaling method is as follows:

1. Draw a simple random sample $\{\mathbf{y}_{hi}^*\}_{i=1}^{n_h^*}$ ($n_h^* \geq 1$) with replacement from $\{\mathbf{y}_{hi}\}_{i=1}^{n_h}$. Let $C_h = \sqrt{n_h^*}(n_h - 1)^{-1/2}(1 - f_h)^{1/2}$, and calculate the following:

$$\tilde{\mathbf{y}}_h^* = \bar{\mathbf{y}}_h + C_h(\bar{\mathbf{y}}_h^* - \bar{\mathbf{y}}_h),$$

$$\tilde{\mathbf{y}}^* = \sum_{h=1}^{L} W_h\tilde{\mathbf{y}}_h^*, \quad \text{and} \quad \hat{\theta}^* = g(\tilde{\mathbf{y}}^*).$$

2. Repeat step 1 a large number of times, B, to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$, and estimate $\mathcal{V}ar\ \hat{\theta}$ with

$$\mathrm{v}_r = \mathcal{E}_*(\hat{\theta}^* - \mathcal{E}_*\hat{\theta}^*)^2,$$

or its Monte Carlo approximation

$$\mathrm{v}_r = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2$$

($\mathcal{E}_*\hat{\theta}^*$ and $\hat{\theta}_{(\cdot)}^*$ can be replaced by $\hat{\theta}$), where $\mathcal{E}_*$ denotes expectation with respect to the resampling (this notation will be used throughout).

If $\tau = 1$, and $\hat{\theta} = \bar{y}$, then $\mathrm{v}_r(\tilde{y}^*) = \mathrm{var}\ \bar{y}$, the usual unbiased estimator of $\mathcal{V}ar\ \bar{y}$. To reduce computations the entire sample could be rescaled before initiating the bootstrap step.

Rao and Wu (1988) develop various rescaling algorithms for a wide class of sampling designs, including stratified sampling, two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) method of pps sampling. They all yield the usual variance estimates in the linear case. Rao and Wu show that for stratified sampling the method gives consistent variance estimates for $\hat{\theta} = g(\bar{\mathbf{y}})$, and in the linear case, with appropriate choice of n_h^* in each stratum, the bootstrap histogram captures the second-order term of the Edgeworth expansion of $\bar{y}$ as $L \longrightarrow \infty$.

Some limitations of the method do exist: (1) to calculate the rescaling factors various summary statistics of subsets of the sample are necessary, which, for large complex designs, may be undesirable; (2) if one chooses the resample sizes to match the second-order term of the Edgeworth expansion of $\bar{y}$, rescaling of each data point at each bootstrap iteration is necessary, and the number of required calculations is thus multiplied by n; and (3) for some choices of n_h^* it is possible for $\hat{\theta}^*$ to be negative even if $\hat{\theta} \geq 0$ by definition.

2.2. *The Mirror-Match Method.*

Sitter (1992) proposed a bootstrap method which we will call the mirror-match method. Generally, the method entails taking a resample without replacement from the sample to mirror the original sampling scheme, and then repeating this with replacement to match the usual variance estimates in the linear case. In the context of stratified random sampling, the mirror-match method is:

1. Choose $1 \leq n_h' < n_h$, and resample n_h' (SRSWOR) from stratum h to get $\mathbf{y}_{h1}^*, \mathbf{y}_{h2}^*, \ldots, \mathbf{y}_{hn_h'}^*$.
2. Repeat step 1, $k_h = n_h(1-f_h^*)/n_h'(1-f_h)$ times independently, replacing the resamples of size n_h' each time, to get $\mathbf{y}_{h1}^*, \mathbf{y}_{h2}^*, \ldots, \mathbf{y}_{hn_h^*}^*$, where $f_h^* = n_h'/n_h$ and $n_h^* = n_h(1-f_h^*)/(1-f_h)$ (if k_h is noninteger, a randomization between bracketing integers is used).
3. Repeat steps 1 and 2 independently for each stratum to get $\mathrm{S}^* = \{\mathbf{y}_{hi}^* : h = 1, \ldots, L;$ $i = 1, \ldots, n_h^*\}$, and let $\hat{\theta}^* = \hat{\theta}(\mathrm{S}^*)$.
4. Repeat steps 1–3 a large number of times, B, to get $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$, and estimate $\mathcal{V}ar$ $\hat{\theta}$ with

$$\mathrm{v}_m = \mathcal{E}_*(\hat{\theta}^* - \mathcal{E}_*\hat{\theta}^*)^2,$$

or its Monte Carlo approximation $\mathrm{v}_m = \sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2/B$ ($\mathcal{E}_*\hat{\theta}^*$ and $\hat{\theta}_{(\cdot)}^*$ can be replaced by $\hat{\theta}$).

If $\tau = 1$ and $\hat{\theta} = \bar{y}$, then $\mathrm{v}_m = \mathrm{var}\ \bar{y}$, the usual variance estimate. If $f_h \geq 1/n_h$, then choosing $n_h' = f_h n_h$ implies the resampling fraction f_h^* in step 1 is the same as the original sampling fraction f_h (mirror). This choice has theoretical justification using Edgeworth expansions which is described in the next paragraph.

Using this basic idea, Sitter (1992) gives extensions to two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) method of pps sampling. For stratified sampling he also shows that the method yields consistent variance estimates for nonlinear statistics, and when $\hat{\theta} = \bar{y}$, for appropriate choice of n_h' ($f_h^* = f_h$), the bootstrap histogram captures the second-order term of the Edgeworth expansion as $L \longrightarrow \infty$, as well as when L is bounded and n, $N \longrightarrow \infty$. The with-replacement bootstrap (BWR), proposed by McCarthy and Snowden (1985), is a special case of the mirror-match method, for the sampling designs where the BWR has been developed (i.e. for stratified let $n_h' = 1$).

Although the mirror-match method does not share the problems indicated for the rescaling method, it does have the disadvantage that even if the n_h''s are chosen integers, a randomization between bracketing integers for k_h is necessary.

2.3. *The BWO Method.*

The without-replacement bootstrap (BWO) method was proposed by Gross (1980) for variance estimation in SRSWOR. Assuming that $N = kn$ for some integer k, the method is as follows:

1. Construct a pseudopopulation by replicating the sample k times.
2. Draw a SRSWOR of size n from this pseudopopulation to get $\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*$, and let $\hat{\theta}^* = \hat{\theta}(\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*)$.
3. Repeat step 2 a large number of times, B, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, and estimate $\mathcal{V}ar$ $\hat{\theta}$ with

$$\mathrm{v}_{\mathrm{bwo}} = \mathcal{E}_*(\hat{\theta}^* - \mathcal{E}_*\hat{\theta}^*)^2,$$

TABLE 1: Values of p for the Bickel-Freedman BWO.

	p					
n	$N = 25$	30	40	50	75	100
2	−253	−404	−739	−1174	−2628	−4849
4	−12.0	−17.6	−37.7	−57.0	−134	−263
6	−1.90	−3.48	−6.79	−12.5	−30.8	−56.8
8	−0.04	−0.81	−2.07	−4.03	−11.0	−20.9
10	0.14	0.36	−0.36	−1.33	−4.66	−9.90
15	0.17	0.89	0.07	0.16	−0.45	−2.00
20	0.59	0.32	0.92	0.28	−0.15	.05

or its Monte Carlo approximation $\mathrm{v}_{\mathrm{bwo}} = \sum_{i=1}^{B}(\hat{\theta}^* - \hat{\theta}^*_{(\cdot)})^2/B$ ($\mathcal{E}_*\hat{\theta}^*$ and $\hat{\theta}^*_{(\cdot)}$ can be replaced by $\hat{\theta}$).

If $\tau = 1$ and $\hat{\theta} = \bar{y}$, then

$$\mathrm{v}_{\mathrm{bwo}} = \frac{n-1}{n-f}\,\frac{1-f}{n}\,s^2.$$

So the BWO method does not yield an unbiased estimate of variance even in the simplest situation. Of course, a correction factor could be used in this case, but if one applied the BWO method independently to each stratum in stratified sampling, a simple correction factor would not be available.

The following procedure was proposed by Bickel and Freedman (1984) as an extension of the BWO method to stratified sampling. It was further discussed by McCarthy and Snowden (1985).

Suppose $N_h = k_h n_h + r_h$ for $0 \le r_h \le n_h - 1$ for each stratum, with k_h and r_h integers. Construct two pseudopopulations for each stratum: population 1 by replicating $\{y_{hi}\}_{i=1}^{n_h}$, k_h times, and population 2 by replicating it $k_h + 1$ times. For each stratum independently, a SRSWOR of size n_h is drawn from population 1 with probability

$$p_h = \frac{(1-f_h)/(n_h - 1) - a_{h2}}{a_{h1} - a_{h2}},$$

where $a_{hj} = \{k_h + j - 1\}/\{(k_h + j)n_j - 1\}$ for $j = 1,2$, and from population 2 with probability $1 - p_h$. If this procedure is feasible, then $\mathcal{V}ar_*\ \bar{y}^* = \mathrm{var}\ \bar{y}$. Unfortunately, it is possible that $p_h < 0$, which would make the procedure infeasible. An example of this is given by McCarthy and Snowden (1985) for a single stratum. To illustrate the limitations of this extension, Table 1 gives values of p for various n- and N-values of a single stratum in stratified sampling.

3. AN EXTENDED BWO METHOD

The BWO method for SRSWOR is intuitively appealing, avoids any rescaling, and is easily applicable. Unfortunately, it does not extend to more complicated sampling designs. Due to this limited applicability, it would be difficult to justify consideration of the BWO as a possible competitor to the rescaling and mirror-match methods. With this in mind we propose, in this section, extensions of the BWO method applicable to stratified sampling, two-stage cluster sampling, and the Rao-Hartley-Cochran method of unequal-probability sampling. The extensions give the usual variance estimates in the linear case.

3.1. Stratified Sampling.

Ignoring, for the moment, all integer restrictions, let

$$n'_h = n_h - (1 - f_h) \quad \text{and} \quad k_h = \frac{N_h}{n_h}\left(1 - \frac{1 - f_h}{n_h}\right) \tag{3.1}$$

for $h = 1, \ldots, L$. Our method is as follows:

1. Create the hth stratum of the pseudopopulation by replicating $\{\mathrm{y}_{hi}\}_{i=1}^{n_h}$ k_h times. Do this for each stratum.
2. Resample n'_h units from stratum h without replacement to get $\{\mathrm{y}^*_{hi}\}_{i=1}^{n'_h}$ for $h = 1, \ldots, L$, and let $\hat{\theta}^* = \hat{\theta}(\mathrm{S}^*)$, where $\mathrm{S}^* = \{\mathrm{y}_{hi} : h = 1, \ldots, L;\ i = 1, \ldots, n'_h\}$.
3. Repeat 2 and 3 a large number of times, B, to get $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$, and estimate $\mathcal{V}ar\ \hat{\theta}$ with

   $$\mathrm{v}_{\mathrm{bwo}} = \mathcal{E}_*(\hat{\theta}^* - \mathcal{E}_*\hat{\theta}^*)^2,$$

   or its Monte Carlo approximation $\mathrm{v}_{\mathrm{bwo}} = \sum_{i=1}^{B}(\hat{\theta}^*_i - \hat{\theta}^*_{(\cdot)})^2/B$ ($\mathcal{E}_*\hat{\theta}^*$ and $\hat{\theta}^*_{(\cdot)}$ can be replaced by $\hat{\theta}$).

If $\tau = 1$ and $\hat{\theta} = \bar{y}$, then $\mathrm{v}_{\mathrm{bwo}} = \mathrm{var}\ \bar{y}$, the usual variance estimate. The method is similar to the Bickel-Freedman BWO, but the restrictions that $N_h = k_h n_h$, and that the resample size be n_h are removed. Instead n'_h and k_h are chosen to satisfy the following:

$$f^*_h = f_h \quad \text{and} \quad \mathcal{V}ar_*\ \bar{y}^*_h = \frac{1 - f_h}{n_h} s^2_h, \tag{3.2}$$

where $f^*_h = n'_h/(k_h n_h)$ is the resampling fraction. Ignoring all integer restrictions, these two equations are satisfied by (3.1). Clearly, the extension will be very close to the standard BWO in many situations.

Of course this only makes sense if n'_h and k_h are both integers for all h, and clearly, since $0 \le f_h \le 1$, $n'_h = n_h - (1 - f_h)$ is noninteger unless $f_h = 0$ or 1. To avoid this problem one randomizes between bracketing integer values. Since the same technique will be used independently for each stratum, the method for a single stratum with the h subscript suppressed is described below.

Let $k_1 = \lceil k \rceil$ and $k_2 = \lfloor k \rfloor$ ($\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the greatest integer smaller than, and the smallest integer greater than, respectively), where k is given in (3.1), and let $n'_1 = n - 1$ and $n'_2 = n$. Also let

$$p = \frac{(1 - f)/n(n - 1) - a_{h2}}{a_1 - a_2}, \tag{3.3}$$

where $a_j = \{k_j(1 - n'_j/nk_j)\}/\{n'_j(nk_j - 1)\}$ for $j = 1, 2$. It is straightforward to show that, for $n \ge 2$, $0 \le p \le 1$. So the method is generally applicable.

If $n \ge 2$, and if, to select the samples, (k_1, n'_1) are used with probability p, given in (3.3), and (k_2, n'_2) with probability $1 - p$, then, for $\tau = 1$,

$$\mathcal{V}ar_*\ \bar{y} = \mathrm{var}\ \bar{y}.$$

To see this, consider k and n' as described above. Let $\mathcal{E}_{2*}$ and $\mathcal{V}_{2*}$ denote the conditional expectation and variance under the resampling given k and n'. Similarly, let $\mathcal{E}_{1*}$ and $\mathcal{V}_{1*}$ denote the expectation and variance with respect to the randomization of (n', k). Then

$$\begin{aligned}\mathcal{V}ar_*\ \bar{y} &= \mathcal{E}_{1*}\mathcal{V}_{2*}(\bar{y}^*) + \mathcal{V}_{2*}\mathcal{E}_{2*}(\bar{y}^*) = \mathcal{E}_{1*}\mathcal{V}_{2*}(\bar{y}^*)\\ &= p\mathcal{V}_{2*}(\bar{y}^*|k_1, n'_1) + (1 - p)\mathcal{V}_{2*}(\bar{y}^*|k_2, n'_2)\\ &= (n - 1)s^2\{pa_1 + (1 - p)a_2\} = \frac{1 - f}{n}s^2.\end{aligned}$$

3.2. Two-Stage Sampling.

The same general idea can be used at each stage of two-stage cluster sampling, where each stage is a SRSWOR, to extend the BWO method to that situation. In two-stage cluster sampling, the population consists of N clusters, with M_i units in the ith cluster. A SRSWOR of n clusters is drawn, and then a SRSWOR of size m_i is drawn from each obtained cluster. The $\mathbf{y}_{ij} = (y_{1ij}, \ldots, y_{\tau ij})^{\mathrm{T}}$ denotes some vector of characteristics of the jth unit from the ith cluster. The usual unbiased estimate of the population mean is $\hat{\hat{\mathbf{Y}}} = \sum_1^n M_i \bar{\mathbf{y}}_i / n\bar{M}_0$, where $\bar{\mathbf{y}}_i = \sum_{j=1}^{m_i} \mathbf{y}_{ij}$ and $\bar{M}_0 = \sum_{i=1}^N M_i/N$. For $\tau = 1$, an unbiased estimate of $\mathcal{V}ar\ \hat{\hat{Y}}$ is

$$\operatorname{var} \hat{\hat{Y}} = \frac{1-f_1}{n} s_1^2 + \sum_{i=1}^{n} \frac{f_1(1-f_{2i})}{nm_i} s_{2i}^2, \tag{3.4}$$

where $f_1 = n/N, f_{2i} = m_i/M_i$,

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i \bar{y}_i}{\bar{M}_0} - \hat{\hat{Y}} \right)^2, \quad \text{and} \quad s_{2i}^2 = \frac{1}{n(m_i - 1)} \sum_{j=1}^{m_i} \left(\frac{M_i}{\bar{M}_0} \right)^2 (y_{ij} - \bar{y}_i)^2.$$

Let k_1, n', m_i', and k_{2i}, for $i = 1, \ldots, n$, be integers. The method is as follows:

1. Create a pseudopopulation by replicating each cluster in the sample k_1 times and each unit within the ith cluster k_{2i} times.
2. From this pseudopopulation take a SRSWOR of n' clusters, and a SRSWOR of m_i' units from each of the resampled clusters which was a replicate of the ith cluster, and let $\hat{\theta}^{**} = \hat{\theta}(\mathrm{S}^*)$, where $\mathrm{S}^* = \{\mathbf{y}_{ij} : i = 1, \ldots, n'; j = 1, \ldots, m_i'\}$.
3. Repeat step 2 a large number of times, B, to get $\hat{\theta}_1^{**}, \ldots, \hat{\theta}_B^{**}$, and estimate $\mathcal{V}ar\ \hat{\theta}$ with

$$\mathrm{v}_{\mathrm{bwo}} = \mathcal{E}_{**}(\hat{\theta}^{**} - \mathcal{E}_{**}\hat{\theta}^*)^2,$$

or its Monte Carlo approximation $\mathrm{v}_{\mathrm{bwo}} = \sum_{i=1}^{B} (\hat{\theta}_i^{**} - \hat{\theta}_{(\cdot)}^{**})^2 / B$ ($\mathcal{E}_{**}\hat{\theta}^{**}$ and $\hat{\theta}_{(\cdot)}^{**}$ can be replaced by $\hat{\theta}$), where $\mathcal{E}_{**}$ denotes expectation under the two stages of resampling.

If $\tau = 1$ and $\hat{\theta} = \hat{\hat{Y}}$, then

$$\mathrm{v}_{\mathrm{bwo}} = \mathcal{V}ar_{**}\ \hat{\hat{Y}}^{**} = \frac{k_1(n-1)}{k_1 n - 1} \frac{1 - f_1^*}{n'} s_1^2 + \sum_{i=1}^{n} \frac{k_{2i}(m_i - 1)}{k_{2i} m_i - 1} \frac{1 - f_{2i}^*}{n' m_i'} s_{2i}^2, \tag{3.5}$$

where $f_1^* = n'/k_1 n$ and $f_{2i}^* = m_i'/k_{2i} m_i$, for any choices of k_1, k_{2i}, n', and m_i'. Ignoring integer restrictions, one can choose n' and k_1 to satisfy

$$f_1^* = f_1 \quad \text{and} \quad \frac{k_1(n-1)}{n'(k_1 n - 1)} = \frac{1}{n},$$

and m_i' and k_{2i} to satisfy

$$f_{2i}^* = f_{2i} \quad \text{and} \quad \frac{k_{2i}(m_i - 1)}{n' m_i' (k_{2i} m_i - 1)} = \frac{f_1}{nm_i}$$

for each i. If this is done, $\mathcal{V}ar_{**}\ \hat{\hat{Y}}^{**} = \operatorname{var} \hat{\hat{Y}}$. To avoid the problem of noninteger resample sizes in this extension of the BWO method, one can randomize between bracketing integer values and still retain the unbiasedness of the variance estimate in the linear case (see Sitter 1989).

3.3. The Rao-Hartley-Cochran (RHC) Method for PPS Sampling.

A simple method of sampling with unequal probabilities without replacement was proposed by Rao, Hartley, and Cochran (1962); see also Cochran (1977). The sampling method is as follows:

1. Randomly partition the population of N units into n groups $\{G_g\}_{g=1}^{n}$ of sizes $\{N_g\}_{g=1}^{n}$, respectively.
2. Draw one unit from each group with probability z_j/Z_g for the gth group, where $z_j = x_j/X$, $Z_g = \sum_{j\in G_g} z_j$, x_j = some measure of the size of the jth unit, and $X = \sum_{j=1}^{N} x_j$.

An unbiased estimator of $\bar{Y}$, the population mean, is $\hat{\bar{Y}} = \sum_{g=1}^{n} w_g y_g/n$, where $w_g = f/\pi_g$, $\pi_g = z_g/Z_g$ is the probability of selection of the particular unit sampled from group $g, f = n/N$ is the probability of selection under simple random sampling without replacement, and y_g and z_g denote the values for the unit selected from the gth group. Note that by definition $\sum_{g=1}^{n} Z_g = 1$. An unbiased estimate of $\mathcal{V}ar\ \hat{\bar{Y}}$ is

$$\operatorname{var}\hat{\bar{Y}} = \frac{\sum_{g=1}^{n} N_g^2 - N}{N^2 - \sum_{g=1}^{n} N_g^2} \sum_{g=1}^{n} Z_g \left(\frac{y_g}{N z_g} - \hat{\bar{Y}}\right)^2 . \tag{3.6}$$

To extend the BWO method to this situation, let $\hat{Y} = N\hat{\bar{Y}} = \sum_{g=1}^{n} Z_g y_g/z_g$. A BWO procedure for this situation is:

1. Replicate (z_g, y_g) $k_g = Z_g/z_g$ times for $g = 1, \ldots, n$ to create a pseudopopulation.
2. Randomly partition the pseudopopulation of $N^* = \sum_{g=1}^{n} k_g$ units into n^* groups, $\{\Gamma_g^*\}_{g=1}^{n^*}$, of sizes $\{N_g^*\}_{g=1}^{m}$.
3. Randomly select one (z_i^*, y_i^*) pair from each of the n^* groups with probability z_i^*/Z_g^*, where $Z_g^* = \sum_{i\in\Gamma_g^*} z_i^*$, and let $\hat{\theta}^* = \hat{\theta}(\mathbf{z}^*, \mathbf{y}^*)$.
4. Repeat steps 1–3 a large number of times, B, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$, and estimate $\mathcal{V}ar\ \hat{\theta}$ with

$$\mathrm{v_{bwo}} = \mathcal{E}_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2,$$

or its Monte Carlo approximation $\mathrm{v_{bwo}} = \sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2/B$ ($\mathcal{E}_*\hat{\theta}^*$ and $\hat{\theta}_{(\cdot)}^*$ can be replaced by $\hat{\theta}$).

If $\theta = Y$ and $\hat{\theta} = \hat{Y}$, we have

$$\mathrm{v_{bwo}} = \mathcal{V}ar_*\ \hat{Y}^* = \left(\frac{\sum_{g=1}^{n^*} N_g^{*2} - N^*}{\sum_{g=1}^{n} N_g^2 - N}\right)\left(\frac{N^2 - \sum_{g=1}^{n} N_g^2}{N^*(N^* - 1)}\right) \operatorname{var}\hat{Y}. \tag{3.7}$$

This suggests two alternatives. The simplest is to merely correct the variance estimate by an appropriate multiplicative factor. The second is to choose n^* and N_g^* for $g = 1, \ldots, n^*$ so that the product of the first two factors in (3.7) is close to one. In many situations it should be possible to find bracketing values of n^* and N_g^* and randomize between them, but a closed-form solution seems unavailable. This is a limitation brought on by the complexity of the original sampling [for a discussion of various situations, and appropriate randomization schemes, see Sitter (1989)].

4. COMPETITORS TO BOOTSTRAPPING

This section briefly describes three competitors to bootstrapping for variance estimation and confidence intervals in stratified sampling, where a SRSWOR is taken within each stratum. These three methods are used in the simulation study of Section 5. The first is a method based on Woodruff's (1952) confidence intervals for quantiles. The two others are the linearization method and the delete-one jackknife method, which are most commonly used when $\hat{\theta} = g(\bar{\mathbf{y}})$, a function of means.

4.1. Woodruff-Based Method for Quantiles.

A method, proposed by Woodruff (1952) for obtaining confidence intervals for quantiles in general survey designs, is described here for stratified sampling, though it is applicable to general sampling designs. Let $P(z) = \sum_{h=1}^{L} W_h P_h(z)$ be a function of z, where $P_h(z) = (1/n_h)\sum_{i=1}^{n_h} I_{[y_{hi} \leq z]}$ is the proportion of sampled observations less than or equal to z, and I is the indicator function. the pth percentile, Q_p, is then estimated by

$$q_p = P^{-1}(p) = \inf\{z : P(z) \geq p\}. \tag{4.1}$$

In practice, if $y_{(1)}, \ldots, y_{(n)}$ are the sample values arranged in ascending order, and $w_{(1)}, \ldots, w_{(n)}$ their corresponding values of W_h/n_h, then $q_p = y_{(k)}$, where k is the smallest integer such that $\sum_{j=1}^{k} w_{(j)} > p$.

An approximate $(1 - \alpha)$-level confidence interval for q_p is given by

$$[P^{-1}(p - z_{\alpha/2}s_p),\ P^{-1}(p + z_{\alpha/2}s_p)],$$

where $s_p^2 = \sum_{h=1}^{L} W_h^2(1 - f_h)P_h(q_p)\{1 - P_h(q_p)\}/(n_h - 1)$, and $z_{\alpha/2}$ is the upper $\alpha/2\%$ cutoff point of the standard normal distribution. If $p = 0.5$, this gives a confidence interval for the median.

One can obtain a variance estimate from this confidence interval by equating it to the normal-theory interval. For the median we get $v_\alpha = \{L(\alpha)/2z_{\alpha/2}\}^2$, where $L(\alpha)$ is the length of the Woodruff confidence interval (see Francisco and Fuller 1991, Rao and Wu 1987). Note that this variance estimate depends on α, so throughout the simulation study of Section 5, various α-values are used in an attempt to get some empirical indication of the optimum choice.

4.2. The Linearization and Jackknife Methods.

If $\hat{\theta} = g(\bar{\mathbf{y}})$, a smooth function of means, then the linearization estimate of variance is obtained by taking the first two terms of the Taylor series expansion of

$$g(\bar{\mathbf{y}}) = g(\bar{\mathbf{Y}}) + (\bar{\mathbf{y}} - \bar{\mathbf{Y}})^{\mathrm{T}} g'(\bar{\mathbf{Y}})$$

about $\bar{\mathbf{Y}}$, the expected value of $\bar{\mathbf{y}}$, where $g'(\bar{\mathbf{Y}})$ is the gradient of g evaluated at $\bar{\mathbf{Y}}$. Taking the variance on both sides of this expansion, and substituting the usual mean and variance estimates, one has

$$\widehat{\mathcal{V}ar}\ g(\bar{\mathbf{y}}) = g'(\bar{\mathbf{y}})^{\mathrm{T}}(\widehat{\mathcal{V}ar}\ \bar{\mathbf{y}})g'(\bar{\mathbf{y}}).$$

The delete-one jackknife (or just the jackknife) method recalculates the estimate $\hat{\theta}$ after deleting one observation from the sample, and repeats this for each sampled observation. The resulting delete-one estimates are then used to obtain an estimate of the variance of $\hat{\theta}$. In the study by Kovar, Rao, and Wu (1988), very little difference was found between the six jackknife definitions which they considered for stratified sampling with replacement. In our case, the sampling within each stratum is without replacement, but in view of their results it seemed unlikely that there would be a substantial difference in the performance of these six jackknifes. With this in mind, only one jackknife, defined below, was used (their $\mathrm{v}_{\mathrm{J}2}$).

With $\mathrm{S} = \{\mathrm{y}_{hi} : h = 1,\ldots,L;\ i = 1,\ldots,n_h\}$ the sample, and $\hat{\theta} = \hat{\theta}(\mathrm{S})$, let $\hat{\theta}_{(hi)} = \hat{\theta}(\mathrm{S}_{(hi)})$, where $\mathrm{S}_{(hi)}$ is the sample with the ith unit from the hth stratum removed. Then the jackknife estimate of the variance of $\hat{\theta}$ is

$$\mathrm{v}_{\mathrm{J}}(\hat{\theta}) = \sum_{h=1}^{L} \frac{(1-f_h)(n_h-1)}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{(hi)} - \hat{\theta})^2.$$

4.3. Confidence Intervals.

For the linearization and jackknife methods, confidence intervals are obtained using normal theory, and are given by

$$[\hat{\theta} - z_{\alpha/2}\sqrt{\mathrm{v}},\ \hat{\theta} + z_{\alpha/2}\sqrt{\mathrm{v}}],$$

where v is the corresponding variance estimate of the particular method, and $z_{\alpha/2}$ is the upper $\alpha/2\%$ cutoff for the standard normal distribution. For the various bootstrap methods, one can obtain a histogram of the B bootstrap estimates $\hat{\theta}_b^*$. Confidence intervals can be constructed using these values by the percentile method (Efron 1982). This method is used for the median. For nonlinear functions of means the bootstrap-t method is used. Here the bootstrap histogram of $t_b^* = (\hat{\theta}_b^* - \hat{\theta})/\sqrt{\mathrm{v}_{\mathrm{J}}^*}$, where $\mathrm{v}_{\mathrm{J}}^*$ is a jackknife variance estimator (using the delete-one jackknife described in the previous section) applied to the resampled data, is used to obtain the upper and lower $\alpha/2\%$ cutoff points t_{U}^* t_{L}^* for t^*. The bootstrap-t $(1-\alpha)\%$ confidence interval is then given by

$$[\hat{\theta} - t_{\mathrm{U}}^*\sqrt{\mathrm{v}_{\mathrm{J}}},\ \hat{\theta} - t_{\mathrm{L}}^*\sqrt{\mathrm{v}_{\mathrm{J}}}],$$

where v_{J} is the same jackknife variance estimate applied to the whole sample. Abromovitch and Singh (1985) demonstrated the superiority of the one-sided bootstrap-t intervals over the normal-theory intervals for the i.i.d. case using asymptotic expansions.

5. A SIMULATION STUDY

Theoretically there is little difference between the three bootstrap methods described. Rao and Wu (1988) and Sitter (1992) prove consistency of the variance estimates for the rescaling and mirror-match methods, respectively, where $\hat{\theta} = g(\bar{\mathbf{y}})$, a function of means. A similar result can be easily obtained for the BWO extension of Section 3 from the work of Bickel and Freedman (1984). There is no similar result for any of the three for the median. In terms of confidence intervals, Rao and Wu (1988) and Sitter (1992) prove that the rescaling and mirror-match methods, respectively, capture the second-order term of the Edgeworth expansion of the distribution of $\bar{y}$ as $L \longrightarrow \infty$. There are no results for $\hat{\theta} = g(\bar{\mathbf{y}})$ or the median.

This section describes a simulation study that was done to compare the performance of the three bootstrap methods for variance estimation and confidence-interval estimation for stratified random sampling without replacement. Eight finite populations are considered. The estimators considered in this study are the ratio, the regression coefficient, the correlation coefficient, and the median. Assuming the notation for stratified sampling of Section 2, let z_{hi} denote the characteristic of interest of the ith observation from the hth stratum, and let x_{hi} denote a related concomitant variable. Let $y_{1hi} = x_{hi}$, $y_{2hi} = z_{hi}$, $y_{3hi} = x_{hi}z_{hi}$, $y_{4hi} = x_{hi}^2$, and $y_{5hi} = z_{hi}^2$. The sample estimates of the corresponding population parameters used were: (1) the ratio $r = \bar{y}_2/\bar{y}_1$; (2) the regression coefficient $b = (\bar{y}_3 - \bar{y}_1\bar{y}_2)/(\bar{y}_4 - \bar{y}_1^2)$; (3) the correlation coefficient $c = \{\bar{y}_3 - \bar{y}_1\bar{y}_2\}/\{(\bar{y}_4 - \bar{y}_1^2)(\bar{y}_5 - \bar{y}_2^2)\}^{\frac{1}{2}}$; and (4) the median m, which was given in (4.1) with $p = \frac{1}{2}$; here $\bar{y}_j = \sum_{h=1}^{L} W_h\bar{y}_{jh}$ with $\bar{y}_{jh} = \sum_{i=1}^{n_h} y_{jhi}/n_h$ for $j = 1, \ldots, 5$.

For completeness, the linearization and jackknife methods of Section 4.2 are included for $\hat{\theta} = g(\bar{\mathbf{y}})$ (i.e., r, b, and c), and the Woodruff-based method of Section 4.1 is included for the median. These methods are much less computer-intensive. In this simulation study, the actual computation times for the various bootstrap methods were similar, except for the rescaling method when the resample size was chosen to match third moments. When this was done the rescaling method took much longer. This is due to the fact that when a within-stratum sampling fraction is near $\frac{1}{2}$, the resample size advocated can be much larger than the original stratum sample size (see Rao and Wu 1988).

5.1. Generating the Finite Populations.

The finite populations used are based on hypothetical populations originally given by Hansen and Tepping (1985) and Hansen, Madow, and Tepping (1983). These hypothetical populations were used as the basis for a large simulation study by Kovar, Rao, and Wu (1988), in which they compare the rescaling bootstrap to various nonbootstrapping methods for stratified sampling, where a simple random sample *with* replacement (SRSWR) is taken within each stratum. The population in Hansen and Tepping (1985) was put forth as an approximation to real populations encountered in the National Assessment of Educational Progress study. This population consisted of $L = 32$ strata, with (x_h, y_h) bivariate normal variates distributed $N_2(\mu_{xh}, \mu_{yh}, f_x\sigma_{xh}^2, f_y\sigma_{yh}^2, \rho)$ for $h = 1, \ldots, L$, where ρ is constant across strata and f_x and f_y are constant multipliers used to increase the variability of the underlying population. The W_h's range from 0.013 to 0.042, and the stratum means and variances are such that the coefficients of variation of x and y are approximately 10% and 30% respectively. The parameter settings, as well as the stratum weights, are given in Sitter (1989). Kovar, Rao, and Wu (1988) extended this hypothetical population to gamma distributions with the same mean and covariance structure in the following manner. Let x_h be distributed Gamma(α, β) with shape parameter $\alpha = \mu_{xh}^2/f_x\sigma_{xh}^2$ and scale parameter $\beta = f_x\sigma_{xh}^2/\mu_{xh}$, so that x_h has mean and variance μ_{xh} and $f_x\sigma_{xh}^2$. Using a linear regression model, y_h was obtained from x_h by

$$y_h = \frac{\sqrt{f_y}\sigma_{yh}\rho}{\sqrt{f_x}\sigma_{xh}} x_h + \epsilon_h,$$

where ϵ_h is normally distributed with mean $\mu_{yh} - \sqrt{f_y}\sigma_{yh}\mu_{xh}\rho/\sqrt{f_x}\sigma_{xh}$ and variance $f_y\sigma_{yh}^2(1-\rho^2)$. So (x_h, y_h) has joint distribution with the prespecified moments.

Kovar, Rao, and Wu (1988) generate 60 populations using the values of μ_{xh}, μ_{yh}, σ_{xh}, σ_{yh}, and W_h in Hansen and Tepping (1985), by varying the choice of f_x and f_y. Thirty were generated using bivariate normals, and thirty using gammas, as described above. In addition to these 60 populations, Kovar, Rao, and Wu (1988) use the method described in Hensen, Madow, and Tepping (1983) to generate a bivariate population where (x, y) are highly correlated. They then divided the y-values into $L = 32$ strata using the

x-values, and finally they calculated the resulting μ_{yh}, σ_{yh}, and W_h values. They used these parameter values as the basis for comparing methods when $\hat{\theta}$ = median.

The study by Kovar, Rao, and Wu (1988) dealt with SRSWR, so they could merely use the values of μ_{xh}, μ_{yh}, $f_y\sigma_{yh}^2$, $f_y\sigma_{yh}^2$, and W_h, along with the random mechanism for generation, to do their simulation. Here, sampling without replacement is being considered; thus finite populations must first be generated from hypothetical populations, and then simulations performed on the obtained finite populations. The methods under consideration include a number of different bootstraps, which are computer-intensive. Also, the stratum sampling fractions could be important in comparing the different methods, which suggests using a number of different stratum sample sizes on each finite population. With these points in mind, it is clear that cost constraints will limit the number of populations that can be considered. Some of the general results of Kovar, Rao, and Wu (1988) were used to choose hypothetical population parameters in an attempt to create finite populations which would emphasize differences in the performance of the various methods.

To compare the performance of the various methods for variance estimation and confidence intervals for r, b, and c, four finite populations, numbered 1 through 4, were generated:

(1) Finite populations 1 and 2 consist of $N = 800$ units divided into $L = 16$ strata. The two populations were generated using bivariate normals and gammas, respectively, as described above. The parameter settings used were varied so that $75 \leq \mu_{xh} \leq 95$, $45 \leq \mu_{yh} \leq 75$, $33.5 \leq \sigma_{xh} \leq 42.5$, $12 \leq \sigma_{yh} \leq 25$, $22 \leq N_h \leq 69$, $\rho = 0.8$, $f_x = 20$, and $f_y = 1$. The number of strata was reduced from $L = 32$ to $L = 16$ by averaging adjacent strata.

(2) Populations 3 and 4 were generated using a bivariate normal distribution and similar parameter settings, but aggregated into fewer strata. These two populations consist of $N = 760$ and $N = 516$ units respectively divided into $L = 6$ strata, with the strata sizes N_h ranging from 80 to 200 and from 36 to 144 respectively.

Thus populations 1 and 2 have a larger number of strata with fewer units per stratum than populations 3, and 4.

To compare the performance of the methods for variance estimation and confidence interval for the estimated median m, four finite populations, numbered 5 through 8, were generated. All four were generated using normal distributions in a similar fashion to the above, but with only y_h-values:

(1) Population 5 is similar to population 1, consisting of $N = 800$ units divided into $L = 16$ strata, with the same μ_{yh}, σ_{yh}, and N_h settings, but with $f_y = 0.5$. This corresponds to the parameter settings used by Kovar, Rao, and Wu (1988) to consider the median.

(2) Population 6 is similar to Population 4, consisting of $N = 516$ units divided into $L = 6$ strata, with the same μ_{yh}, σ_{yh}, and N_h settings, but with $f_y = 0.01$.

(3) Population 7 consists of $N = 800$ units divided into $L = 32$ strata, with the same parameter settings as those generated by Kovar, Rao, and Wu (1988) from the population in Hansen, Madow, and Tepping (1983). The parameter settings varied over $2 \leq \mu_{yh} \leq 42$, $0.1 \leq \sigma_{yh} \leq 6.7$, and $8 \leq N_h \leq 256$. From the parameter settings of Population 7 (see Sitter 1989), it is clear that a finite population generated from these settings will have little or no overlap of values across strata. So from the W_h-values it is clear that the resulting median will likely be near the center of stratum 6.

(4) Population 8 was generated in the same manner as Population 7, but with slightly

different N_h-values to create a population with median near a stratum boundary.

The resulting stratum means and standard deviations, as well as the R, B, C, and median, where applicable, for the above eight finite populations are given in Sitter (1989).

5.2. Measures of Performance.

To simplify the study, equal within-stratum sample sizes ($n_h = n_0$ for $h = 1, \ldots, L$) are used, so the total sample size is $n = n_0 L$. For all the bootstrap variance and confidence-interval estimation methods under consideration, $\hat{\theta}^*_{(\cdot)}$ is replaced by $\hat{\theta}$ in the definitions of their Monte Carlo approximations, and all methods are compared with the true MSE of the estimator of interest in the following manner. First, the true MSE was estimated by selecting 3000 stratified random samples without replacement, and using

$$\mathrm{MSE} = \frac{\sum_{b=1}^{3000} (\hat{\theta}_b - \theta)^2}{3000}.$$

The simulation consisted of selecting S new stratified random samples, and for each sample calculating the various variance and confidence limit estimates. The variance estimators were then compared in terms of their relative bias and relative instability. The relative bias and relative instability of a particular variance estimator v were measured by

$$\text{bias} = \frac{\bar{\mathrm{v}} - \mathrm{MSE}}{\mathrm{MSE}} \quad \text{and} \quad \text{instab} = \frac{\sigma_{\mathrm{v}}}{\mathrm{MSE}},$$

respectively, where $\bar{\mathrm{v}} = \sum_s v_s / S$ and $\sigma^2_{\mathrm{v}} = \sum_s (\mathrm{v}_s - \mathrm{MSE})^2 / S$. These variance estimators were then compared with regard to their relative bias and relative instability.

For comparing the various confidence interval estimators, the error rates in the upper and lower tails were compared to the nominal error rates of 5% and 10% in each tail. The standardized interval lengths of the methods were also compared. The upper- and lower-tail error rates and the standardized interval lengths were obtained using

$$U = \frac{\text{No. of samples with } \theta > \theta_{\mathrm{Us}}}{S}, \qquad L = \frac{\text{No. of samples with } \theta < \theta_{\mathrm{Ls}}}{S},$$

$$\text{length} = \frac{\sum_s (\theta_{\mathrm{Us}} - \theta_{\mathrm{Ls}})/S}{2 z_{\alpha/s} \sqrt{\mathrm{MSE}}},$$

where $(\theta_{\mathrm{Ls}}, \theta_{\mathrm{Us}})$ is the estimated confidence interval from the sth sample, and $z_{\alpha/2}$ is the upper $\alpha/2\%$ cutoff point of a standard normal distribution.

For the nonlinear statistics r, b, and c, the bootstrap variance estimates and confidence intervals were based on $B = 200$ bootstrap replicates, and $S = 1000$ simulation replicates were used. With $S = 1000$ the empirical coverage probabilities in each tail pertaining to nominal levels 0.05 and 0.1 will be within 0.01 and 0.02, respectively, 95% of the time. For the median, the bootstrap variance estimates and confidence intervals were based on $B = 500$ bootstrap replicates, and $S = 500$ replicates were used. With $S = 500$ the empirical coverage probabilities in each tail pertaining to the nominal levels 0.05 and 0.1 will be within 0.01 and 0.02 respectively 85% of the time.

5.3. Nonlinear Statistics.

Recall from Section 5.1 that populations 1 through 4 were used to compare the methods on nonlinear functions of means: r, the ratio; b, the regression coefficient; and c, the

TABLE 2: Average error rates in the upper and lower tails, average absolute relative bias, and relative instability over considered populations (ratio, regression, and correlation).

	Nominal error rate							
	5%			10%				
Method	$\bar{L}$	$\bar{U}$	$\bar{L}+\bar{U}$	$\bar{L}$	$\bar{U}$	$\bar{L}+\bar{U}$	·bias[a]·	instab[b]
				Ratio				
lin	4.56	5.45	10.01	9.61	10.83	20.44	0.029	0.22
jack	4.54	5.44	9.98	9.58	10.83	20.41	0.029	0.22
ext bwo	5.15	4.54	9.69	10.11	9.99	20.10	0.032	0.25
m-m[c] 1	5.30	4.69	9.99	10.09	9.98	20.07	0.034	0.25
m-m 2	5.30	4.63	9.93	10.11	10.18	20.29	0.032	0.24
rs[d] 1	4.83	4.16	8.99	9.94	9.65	19.59	0.034	0.24
rs 2	5.23	4.58	9.81	9.91	9.74	19.65	0.034	0.24
				Regression				
lin	5.71	6.25	11.96	10.49	10.76	21.25	0.051	0.30
jack	5.30	5.71	11.01	9.73	10.09	19.82	0.036	0.34
ext bwo	4.51	5.35	9.86	9.06	10.15	19.21	0.034	0.33
m-m 1	4.66	5.41	10.07	9.41	10.30	19.71	0.033	0.32
m-m 2	4.59	5.34	9.93	9.16	10.20	19.36	0.033	0.32
rs 1	4.21	4.98	9.19	9.20	9.56	18.76	0.031	0.32
rs 2	4.51	5.05	9.56	8.99	10.19	19.18	0.033	0.32
				Correlation				
lin	7.44	4.14	11.58	13.10	8.91	22.01	0.041	0.38
jack	6.78	3.83	10.61	12.20	8.36	20.56	0.034	0.42
ext bwo	3.93	5.65	9.58	9.01	10.56	19.57	0.034	0.42
m-m 1	4.34	5.88	10.22	8.91	10.83	19.74	0.029	0.42
m-m 2	4.33	5.50	9.83	8.90	10.68	19.59	0.025	0.40
rs 1	4.19	5.30	9.49	8.35	10.33	18.68	0.023	0.41
rs 2	4.05	5.48	9.54	8.63	10.45	19.08	0.026	0.42

[a] Average(·rel. bias·) over populations and strata sample sizes considered.

[b] Average(rel. instability).

[c] m-m 1: $n'_h = f_h n_h$; m-m 2: $n'_h = 1$.

[d] Rescale 1: resample size chosen to match 3rd moments; rescale 2: resample sizes = $n_h - 1$.

correlation coefficient. For populations 1 and 2, stratum sample sizes of 5, 6, and 7 were considered for each, while for populations 3 and 4 stratum sample sizes of 20 and 12 were used respectively.

Table 2 gives the average observed error rates for both one-tailed and two-tailed confidence intervals, where the average is over the four populations and the various stratum sample sizes considered. Table 2 also gives the average absolute relative bias and the average relative instability of the variance estimates. The abbreviations in Table 2 represent the following: lin is the linearization method; jack is the jackknife; ext bwo is the extended BWO method of Section 3; m-m 1 is the mirror-match method with n'_h chosen to match third moments; m-m 2 is the mirror-match method with $n'_h = 1$; rs1 is the rescaling method with n^*_h chosen to match third moments; and rs2 is the rescaling method with $n^*_h = n_h - 1$.

Figure 1 is a dot plot of the absolute relative bias for r, b, and c for each method and each population. For populations 1 and 2, the three lines represent the sample sizes 5, 6,

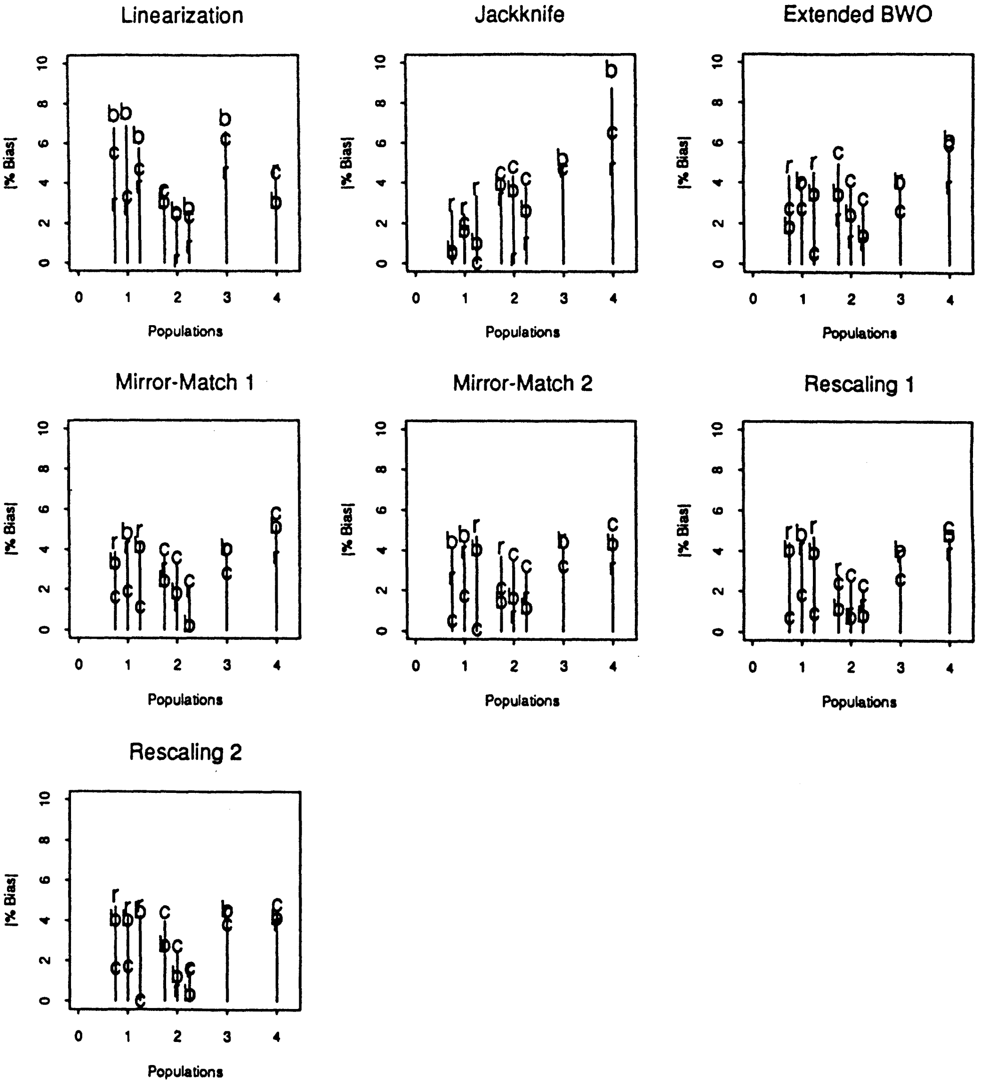
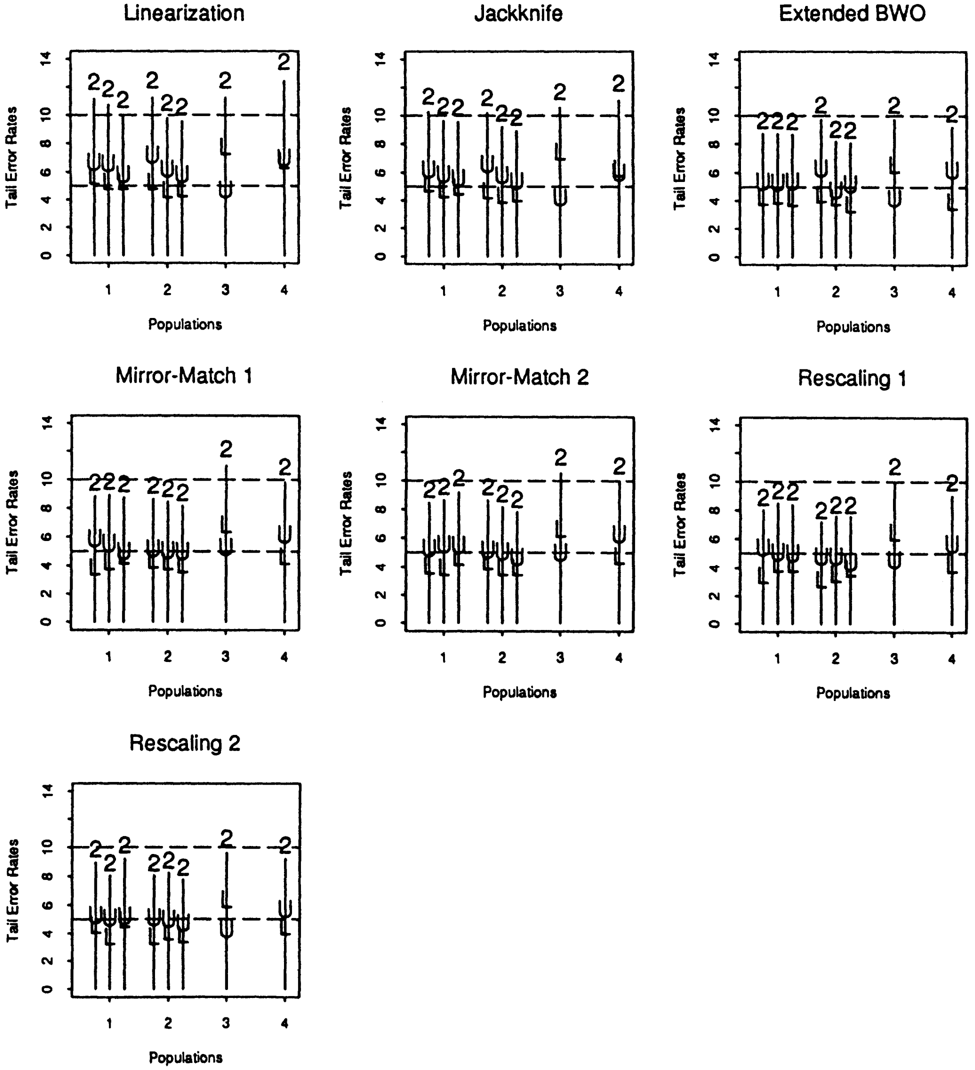


FIGURE 1: Dot plot of absolute relative bias of variances in percentage; the three lines each for populations 1 and 2 refer to the strata sample sizes of 5, 6, and 7 used for these two populations.

and 7 used. Figure 2 is a dot plot of the lower-tail (L), upper-tail (U), and two-tail (2) error rates (5% nominal) for *b*, for each method and each of the four populations and various sample sizes. The horizontal dashed lines indicate the one- and two-tail nominal levels. Similar dot plots for *r* and *c* are not included due to space considerations. The relative differences are similar, but less pronounced for the ratio, and more pronounced for the correlation coefficient. Some general comments based on Table 2 and Figures 1 and 2 are:

(1) *Variance estimates:* (a) For the ratio, the linearization and jackknife methods perform equivalently, and slightly better than the bootstrap methods in terms of both relative bias and instability; (b) there is little to choose between the bootstrap methods and the jackknife method for *b*, but the bootstrap methods do slightly better for *c*; (c) the linearization method yields higher relative bias, but slightly lower relative instability



Linearization

Jackknife

Extended BWO

Mirror-Match 1

Mirror-Match 2

Rescaling 1

Rescaling 2

FIGURE 2: Dot plot of one- and two-tail error rates for b, the regression coefficient; the L, U, and 2 refer to the error rates for the lower, upper, and two tails. The three lines each for populations 1 and 2 refer to the strata sample sizes of 5, 6, and 7 used for these two populations. The horizontal dashed lines mark the nominal error levels of 5% and 10%.

for b and c; and (d) Figure 1 shows that the bootstrap methods are more robust over the four populations considered, in terms of relative bias, than are the linearization or jackknife methods.

(2) *Confidence intervals:* (a) For the ratio, the methods perform equally well. As the nonlinearity increases ($r \longrightarrow b \longrightarrow c$), the bootstrap methods track the one-tail error rates better than do the linearization and jackknife methods. Kovar, Rao, and Wu (1988) observed a similar trend in their study. (b) On average, the rescaling method tends to have tail probabilities below the nominal level. (c) Figure 2 shows that for b the population-to-population differences are similar for all the bootstrap methods. This holds true for r and c as well, though they are not shown here. (d) As nonlinearity increases, choosing the resample sizes in the mirror-match method to match the second-order term of the

TABLE 3: Population 6: variances and C.I.'s for the median.

	Nominal error rate									
	5%			10%			Std. length			
Method	L	U	$L+U$	L	U	$L+U$	5%	10%	bias	instab
wood:	3.8	4.8	8.6	8.2	7.4	15.6	0.96	1.02		
$\alpha = 0.01$									0.227	0.64
$\alpha = 0.025$									0.074	0.59
$\alpha = 0.05$									0.066	0.63
$\alpha = 0.1$									0.027	0.68
$\alpha = 0.2$									0.178	0.88
ext bwo	11.4	3.4	14.8	16.0	6.6	22.6	0.92	0.92	0.172	0.85
m-m[a] 1	11.0	2.8	13.8	15.6	6.6	22.2	0.96	0.92	0.222	0.85
m-m 2	7.2	4.8	12.0	14.4	8.2	22.6	0.96	0.93	0.300	0.95
rs[b] 1	10.2	4.0	14.2	15.0	6.6	21.6	0.94	0.94	0.250	0.87
rs 2	10.6	3.8	14.4	15.6	7.2	22.8	0.96	0.99	0.387	1.00

[a] Mirror-match 1: $n'_h = f_h n_h$; mirror-match 2: $n'_h = 1$.
[b] Rescale 1: resample size chosen to match 3rd moments; rescale 2: resample size $= n_h - 1$.

Edgeworth expansion of the mean seems to become more beneficial. Though the same is not true for the rescaling method, this choice of resample size performs much better than was shown in the study by Kovar, Rao, and Wu (1988), where choosing the resample size in the rescaling method to match the Edgeworth expansion of the mean yielded very poor performance, which is contrary to what one would expect.

5.4. The Median.

The four populations considered for the median, described in Section 5.1, group naturally into populations 5 and 6 and populations 7 and 8. Populations 7 and 8 represent the situation where stratification is done using a highly correlated concomitant variable. For populations 5 and 6 stratum sample sizes of 5 and 6 and of 12 were considered, respectively. For populations 7 and 8, stratum sample sizes of 4 through 7 were considered for both. Throughout the tables the abbreviations used for the methods are the same as those used in Table 2, which were described in Section 5.3. In addition, "wood" represents the Woodruff-based method, with various α-values given.

Table 3 summarizes the results for population 6 (population 5 gave qualitatively similar results). It gives the observed one-tail and two-tail error rates, and the standardized lengths for nominal levels of 5% and 10%, as well as the relative bias and the relative instability for the variance estimates. The Woodruff-based variance estimate's performance depends on the choice of α, but for this population it outperforms the bootstrap methods for a fairly wide range. In terms of confidence intervals, no clear winner is evidenced. The bootstrap methods overstate the coverage probabilities and have shorter standardized length than the Woodruff intervals, which understate the coverage probabilities. It should be noted that the bootstrap methods do not track the one-tail error rates well.

Table 4 summarizes the variance-estimate results for populations 7 and 8. It illustrates the greater robustness of the extended BWO of Section 3 and the mirror-match methods to different stratifications than for the Woodruff-based and the rescaling methods. It should be noted that though Kovar, Rao, and Wu (1988) include the median in their simulation study, the rescaling method was introduced by Rao and Wu (1988) only for functions of means. Table 5 summarizes the confidence-interval results for population 7 with stratum sample sizes 5 and 7. For this population, the mirror-match method performed the best.

TABLE 4: Variances for the median.

	Population 7						Population 8					
	$n_0 = 5$		6		7		5		6		7	
Method	bias	instab	bias	instab	bias	instab	bias	instab	bias	instab	bias	instab
wood:												
$\alpha = 0.01$	2.21	2.46	0.63	0.89	0.43	0.65	6.12	6.75	6.75	7.00	9.01	9.69
$\alpha = 0.025$	−0.09	0.44	0.98	1.25	−0.29	0.46	7.71	6.27	8.10	8.73	5.51	7.92
$\alpha = 0.05$	0.19	0.59	−0.12	0.59	−0.06	0.47	4.10	5.58	5.84	7.86	7.52	0.11
$\alpha = 0.1$	0.69	1.05	0.26	0.86	0.33	0.74	6.24	8.23	8.72	11.47	15.23	21.18
$\alpha = 0.2$	−0.28	0.70	1.08	1.73	−0.42	0.67	10.96	14.10	15.05	19.45	15.23	21.18
ext bwo	0.09	0.86	0.74	1.48	0.16	0.87	0.03	1.32	0.07	1.34	0.19	1.17
m-m[a] 1	0.14	0.84	0.55	1.25	0.09	0.81	0.19	1.39	0.09	1.30	0.06	1.01
m-m 2	0.23	0.95	0.76	1.41	0.18	0.79	0.03	1.26	−0.20	1.09	−0.36	0.76
rs[b] 1	0.38	1.19	2.14	3.79	9.61	13.15	0.26	0.83	10.05	12.33	large	large
rs 2	0.08	0.87	1.22	1.93	0.01	0.68	1.44	2.49	1.89	2.86	large	large

[a] Mirror-match 1: $n'_h = f_h n_h$; mirror-match 2: $n'_h = 1$.

[b] Rescale 1: resample size chosen to match 3rd moments; rescale 2: resample size = $n_h - 1$.

TABLE 5: Population 7: C.I.'s for the median.

	Nominal error rate in each tail							
	5%			10%			Std. length	
Method	L	U	$L+U$	L	U	$L+U$	$1-\alpha=0.8$	0.9
				$n_0 = 5$				
wood	6.0	2.2	8.2	5.8	17.6	23.4	1.26	0.77
ext bwo	5.8	7.0	12.8	5.8	17.6	23.4	0.76	0.77
m-m[a] 1	5.8	2.2	8.0	5.8	15.4	21.2	0.83	1.04
m-m 2	5.8	2.2	8.0	5.8	13.6	19.4	0.83	0.85
rs[b] 1	7.2	17.6	24.8	7.2	17.6	24.8	0.81	1.00
rs 2	9.0	6.2	15.2	9.2	6.2	15.2	0.69	0.89
				$n_0 = 7$				
wood	0.2	2.6	2.8	6.4	19.0	25.4	1.11	0.69
jack	18.6	8.0	26.6	21.6	8.0	29.6	1.06	1.06
ext bwo	9.0	2.0	11.0	9.0	2.0	11.0	0.77	0.98
m-m 1	9.0	2.0	11.0	9.0	12.8	21.8	0.76	0.96
m-m 2	1.8	2.2	4.0	9.0	15.8	23.8	1.02	0.73
rs 1	2.4	63.0	65.4	5.2	91.0	97.2	1.01	0.20
rs 2	2.6	10.2	12.8	11.2	11.8	23.0	0.83	0.74

[a] Mirror-match 1: $n'_h = f_h n_h$; mirror-match 2: $n'_h = 1$.

[b] Rescale 1: resample size chosen to match 3rd moments; rescale 2: resample size $= n_h - 1$.

In general, the results in Tables 3–5 indicate that the Woodruff method yields good results if the population values, at least moderately, overlap across strata (populations 5 and 6). Choosing $\alpha = 0.05$ seems to perform well on average. If the stratification is done using a highly correlated concomitant variable (populations 7 and 8), and thus there is little overlap across strata, then the Woodruff method performs poorly. The problems are due to the dependence of the median on only a few strata near the center of the population. This could occur in less extreme populations if the estimation of a small or large percentile were of interest. The BWO and mirror-match methods are much more robust to this situation. Note, however, the leap in relative bias for population 7, for all of the methods, as the within-stratum sample size increases from 5 to 6. This is a result of this same nonoverlapping structure. The median definition does not adjust for even and odd stratum sample sizes.

6. CONCLUSIONS

The mirror-match and BWO methods perform better on average over the situations considered here. It is not surprising that the rescaling method does well for stratified random sampling with the nonlinear functions of means considered. In such a simple situation a rescaling of i.i.d. resampling adequately approximates without-replacement sampling. One would expect that as the estimators and the sampling design become more complex, methods which better mimic the original sampling will perform better. From this viewpoint the mirror-match method and (perhaps even more) the extended BWO method are promising. The results for the median support this, especially where stratification is done by a highly correlated concomitant variable. Clearly, both theoretical and empirical study of these methods for more complex sampling plans is needed, and it is being pursued.

REFERENCES

Abromovitch, L., and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap, *Ann. Statist.*, 13, 116–132.

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

Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New York.

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

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Ann. Statist.*, 19, 454–469.

Gross, S. (1980). Median estimation in sample surveys. *Proc. Sect. Survey Res. Methods Amer. Statist. Assoc.*, 181–184.

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. (1975). Constructing orthogonal replications for variance estimation. *J. Amer. Statist. Assoc.*, 70, 819–821.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *J. Amer. Statist. Assoc.*, 78, 776–807.

Hansen, M.H., and Tepping, B.J. (1985). Estimation of variance in *N AEP*. Unpublished manuscript.

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.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Ann. Statist.*, 9, 1010–1019.

McCarthy, P.J., and Snowden, C.B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statist., Ser. 2, No. 95, Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc. Ser. B*, 24, 482–491.

Rao, J.N.K., and Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: Some recent work. *Bull. Internat. Statist. Inst.*, Proceedings of the 46th session, Book 3, 5–19.

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

Sitter, R. (1989). Resampling procedures for complex survey data. Ph.D. Thesis, Dept. of Statistics and Actuarial Science, University of Waterloo.

Sitter, R. (1992). A resampling procedures for complex survey data. *J. Amer. Statist. Assoc.*, 87, to appear.

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

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

Received 19 July 1990
Revised 3 July 1991
Accepted 29 July 1991

Department of Mathematics and Statistics
Carleton University
Ottawa, Ontario K1S 5B6