BOOTSTRAP INFERENCE FOR SAMPLE SURVEYS

J.N.K. Rao, Carleton University and C.F.J. Wu, University of Wisconsin-Madison

The bootstrap method of inference is extended to stratified sampling, involving a large number of strata with relatively few (primary) units sampled within strata, when the parameter of interest, θ, is a nonlinear function of the population mean vector. The bootstrap estimate of bias of the estimate, $\hat{\theta}$, of θ and the estimate of variance of $\hat{\theta}$ are obtained. Bootstrap confidence intervals for θ are also given, utilizing the percentile method or the bootstrap histogram of the t-statistic. Extensions to unequal probability sampling without replacement and two-stage sampling are also obtained.

1. Introduction

Resampling methods, including the jackknife and the bootstrap, provide standard error estimates and confidence intervals for the parameters of interest. These methods are simple and straightforward but are computer-intensive, especially the bootstrap. Efron (1982) has given an excellent account of resampling methods in the case of an independent and identically distributed (i.i.d.) sample of fixed size n from an unknown distribution F, and the parameter of interest $\theta = \theta(F)$. Limited empirical evidence (see Efron, 1982, p.18) has indicated that the bootstrap standard error estimates are likely to be more stable than those based on the jackknife and also less biased than those based on the customary delta (linearization) method. Moreover, the bootstrap histogram of t-statistic approximates the true distribution of t with a remainder term of $O_p(n^{-1})$ in the Edgeworth expansion (Abramovitch and Singh, 1984), unlike the usual normal approximation with a remainder term of $O_p(n^{-\frac{1}{2}})$. Empirical results for the ratio estimate (Hinkley and Wei, 1984) also indicate that the confidence intervals from the bootstrap histogram of t (using the linearization variance estimator) perform better than those based on the normal approximation.

The main purpose of this article is to propose an extension of the bootstrap method to stratified samples, in the context of sample survey data; especially to data obtained from stratified cluster samples involving large numbers of strata, L, with relatively few primary sampling units (psu's) sampled within each stratum. For nonlinear statistics $\hat{\theta}$ that can be expressed as functions of estimated means of p (≥ 1) variables, Krewski and Rao (1981) established the asymptotic consistency of the variance estimators from the jackknife, the delta and the balanced repeated replication (BRR) methods as $L \to \infty$ within the context of a sequence of finite populations $\{\Pi_L\}$ with L strata in Π_L. Their result is valid for any multistage design in which the psu's are selected with replacement and in which independent subsamples are selected within those psu's sampled more than once. Rao and Wu (1983) obtained second order asymptotic expansions of these variance estimators under the above set up and made comparisons in terms of their biases.

The proposed bootstrap method for stratified samples is described in Section 2 and the properties of the resulting variance estimator are studied. The bootstrap estimate of bias of is also obtained. Section 3 provides bootstrap confidence intervals for θ. The results are extended to stratified simple random sampling without replacement in Section 4. Finally, the method is extended to unequal probability sampling without replacement in Section 5 and to two-stage cluster sampling without replacement in Section 6.

2. The Bootstrap Method

The parameter of interest θ is a nonlinear function of the population mean vector $\bar{Y} = (\bar{Y}_1,\ldots,\bar{Y}_p)^T$, say $\theta = g(\bar{Y})$. This form of θ includes ratios, regression and correlation coefficients. If n_h (≥ 2) psu's are selected with replacement with probabilities p_{hi} in stratum h, then Krewski and Rao (1981) have shown that the natural estimator $\hat{\theta} = g(\hat{\bar{Y}})$ can be expressed as $\hat{\theta} = g(\bar{y})$. Here $\hat{\bar{Y}}$ is a design-unbiased linear estimator of $\bar{Y} = \Sigma W_h \bar{Y}_h$ and $\bar{y} = \Sigma W_h \bar{y}_h$ where W_h and $\bar{Y}_h = (\bar{Y}_{h1},\ldots,\bar{Y}_{hp})^T$ are the h-th stratum weight ($\Sigma W_h = 1$) and population mean vector respectively and $\bar{y}_h$ is the mean of n_h i.i.d. random vectors $y_{hi} = (y_{hi1},\ldots,y_{hip})^T$ for each h with $E(y_{hi}) = \bar{Y}_h$. For $h \neq h'$, y_{hi} and $y_{h'j}$ are independent but not necessarily identically distributed.

2.1. The Naive bootstrap. In the case of an i.i.d. sample $\{y_i\}_1^n$ with $E(y_i) = \bar{Y}$, the bootstrap method is as follows: (i) Draw a simple random sample $\{y_i^*\}_1^n$ with replacement from the observed values $y_1, y_2, \ldots, y_n$ and calculate $\hat{\theta}^* = g(\bar{y}^*)$ where $\bar{y}^* = \Sigma y_i^*/n$. (ii) Independently replicate step (i) a large number, B, of times and claculate the corresponding estimates $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$. (iii) The bootstrap variance estimator of $\hat{\theta} = g(\bar{y})$ is given by

$$v_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}_a^*)^2, \qquad (2.1)$$

where $\hat{\theta}_a^* = \Sigma\, \hat{\theta}^{*b}/B$. The Monte-Carlo estimator $v_b(a)$ is an approximation to

$$v_b = \mathrm{var}_*(\hat{\theta}^*) = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2, \qquad (2.2)$$

where E_* denotes the expectation with respect to bootstrap sampling from a given sample $y_1,\ldots,y_n$. No closed-form expression for $\mathrm{var}_*(\hat{\theta}^*)$ generally exists in the nonlinear case, but in the linear case with $p = 1$, $\hat{\theta}^* = \bar{y}^*$ and v_b reduces to

$$\mathrm{var}_*(\bar{y}^*) = \frac{n-1}{n^2} s^2 = \frac{n-1}{n} \mathrm{var}(\bar{y}), \qquad (2.3)$$

where $(n-1)s^2 = \Sigma(y_i - \bar{y})^2$ and $\mathrm{var}(\bar{y}) = s^2/n$

is the unbiased estimator of variance of $\bar{y}$. The modified variance estimator $[n/(n-1)]\mathrm{var}_*(\hat{\theta}^*)$ exactly equals $\mathrm{var}(\bar{y})$ in the linear case, but Efron (1982) found no advantage in this modification. In any case, $n/(n-1) \doteq 1$ in most applications and $\mathrm{var}_*(\hat{\theta}^*)$ is a consistent estimator of the variance of $\hat{\theta}$, as $n \to \infty$ (Bickel and Freedman, 1981). The bootstrap histogram of $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$ may be used to find confidence intervals for θ . This method (Efron, 1982) is called the percentile method.

Noting the i.i.d. property of the y_{hi}'s within each stratum, a straightforward extension of the usual bootstrap method to stratified samples (referred to as the "naive bootstrap") is as follows: (i) Take a simple random sample $\{y^*_{hi}\}_{i=1}^{n_h}$ with replacement from the given sample $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h , independently for each stratum. Calculate $\bar{y}^*_h = n_h^{-1}\Sigma_i y^*_{hi}$, $\bar{y}^* = \Sigma W_h \bar{y}^*_h$ and $\hat{\theta}^* = g(\bar{y}^*)$. (ii) Independently replicate step (i) a large number, B , of times and calculate the corresponding estimates $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$. (iii) The bootstrap variance estimator of $\hat{\theta} = g(\bar{y})$ is given by

$$v_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^*_a)^2 , \tag{2.4}$$

where $\bar{y} = \Sigma W_h \bar{y}_h$ and $\hat{\theta}^*_a = \Sigma\hat{\theta}^{*b}/B$. The Monte-Carlo estimator $v_b(a)$ is an approximation to

$$v_b = \mathrm{var}_*(\hat{\theta}^*) = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2 , \tag{2.5}$$

where E_* denotes the expectation with respect to bootstrap sampling. In the linear case with $p = 1$, $\hat{\theta}^* = \Sigma W_h \bar{y}^*_h = \bar{y}^*$ and v_b reduces to

$$\mathrm{var}_*(\bar{y}^*) = \Sigma \frac{W_h^2}{n_h}\left(\frac{n_h - 1}{n_h}\right)s_h^2 , \tag{2.6}$$

where $(n_h-1)s_h^2 = \Sigma_i(y_{hi} - \bar{y}_h)^2$. Comparing (2.6) with the unbiased estimator of variance of $\bar{y}$, $\mathrm{var}(\bar{y}) = \Sigma W_h^2 s_h^2/n_h$, it immediately follows that $\mathrm{var}_*(\bar{y}^*)/\mathrm{var}(\bar{y})$ does not converge to 1 in probability, unless L is fixed and $n_h \to \infty$ for each h . Hence, $\mathrm{var}_*(\bar{y}^*)$ is not a consistent estimator of the variance of $\bar{y}$. It also follows that v_b is not a consistent estimator of the variance (or mean square error) of a general nonlinear statistic. There does not seem to be an obvious way to correct this scaling problem except when $n_h = k$ for all h in which case $k(k-1)^{-1}\mathrm{var}_*(\hat{\theta}^*)$ will be consistent. Bickel and Freedman (1984) also noticed the scaling problem, but they were mainly interested in bootstrap confidence intervals in the linear case $(p = 1)$. They have established the asymptotic $N(0,1)$ property of the distribution of $t = (\bar{y} - \bar{Y})/[\mathrm{var}(\bar{y})]^{\frac{1}{2}}$ and of the conditional distribution of $(\bar{y}^* - \bar{y})/[\mathrm{var}_*(\bar{y}^*)]^{\frac{1}{2}}$ in stratified simple random sampling with replacement, and also proved that $(\Sigma W_h^2 s_h^{*2}/n_h)/\mathrm{var}_*(\bar{y}^*)$ converges to 1 in probability as $n = \Sigma n_h \to \infty$, where $(n_h-1)s_h^{*2} = \Sigma_i(y^*_{hi} - \bar{y}^*_h)^2$. Their result implies that one could use the bootstrap histogram of $\tilde{t}^{*1},\ldots,\tilde{t}^{*B}$ to find confidence intervals for $\bar{Y}$, where $\tilde{t}^{*b} = (\bar{y}^{*b} - \bar{y})/[\Sigma W_h^2 s_h^{*b2}/n_h]^{\frac{1}{2}}$ where s_h^{*b2} is the value of s_h^{*2} for the b-th bootstrap sample $(b = 1,\ldots,B)$. In the nonlinear case, there does not seem to be a simple way to construct $\tilde{t}^{*b}$-values similar to those of Bickel and Freedman since v_b has no closed form. Moreover, the straightforward extension of the bootstrap (hereafter called the naive bootstrap) does not permit the use of the percentile method based on the bootstrap histogram of $\hat{\theta}^{*1},\ldots,\hat{\theta}^{*B}$.

Although $\tilde{t}^{*b}$ is asymptotically $N(0,1)$ in the linear case, it is not likely to provide as good an approximation to the distribution of t as a statistic whose denominator and numerator are both adequate approximations to their counterparts in t . Such statistics will be proposed in Section 3.2. These are also applicable to the nonlinear case.

Recognizing the scaling problem in a different context, Efron (1982) suggested to draw a bootstrap sample of size $n_h - 1$ instead of n_h from stratum h $(h = 1,\ldots,L)$. In Section 2.2, we will instead propose a different method which includes his suggestion as a special case.

2.2 <u>The proposed method</u>. Our method is as follows: (i) Draw a simple random sample $\{y^*_{hi}\}_{i=1}^{m_h}$ of size m_h with replacement from $\{y_{hi}\}_{i=1}^{n_h}$. Calculate

$$\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_h-1)^{-\frac{1}{2}}(y^*_{hi} - \bar{y}_h)$$

$$\tilde{y}_h = m_h^{-1}\sum_{i=1}^{m_h}\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_h-1)^{-\frac{1}{2}}(\bar{y}^*_h - \bar{y}_h) \tag{2.7}$$

$$\tilde{y} = \Sigma W_h \tilde{y}_h , \quad \tilde{\theta} = g(\tilde{y}) .$$

(ii) Independently replicate step (i) a large number, B , of times and calculate the corresponding estimates $\tilde{\theta}^1,\ldots,\tilde{\theta}^B$. (iii) The bootstrap estimator $E_*(\tilde{\theta})$ of θ can be approximated by $\hat{\tilde{\theta}}_a = \Sigma\tilde{\theta}^b/B$. The bootstrap variance estimator of $\hat{\theta}$ is given by

$$\tilde{\sigma}_b^2 = \tilde{v}_b = \mathrm{var}_*(\tilde{\theta}) = E_*(\tilde{\theta} - E_*\tilde{\theta})^2 \tag{2.8}$$

with its Monte-Carlo approximation

$$\tilde{\sigma}_b^2(a) = \tilde{v}_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\theta}^b - \tilde{\theta}_a)^2. \tag{2.9}$$

One can replace $E_*\tilde{\theta}$ in (2.8) by $\hat{\theta}$.

2.3. <u>Justification of the method</u>. In the linear case, $\theta = \bar{Y}$, $\tilde{v}_b$ reduces to the customary unbiased variance estimator $\mathrm{var}(\bar{y}) =$

$\Sigma W_h^2 s_h^2/n_h$ for any choice m_h. In the nonlinear case, it can be shown that

$$v_b = v_L + O_p(n^{-2}) \quad (2.10)$$

for any m_h, where v_L is the linearization variance estimator (Rao and Wu, 1983). In the linear case (p=1), v_L reduces to $\text{var}(\bar{y})$. Under reasonable regularity conditions, v_L is a consistent estimator of variance of $\hat{\theta}$, $\text{Var}(\hat{\theta})$. Hence, it follows from (2.10) that $\tilde{v}_b$ is also consistent for $\text{Var}(\hat{\theta})$.

The asymptotic $N(0,1)$ property of the conditional distribution of $(\tilde{\theta} - \hat{\theta})/\tilde{\sigma}_b$ can be established, assuming that $0 < \delta_1 \leq m_h/(n_h-1) \leq \delta_2 < \infty$ for all h, i.e. the bootstrap sample size m_h should be comparable to the original size n_h in each stratum.

2.4. Estimate of bias of $\hat{\theta}$. Our bootstrap estimate of bias is

$$\tilde{B}(\hat{\theta}) = E_*(\tilde{\theta}) - \hat{\theta} \quad (2.11)$$

which is approximated by $\tilde{\theta}_a - \theta$. It can be shown that $\tilde{B}(\hat{\theta})$ is a consistent estimator of $B(\hat{\theta})$ (Rao and Wu, 1983). On the other hand, the bias estimate $B(\hat{\theta}) = E_*(\hat{\theta}^*) - \hat{\theta}$, based on the naive bootstrap, is not a consistent estimator of $B(\hat{\theta})$ (Rao and Wu, 1983).

3. Confidence Intervals

We now consider different bootstrap methods for setting confidence intervals for θ.

3.1. Percentile method. For ready reference, we now give a brief account of the percentile method based on the bootstrap histogram of $\tilde{\theta}^1,\ldots,\tilde{\theta}^B$. Define the cumulative bootstrap distribution function as

$$\widehat{CDF}(t) = \#\{\tilde{\theta}^b \leq t;\ b = 1,\ldots,B\}/B . \quad (3.1)$$

For $\alpha \leq 0.5$, define $\tilde{\theta}_{LOW}(\alpha) = \widehat{CDF}^{-1}(\alpha)$ and $\tilde{\theta}_{UP}(\alpha) = \widehat{CDF}^{-1}(1-\alpha)$. Then the interval

$$\{\tilde{\theta}_{LOW}(\alpha),\ \tilde{\theta}_{UP}(\alpha)\} \quad (3.2)$$

is an approximate $(1-2\alpha)$-level confidence interval for θ. It has the central $1-2\alpha$ portion of the bootstrap distribution (Efron, 1982, p.78). One can also consider a bias-corrected percentile method, following Efron (1982, p.82). This method leads to

$$\{\widehat{CDF}^{-1}(\Phi(2z_0 - z_\alpha)),\ \widehat{CDF}^{-1}(\Phi(2z_0 + z_\alpha))\} \quad (3.3)$$

as an approximate $(1-2\alpha)$-level confidence interval for θ, where Φ is the cumulative distribution function of a standard normal, $z_0 = \Phi^{-1}(\widehat{CDF}(\hat{\theta}))$ and $z_\alpha = \Phi^{-1}(1-\alpha)$. The advantage of the interval (3.3) over (3.2) has been demonstrated by Efron (1982) in the i.i.d. case.

3.2. Bootstrap t-statistics. Instead of approximating the distribution of $\hat{\theta}$ by the bootstrap distribution of $\tilde{\theta}$, we can approximate the distribution of the t-statistic $t=(\hat{\theta}-\theta)/\tilde{\sigma}_b$ by its bootstrap counterpart $t^* = (\tilde{\theta} - \hat{\theta})/\tilde{\sigma}_b^*(a)$ where $\tilde{\sigma}_b^{*2}(a) = v_b^*(a)$ is the bootstrap variance estimator obtained from (2.9) by bootstrapping the particular bootstrap sample $\{\tilde{y}_{hi}\}$ i.e. by replacing y_{hi} by $\tilde{y}_{hi}$ in the proposed method. For the second phase bootstrapping one could choose values (m_h', B') different from (m_h, B). This double-bootstrap method thus leads to B values $t^{*1},\ldots,t^{*B}$ of t^*. Utilizing the bootstrap histogram of $t^{*1},\ldots,t^{*B}$, we define $\widehat{CDF}_t(x) = \#\{t^{*b} \leq x\}/B$, $\tilde{t}_{LOW} = \widehat{CDF}_t^{-1}(\alpha)$, $\tilde{t}_{UP} = \widehat{CDF}^{-1}(1-\alpha)$, and construct an approximate $(1-2\alpha)$-level confidence interval for θ given by

$$\{\hat{\theta} - \tilde{t}_{UP}\tilde{\sigma}_b,\ \hat{\theta} - \tilde{t}_{LOW}\tilde{\sigma}_b\} . \quad (3.4)$$

We now provide an asymptotic justification for t^*. Noting that $\tilde{v}_b^*(a)$ is a Monte Carlo approximation to

$$\tilde{\sigma}_b^{*2} = \tilde{v}_v^* = E_{**}(\tilde{\theta}^* - E_{**}\tilde{\theta}^*)^2 , \quad (3.5)$$

where $\tilde{\theta}^*$ is the value of $\tilde{\theta}$ obtained from bootstrapping the particular sample $\{\tilde{y}_{hi}\}$ and E_{**} is the second phase bootstrap expectation, we can write $t^* = (\tilde{\theta}-\hat{\theta})/\tilde{\sigma}_b^*$. In the linear case $\theta = \bar{Y}$ it is easily seen that

$$E_*\tilde{\sigma}_b^{*2} = \tilde{\sigma}_b^2 . \quad (3.6)$$

In the nonlinear case, following Bickel and Freedman (1984), we can show that $\tilde{\sigma}_b^{*2}/\tilde{\sigma}_b^2$ converges to 1 in probability as $n \to \infty$. Hence, it follows that the conditional distribution of t^* is asymptotically $N(0,1)$.

One could use a jackknife t-statistic $t_J = (\hat{\theta} - \theta)/\hat{\sigma}_J$ instead of t, where $\hat{\sigma}_J^2$ is a jackknife variance estimator of $\hat{\theta}$ (see Krewski and Rao, 1981). The corresponding confidence interval is then given by

$$\{\hat{\theta} - \hat{t}_{UP}\hat{\sigma}_J,\ \hat{\theta} - \hat{t}_{LOW}\hat{\sigma}_J\} \quad (3.7)$$

where $\hat{t}_{LOW}$ and $\hat{t}_{UP}$ are the lower and upper α-points of the statistic $t_J^* = (\tilde{\theta} - \hat{\theta})/\hat{\sigma}_J^*$ obtained from the bootstrap histogram of $t_1^{*1},\ldots,t_J^{*B}$, and $\hat{\sigma}_J^{*2}$ is obtained from $\hat{\sigma}_J^2$ by jackknifing the particular bootstrap sample $\{\tilde{y}_{hi}\}$. It can be shown that the confidence interval (3.7) is also asymptotically correct. A confidence interval of this type was considered by Efron (1981) in the case of an i.i.d. sample $\{y_i\}$.

It is possible to replace $\hat{\sigma}_J^2$ by the BRR or the linearization variance estimator and obtain a confidence interval similar to (3.7).

3.3. Choice of m_h. The choice $m_h = n_h$ is a natural one. The choice $m_h = n_h-1$ gives $\tilde{y}_{hi} = y_{hi}^*$ and our method reduces to the naive bootstrap, except that in step (i) of the latter method a simple random sample of size n_h-1 is selected from $\{y_{hi}\}_{i=1}^{n_h}$ is stratum h. For $n_h = 2$, $m_h = 1$, the method reduces to the well-known random half-sample replication and the resulting variance estimators are less stable than those obtained from BRR for the same number, B, of half samples (McCarthy, 1969). It may be worth considering a bootstrap sample size m_h larger than $n_h = 2$, say $m_h = 3$ or 4.

In the linear case $(p=1)$, it can be shown that the bootstrap third moment, $E_*(\tilde{y} - \bar{y})^3$, matches the unbiased estimate of the third moment of $\bar{y}$, if $m_h = (n_h-2)^2/(n_h-1)$, $n_h \geq 4$. This property is not enjoyed by the jackknife or the BRR. Similarly, it can also be shown that the same choice of m_h ensures that the bootstrap histogram of a t-statistic (see section 3.2) approximates the true distribution of t with a remainder term of $O_p(n^{-1})$. Details will be provided in a separate paper.

4. Stratified Simple Random Sampling Without Replacement

All the previous results apply to the case of stratified simple random sampling without replacement by making a slight change in the definition of $\tilde{y}_{hi}$:

$$\tilde{y}_{hi} = \bar{y}_h + m_h^{\frac{1}{2}}(n_h-1)^{-\frac{1}{2}}(1-f_h)^{\frac{1}{2}}(y_{hi} - \bar{y}_h), \quad (4.1)$$

where $f_h = n_h/N_h$ is the sampling fraction in stratum h. It is interesting to observe that, even by choosing $m_h = n_h-1$, $\tilde{y}_{hi} \neq y^*_{hi}$. Hence the naive bootstrap using y^*_{hi} will still have the problem of giving a wrong scale as discussed before. In the special case of $n_h = 2$ for all h, McCarthy (1969) used a finite population correction similar to (4.1) in the context of BRR. It can be shown that the choice

$$m_h = [(n_h-2)^2/(n_h-1)][(1-f_h)/(1-2f_h)^2] \quad (4.2)$$

matches the third moments, but it provides an approximation to the true distribution of the t-statistic with a remainder term of $O_p(n^{-\frac{1}{2}})$ only, as in the case of a normal approximation.

Bickel and Freedman (1984) considered a different bootstrap sampling method in order to recover the finite population correction, $1-f_h$, in the variance formula. This method essentially creates populations consisting of copies of each y_{hi}, $i=1,\ldots,n_h$ and $h=1,\ldots,L$ and then generates $\{y^*_{hi}\}_1^{n_h}$ as a simple random sample without replacement from the created population, independently in each stratum. This "blow-up" bootstrap was first proposed by Gross (1980) and also independently by Chao and Lo (1983). The variance estimator resulting from this method (by working directly with y^*_{hi}), however, remains inconsistent for estimating the true variance of $\hat{\theta}$. It is possible, however, to make the variance estimator consistent by reducing the bootstrap sample size to n_h-1, as in Section 2.

The "blow-up" bootstrap approximates the true distribution of the t-statistic with a remainder term of $O_p(n^{-1})$ (see Abramovitch and Singh, 1984), unlike our method.

5. Unequal Probability Sampling Without Replacement

5.1. The Rao, Hartley and Cochran Method. Rao, Hartley and Cochran (1962) proposed a simple method of sampling with unequal probabilities and without replacement. The population of N units is partitioned at random into n groups $G_1,\ldots,G_n$ of sizes $N_1,\ldots,N_n$ respectively $(\Sigma N_k = N)$, and then one unit is drawn from each of the n groups with probabilities p_t/P_k for the k-th group. Here $p_t = x_t/X$, $P_k = \sum_{t\in G_k} p_t$, x_t = a measure of size of the t-th unit $(t=1,\ldots,N)$ that is approximately proportional to y_t (in the scalar case of $p=1$) and $X = \Sigma x_t$. Their unbiased estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}} = \sum_{k=1}^{n} z_k P_k \quad (5.1)$$

where $z_k = y_k/(Np_k)$ and (y_k, p_k) denote the values for the unit selected from the k-th group $(\Sigma P_k = 1)$. An unbiased estimator of variance of $\hat{\bar{Y}}$ is

$$var(\hat{\bar{Y}}) = \lambda^2 \Sigma P_k (z_k - \hat{\bar{Y}})^2, \quad (5.2)$$

where

$$\lambda^2 = (\Sigma N_k^2 - N)/(N^2 - \Sigma N_k^2). \quad (5.3)$$

In the special case of equal group sizes, $N_k = N/n$, λ reduces to $(1-f)^{\frac{1}{2}}(n-1)^{-\frac{1}{2}}$ where $f = n/N$.

Our bootstrap sampling for the above method is as follows: (1) Attach the probability P_k to the sampled unit from G_k and then select a sample $\{y^*_i, p^*_i\}_{i=1}^{m}$ of size m with replacement with probabilities P_k from $\{y_k, p_k\}_{k=1}^{n}$. Calculate $z^*_i = y^*_i/(Np^*_i)$,

$$\tilde{z}_i = \hat{\bar{Y}} + \lambda m^{\frac{1}{2}}(z^*_i - \hat{\bar{Y}})$$

$$\tilde{\bar{Y}} = m^{-1}\sum_{i=1}^{m} \tilde{z}_i = \hat{\bar{Y}} + \lambda m^{\frac{1}{2}}(\bar{z}^* - \hat{\bar{Y}}) \quad (5.4)$$

$$\tilde{\theta} = g(\tilde{\bar{Y}}),$$

where $\bar{z}^* = \Sigma z^*_i/m$. (2) Implement steps (ii) and (iii) of Section 2.2, using $\tilde{\theta}$ obtained from (5.4).

In the linear case with $p=1$, $\theta = \bar{Y}$, we get

$$E_*(\tilde{\bar{Y}}) = E_*(\tilde{z}_i) = \hat{\bar{Y}} + \lambda m^{\frac{1}{2}}(\hat{\bar{Y}} - \hat{\bar{Y}}) = \hat{\bar{Y}} \quad (5.5)$$

and

$$var_*(\tilde{\bar{Y}}) = \frac{1}{m} var_*(\tilde{z}_i) = \lambda^2 var_*(z^*_i)$$
$$= \lambda^2 \sum_{k=1}^{n} P_k(z_k - \hat{\bar{Y}})^2 = var(\hat{\bar{Y}}). \quad (5.6)$$

Hence, the bootstrap variance estimator, $var_*(\tilde{\bar{Y}})$, reduces to $var(\hat{\bar{Y}})$ in the linear case. The properties of $var_*(\tilde{\theta})$ in the nonlinear case and the bootstrap methods of setting confidence intervals for θ, given in Section 3, are being investigated.

5.2 General Results. A general linear unbiased estimator of the population total Y (in the scalar case of $p=1$) based on a sample s of size n with associated probability of selection $p(s)$, is given by

$$\hat{Y} = \sum_{i\epsilon s} d_i(s) y_i \ , \qquad (5.7)$$

where the known weights $d_i(s)$ can depend on s and the unit i $(i \epsilon s)$. Rao (1979) proved that a nonnegative unbiased quadratic estimator of variance of $\hat{Y}$ is necessarily of the form

$$\text{var}(\hat{Y}) = -\sum_{\substack{i<j\\ \epsilon s}}\sum d_{ij}(s)\, w_i w_j (z_i - z_j)^2, \qquad (5.8)$$

where $z_i = y_i/w_i$, the w_i are known constants such that $\text{Var}(\hat{Y}) = 0$ when $y_i \propto w_i$. Here the known coefficients $d_{ij}(s)$ satisfy the unbiasedness condition

$$Ed_{ij}(s) = \text{Cov}[d_i(s), d_j(s)] \ , \qquad (5.9)$$

where $d_i(s) = 0$ if $i \not\epsilon s$, and $d_{ij}(s) = 0$ if s does not contain both i and j. As an example, the well-known Horvitz-Thompson estimator $\hat{Y}_{HT}$ satisfies (5.7) with $d_i(s) = 1/\pi_i$, $i \epsilon s$ and $w_i = \pi_i$ and

$$-d_{ij}(s) w_i w_j = \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}, \qquad (5.10)$$

where $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$ are the first and second order inclusion probabilities. Let $b_{ij}(s) = -d_{ij}(s)/2$ and assume that $b_{ij}(s) > 0$ for all $i < j \epsilon s$ and all s. This is true for many well-known schemes, including the Rao-Hartley-Cochran (RHC) method.

Our bootstrap sampling is as follows: (1) consider all the $n(n-1)$ pairs (i,j), $i \neq j$ and select m pairs (i^*,j^*) with replacement with probabilities λ_{ij} $(= \lambda_{ji})$ to be specified. Calculate

$$\tilde{Y} = \hat{Y} + \frac{1}{m} \sum_{(i^*,j^*)\epsilon s^*} k_{i^*j^*}(z_{i^*} - z_{j^*}) \qquad (5.11)$$

$$\tilde{\theta} = g(\tilde{Y}) \ ,$$

where $k_{ij} = k_{ji}$ for $i < j \epsilon s$ to be specified, and s^* denotes the set of m bootstrap pairs. (2) Implement steps (ii) and (iii) of Section 2.2 using $\tilde{\theta}$ obtained from (5.11).

In the linear case with $p=1$, $\theta = Y$, we get

$$E_*(\tilde{Y}) = \hat{Y} + E_*\{k_{i^*j^*}(z_{i^*} - z_{j^*})\}$$

$$= \hat{Y} + \sum_{\substack{i\neq j\\ \epsilon s}}\sum k_{ij}\lambda_{ij}(z_i - z_j) \ . \qquad (5.12)$$

Letting $\sum_{j(\neq i)} k_{ij}\lambda_{ij} = c_i$ and noting that $k_{ij}\lambda_{ij} = k_{ji}\lambda_{ji}$, we get

$$E_*(\tilde{Y}) = \hat{Y} + \sum_{i\epsilon s} z_i(c_i - c_i) = \hat{Y} \ .$$

Similarly,

$$\text{var}_*(\tilde{Y}) = \frac{1}{m} E_*\{k_{i^*j^*}(z_{i^*} - z_{j^*})\}^2$$

$$= \frac{1}{m} \sum_{\substack{i\neq j\\ \epsilon s}}\sum k_{ij}\lambda_{ij}(z_i - z_j)^2. \qquad (5.13)$$

We now choose $k_{ij}^2\lambda_{ij}/m$ such that (5.13) is identical to $\text{var}(\hat{Y})$ given by (5.8), i.e.,

$$\text{var}(\hat{Y}) = \sum_{\substack{i\neq j\\ \epsilon s}}\sum b_{ij}(s) w_i w_j (z_i - z_j)^2 \ .$$

Thus

$$\frac{1}{m} k_{ij}^2 \lambda_{ij} = b_{ij}(s)\, w_i w_j \ . \qquad (5.14)$$

The choice of k_{ij} and λ_{ij} satisfying (5.14) is not unique. One simple choice is $\lambda_{ij} = 1/[n(n-1)]$, i.e. equal probabilities for each of the $n(n-1)$ pairs of $i \neq j$. Another choice is

$$\lambda_{ij} = \pi_{ij}/d_s \ , \qquad (5.15)$$

where $d_s = \sum_{\substack{i\neq j\\ \epsilon s}}\sum \pi_{ij}$.

Special case. For the Horvitz-Thompson estimator, we have

$$\frac{1}{m} k_{ij}^2 \lambda_{ij} = \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \ , \qquad (5.16)$$

using (5.10) in (5.14). If $m = n(n-1)$ and $\lambda_{ij} = 1/[n(n-1)]$, then (5.16) reduces to

$$k_{ij}^2 = n^2(n-1)^2 \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \ . \qquad (5.17)$$

Hence, $\tilde{Y}_{HT}$ reduces to

$$\tilde{Y}_{HT} = \hat{Y}_{HT} + \sum_{(i^*,j^*)=1}^{m} \left[\frac{\pi_{i^*}\pi_{j^*} - \pi_{i^*j^*}}{\pi_{i^*j^*}}\right]^{\frac{1}{2}} (z_{i^*} - z_{j^*}). \ (5.18)$$

It follows from (5.13) that $\text{var}_*(\tilde{Y}_{HT})$ reduces to the well-known Yates-Grundy variance estimator:

$$\text{Var}(\hat{Y}_{HT}) = \sum_{i<j\epsilon s}\sum \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} (z_i - z_j)^2. \qquad (5.19)$$

The properties of $\text{var}_*(\tilde{\theta})$ in the nonlinear case and the bootstrap confidence intervals on θ are being investigated.

Remark. In the case of RHC method, we now have two different bootstrap methods: (1) The method of Section 6.1 based on selecting a bootstrap sample with probabilities P_k and with replacement from $\{y_k, p_k\}_1^n$. (2) The present method of selecting a bootstrap sample of pairs with replacement from the $n(n-1)$ pairs $(i,j) \in s$, $i \neq j$, with probabilities λ_{ij}. We are investigating their relative merits, but intuitively method (1) is more appealing.

6. Two-stage Cluster Sampling Without Replacement

Suppose that the population is comprised of N clusters with M_t elements (subunits) in the t-th cluster $(t = 1, \ldots, N)$. The population size M_0 $(= \Sigma M_t)$ is unknown in many applications. A simple random sample of n clusters is selected without replacement, and m_i elements are chosen, again by simple random sampling without replacement, from the M_i elements in i-th cluster if the latter is selected. The customary unbiased estimator of the population total Y is

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i , \qquad (6.1)$$

where $\bar{y}_i$ is the sample mean for i-th sample cluster. The corresponding estimator of θ is written as $\hat{\theta} = g(\hat{\bar{Y}})$, where

$$\hat{\bar{Y}} = \frac{\hat{Y}}{M_0} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{Y}_i}{\bar{M}_0}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{M_i y_{ij}}{\bar{M}_0} , \qquad (6.2)$$

where $\bar{M}_0 = M_0/N$. For instance, if $\theta = Y/M_0$ where M_0 is unknown, then $\hat{\theta} = \hat{\bar{Y}}/\hat{\bar{X}}$ where $\hat{\bar{X}} = \hat{X}/M_0$ and $\hat{X} = (N/n) \Sigma M_i \bar{x}_i$ with $\bar{x}_i = 1$ for all elements j in any cluster i. An unbiased estimator of variance of $\hat{\bar{Y}}$ is given by (Cochran, 1977, p.303)

$$\text{var}(\hat{\bar{Y}}) = \frac{(1-f_1)}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{Y}_i}{\bar{M}_0} - \hat{\bar{Y}}\right)^2 + \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{(1-f_{2i})}{m_i(m_i-1)} \sum_{j=1}^{m_i} \left(\frac{M_i y_{ij}}{\bar{M}_0} - \frac{\hat{Y}_i}{\bar{M}_0}\right)^2 , \qquad (6.3)$$

where y_{ij} is the y-value for j-th sample element in the i-th sample cluster, $f_1 = n/N$ and $f_{2i} = m_i/M_i$.

We employ two-stage bootstrap sampling to obtain $\{y_{ij}^{**}\}$ from $\{y_{ij}\}$ as follows: (1) Select a simple random sample of n clusters with replacement from the n sample clusters and then draw a simple random sample of m_i elements with replacement from the m_i elements in i-th sample cluster if the latter is chosen. (Independent bootstrap subsampling for the same cluster chosen more than once.) We use the following notation: y_{ij}^{**} = y-value of the j-th bootstrap element in the i-th bootstrap cluster; $m_i^* = m_i$-value of the i-th bootstrap cluster (similarly M_i^*), and $\hat{Y}_i^* = \hat{Y}_i$-value of the i-th bootstrap cluster. Calculate

$$\tilde{y}_{ij} = \hat{\bar{Y}} + \lambda_1 \left(\frac{\hat{Y}_i^*}{\bar{M}_0} - \hat{\bar{Y}}\right) + \lambda_{2i}^* \left(\frac{M_i^* y_{ij}^{**}}{\bar{M}_0} - \frac{\hat{Y}_i^*}{\bar{M}_0}\right) \qquad (6.4)$$

and

$$\tilde{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{m} \frac{1}{m_i^*} \sum_{j=i}^{m_i^*} \tilde{y}_{ij}$$

$$= \hat{\bar{Y}} + \frac{\lambda_1}{n} \sum_{i=1}^{n} \left(\frac{\hat{Y}_i^*}{\bar{M}_0} - \hat{\bar{Y}}\right) + \frac{1}{n} \sum_{i=1}^{n} \lambda_{2i}^* \left(\frac{\hat{Y}_i^{**}}{\bar{M}_0} - \frac{\hat{Y}_i^*}{\bar{M}_0}\right), \qquad (6.5)$$

where $\hat{Y}_i^{**} = M_i^* \bar{y}_i^{**}$, $\bar{y}_i^{**} = \Sigma_j\, y_{ij}^{**}/m_i^*$, and

$$\lambda_1^2 = \frac{n}{n-1}(1-f_1) \; ; \; \lambda_{2i}^{*2} = f_1(1-f_{2i}^*) \frac{m_i^*}{m_i^*-1} . \qquad (6.6)$$

(2) Implement steps (ii) and (iii) of Section 2.2 with $\tilde{\theta} = g(\tilde{\bar{Y}})$.

Let E_{2*} and var_{2*} respectively denote the conditional bootstrap expectation and variance, for a given bootstrap sample of clusters. Similarly, E_{1*} and var_{1*} denote the bootstrap expectation and variance respectively for the sample clusters. Then

$$E_{2*}(\tilde{\bar{Y}}) = \hat{\bar{Y}} + \frac{\lambda_1}{n} \Sigma \left(\frac{\hat{Y}_i^*}{\bar{M}_0} - \hat{\bar{Y}}\right)$$

and

$$E_*(\tilde{\bar{Y}}) = E_{1*}E_{2*}(\tilde{\bar{Y}}) = \hat{\bar{Y}} + \lambda_1(\hat{\bar{Y}} - \hat{\bar{Y}}) = \hat{\bar{Y}} . \qquad (6.7)$$

Similarly,

$$\text{var}_{1*}E_{2*}(\tilde{\bar{Y}}) = \lambda_1^2 \, \text{var}_{1*}\left(\frac{\hat{Y}_i^*}{\bar{M}_0} - \hat{\bar{Y}}\right)$$

$$= \frac{\lambda_1^2}{n} \Sigma \left(\frac{\hat{Y}_i}{\bar{M}_0} - \hat{\bar{Y}}\right) \qquad (6.8)$$

Also

$$\sum_{i=1}^{n} \lambda_{2i} \frac{1}{m_i{}^2} \sum_{j=1}^{m_i} \left(\frac{M_i y_{ij}}{\bar{M}_0} - \frac{\hat{Y}_i}{\bar{M}_0}\right)^2 , \qquad (6.9)$$

where $\lambda_{2i}^2 = f_1(1-f_{2i})m_i/(m_i-1)$. Hence, combining (6.8) and (6.9) we get

$$\text{var}_*(\tilde{\bar{Y}}) = \text{var}_{1*}E_{2*}(\tilde{\bar{Y}}) + E_{1*}\text{var}_{2*}(\tilde{\bar{Y}}) = \text{var}(\hat{\bar{Y}}) .$$

Hence, the bootstrap variance estimator, $\text{var}_*(\tilde{\bar{Y}})$, reduces to $\text{var}(\hat{\bar{Y}})$ in the linear case. Thus the variability between bootstrap estimates correctly accounts for the between-cluster and within-cluster components of the variance without the necessity of estimating them separately as in the case of BRR or the jackknife.

Acknowledgement.
This research was supported by a grant from

the Natural Sciences and Engineering Research Council of Canada. C.F.J. Wu is partially supported by NSF grant No. MCS-83-00140. Our thanks are due to David Binder for useful discussions on the bootstrap for unequal probability sampling without replacement.

References

Abramovitch, L. and Singh, K. (1984). Edgeworth corrected pivotal statistics and the bootstrap. Preprint.

Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9, 1196-1217.

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

Chao, M.T. and Lo, S.H. (1983). A bootstrap method for finite population. Preprint.

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

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian J. Statist.*, 9, 139-172.

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

Gross, S. (1980). Median estimation in sample surveys. *Proc. Amer. Statist. Assoc.*, Section on Survey Research Methods, 181-184.

Hinkley, D. and Wei, B. (1984). Improvements of jackknife confidence limit methods. *Biometrika*, 71, 331-339.

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. (1969). Pseudoreplication: half samples. *Rev. Int. Statist. Inst.*, 37, 239-264.

Rao, J.N.K. and Wu, C.F.J. (1983). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, No.7.

Rao, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators in finite population sampling. *J. Indian Statist. Assoc.*, 17, 125-136.

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

Rao, J.N.K. and Wu, C.F.J. (1983). Bootstrap inference with stratified samples. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, No. 19.