A Resampling Technique for Complex Survey Data[1]

Tauqueer Ahmad
IASRI, New Delhi

SUMMARY

In this article a new bootstrap technique of variance estimation for complex survey data known as "Rescaling Bootstrap Without Replacement" has been developed for some important sampling designs. The comparison of proposed bootstrap variance estimation technique has been done with different methods of bootstrap and jackknifed variance estimation through simulation. It has been observed that for estimation of variance of a non-linear statistic, the proposed rescaling bootstrap without replacement technique works well and its performance is almost comparable with the jackknifed variance estimator for most of the non-linear statistics considered in this study.

Key words: Bootstrap, Jackknife, Linearization, Balanced repeated replications, Variance estimation, Probability sampling, Two-stage cluster sampling.

1. Introduction

There are several methods of variance estimation out of which bootstrap method is the most recent technique of variance estimation for complex sample surveys. The other commonly used methods for variance estimation are the method of random groups, linearization or Taylor series method, the jackknife method and balanced repeated replications (BRR) etc. Recently, considerable attention has been paid to the problems of estimation of variance for non-linear statistics such as ratio, regression and correlation coefficients etc. Relatively, recent in origin, the bootstrap variance estimation technique is highly computer-intensive resampling procedure which substitutes considerable amount of computation in place of theoretical analysis. The bootstrap method needs no prior assumptions about the distribution of observations as well as the estimators. It provides estimates of bias and standard error apart from estimates of other distributional properties of the estimators however complex it may be.

---

1 The paper was presented on the 1st July 1997 at a special session of the Society held at IASRI, New Delhi and adjudged best for "ISAS Young Scientist Award, 1996".

The asymptotic consistency of some of these variance estimation techniques for non-linear functions have been shown by Krewski and Rao [8]. Kish and Frankle [7], Bean [1], Campbell and Mayor [3] etc. have studied the properties of these estimators. Rao and Wu [10] made second order asymptotic comparisons of some of these variance estimators. The naive bootstrap technique was suggested by Efron [4]. Efron [5] indicated that bootstrap may be better than its competitors, but there is a dispute on this issue. Attempts have been made to develop resampling techniques specially in case of bootstrap for variance estimation and confidence intervals in sample survey data (where the sampling is without replacement) such that resampling procedure parallels the original sampling scheme as closely as possible by Gross [6], McCarthy and Snowden [9], Rao and Wu [11], Sitter [12] etc. The different bootstrap variance estimation techniques available in the literature in case of complex surveys can be broadly classified into two categories (i) rescaling of pseudo values according to original sampling design and (ii) attempt to develop resampling procedures close to the original sampling design. The ultimate aim of both the approaches is to get approximately unbiased estimator of variance for the parameter of interest. It has been proved in different situations that in case of linear statistics, the resampling variance estimator reduces to the customary unbiased variance estimator under both the approaches. In this article an attempt has been made to develop a new bootstrap variance estimation technique namely "Rescaling Bootstrap Without Replacement" by combining both the approaches. An optimum choice of bootstrap sample size in case of proposed technique has also been obtained which is relatively smaller than other existing techniques. Further, this technique has been extended for some important sampling designs such as (a) Stratified Random Sampling Without Replacement and (b) Two Stage Cluster Sampling Without Replacement. Also, the proposed bootstrap technique has been compared with different existing methods of bootstraps and jackknifed variance estimation for non-linear statistics through simulation.

2. *Existing Bootstrap Methods*

Let us consider stratified random sampling in case of finite population, consisting of N units, which is partitioned into L nonoverlapping strata of N_1, N_2, ... , N_L units; thus, $N_1 + N_2 + ... + N_L = N$. A simple random sample without replacement (SRSWOR) is taken independently from each stratum. The sample size within each stratum are denoted by n_1, n_2, ..., n_L and the total sample size is $n = n_1 + n_2 + ... + n_L$. The parameter of interest θ is nonlinear function of the population mean vector $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, .., \bar{Y}_p)$ say $\theta = g(\bar{Y})$. This form includes ratios, regression and correlation coefficients. In this case, the unbiased estimator of $\bar{\bar{Y}}$ is

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)$$

where $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = (\bar{y}_{1h}, \bar{y}_{2h}, \ldots, \bar{y}_{ph})$ and $W_h = \frac{N_h}{N}$

For $p = 1$, an unbiased estimate of Var $(\bar{y})$ is

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

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

If the standard i. i. d. bootstrap is applied to the sample data $\left(y_{hi}^*\right)_{i=1}^{n_h}$ in each stratum, then the resulting resampling algorithm would take the following form :

1. Draw a simple random sample $\left(y_{hi}^*\right)_{i=1}^{n_h}$ with replacement from the original sample $\left(y_{hi}\right)_{i=1}^{n_h}$ in stratum h, independently for each stratum.

   Calculate

$$\bar{y}_h^* = \frac{1}{n_h} \sum_i y_{hi}^* , \; \bar{y}^* = \sum W_h \bar{y}_h^* \text{ and } \hat{\theta}^* = g(\bar{y}^*)$$

2. Repeat step 1, a large number of times, say B, to get $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$.
3. Estimate Var $(\hat{\theta})$ with

$$v_b = E_*(\hat{\theta}^* - E_* \hat{\theta}^*)^2 \tag{2.2}$$

   or its Monte Carlo approximation

$$v_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^*_{(.)})^2 \qquad (2.3)$$

where $\hat{\theta}^*_{(.)} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}$ and E_* refers to the expectation with respect to bootstrap sampling.

In the linear case with $p = 1$, $\hat{\theta}^* = \sum W_h \bar{y}^*_h = \bar{y}^*$, v_b reduces to

$$v_b = var_* (\bar{y}^*) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\frac{n_h - 1}{n_h} \right) s_h^2 \qquad (2.4)$$

Comparing (2.4) with the standard unbiased variance estimate, var $(\bar{y})$, given in equation (2.1), it is seen that v_b in not a consistent estimator for Var $(\bar{y})$. This can be avoided by using a correction factor only if $n_h = k$ and $f_h = f$ for all h, in which case $\frac{k}{k-1}$ (1–f) $var_* (\hat{\theta}^*)$ is consistent.

Recognizing this scaling problem, Efron [5] suggested to draw bootstrap sample of size $n_h - 1$ with SRSWR sampling designs instead of n_h independently from each stratum. Rest of the procedure is same as in the case of naive bootstrap technique.

Rao and Wu [11] proposed a rescaling of the standard bootstrap when $\hat{\theta} = g(\bar{y})$, a non-linear function of means. In this method one applies the previously stated algorithm with a general resample size m_h not necessarily equal to n_h, but rescales the resampled values appropriately so that the resulting variance estimator is the same as the usual unbiased variance estimator in the linear case.

In addition to these with-replacement bootstrap (BWR) techniques, a without replacement bootstrap (BWO) was proposed by Gross [6] in the case of a single stratum. Suppressing the h-th subscript, this method assumes $N = kn$ for some integer k and creates a pseudopopulation of size N by replicating the data k times. Although the BWO method is intuitively appealing, it does not yield the usual unbiased estimate of variance in the linear case. Bickel and Freedman [2] proposed a randomization between two pseudopopulations that corrects this problem and allows an extension to $L > 1$; however, their method is applicable only for some stratified samples.

Sitter [12] developed a method of bootstrapping which retains the desirable properties of BWR and BWO but extends to more complex without replacement sampling designs. In general the method entails:

(a) Selecting a subsample without replacement from the original sample to mirror the original sampling scheme.

(b) Replacing this subsample in the original sample and

(c) Repeating this a specified number of times k_h.

This bootstrapping procedure is repeated a large number of times. The value of k_h is chosen such that the bootstrap estimate of variance matches the usual one in the linear case.

3. *The Proposed Method*

It seems reasonable to design a resampling scheme that parallels the original sampling scheme as closely as possible. This is what so appealing about the BWO method as compared to the BWR methods and the rescaling methods. Therefore a new BWO technique known as "Rescaling Bootstrap Without Replacement (RS BWO)" has been proposed. The proposed technique is as follows :

1. Draw a simple random sample $\left(y_i^*\right)_{i=1}^{m}$ of size m without replacement from the observed values $y_1, y_2, ..., y_n$.
   Calculate

$$\tilde{y}_i = \bar{y} + (1-f)^{1/2} \frac{\sqrt{m}}{\sqrt{n-m}} (y_i^* - \bar{y}) \qquad (3.1)$$

$$\tilde{y} = \frac{1}{m} \sum_{i=1}^{m} \tilde{y}_i \text{ and } \tilde{\theta} = g(\tilde{y}) \text{ where } f = \frac{n}{N}$$

   Obviously, one of the basic assumption of this technique is that n is sufficiently large as compared to m.

2. Replace the sample in the original sample and independently replicate step 1. Repeat this process a large number, say B, of times and calculate the corresponding estimates $\tilde{\theta}^1, \tilde{\theta}^2, \ldots, \tilde{\theta}^B$.

3. The bootstrap variance estimator of $\tilde{\theta} = g(\tilde{y})$ is given by

$$\tilde{v}_b = E_* (\tilde{\theta} - E_* \tilde{\theta})^2 \qquad (3.2)$$

where E_* denotes the expectation with respect to bootstrap sampling from a given sample.

The Monte Carlo estimator $\tilde{v}_b(a)$ as an approximation to $\tilde{v}_b$ is given by

$$\tilde{v}_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\tilde{\theta}^b - \tilde{\theta}_a)^2 \tag{3.3}$$

where $$\tilde{\theta}_a = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^b$$

In the linear case, $\theta = \bar{Y}$ and $p = 1$, $\tilde{v}_b$ reduces to the usual unbiased variance estimator var $(\bar{y})$ for any choice of m, noting that

$$\tilde{v}_b = E_*(\tilde{\bar{y}} - \bar{y}) = \text{var}_*(\tilde{\bar{y}}) = (1-f)\frac{s^2}{n}$$
$$= \text{var}(\bar{y})$$

3.1 The Rescaling Bootstrap Without Replacement Technique for Stratified Random Sampling

The proposed RS BWO technique can easily be extended for stratified sampling, which is as follows:

1. Draw a simple random sample $\left(y^*_{hi}\right)_{i=1}^{m_h}$ of size m_h without replacement from the given sample $\left(y_{hi}\right)_{i=1}^{n_h}$ of size n_h in stratum h, independently for each stratum.

   Calculate

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

$$\tilde{\bar{y}}_h = \frac{1}{m_h} \sum_{i=1}^{m_h} \tilde{y}_{hi}$$
$$= \bar{y}_h \left(1 - (\sqrt{1-f_h}) \frac{\sqrt{m_h}}{\sqrt{n_h - m_h}} \right)$$
$$+ \left((\sqrt{1-f_h}) \frac{\sqrt{m_h}}{\sqrt{n_h - m_h}} \right) \bar{y}^*_h \tag{3.1.1}$$

where $\bar{y}_h^*$ and $\bar{y}_h$ are the bootstrap sample mean and original sample mean for the h-th stratum; n_h is sufficiently large as compared to m_h.

$$\tilde{y} = \sum_{h=1}^{L} W_h \tilde{y}_h, \qquad \tilde{\theta} = g(\tilde{y})$$

2. Replace the sample in the original sample and independently replicate step 1. Repeat this process a large number of times, say B, of times and calculate the corresponding estimates $\tilde{\theta}^1, \tilde{\theta}^2, \ldots, \tilde{\theta}^B$.
3. The bootstrap estimator $E_*(\tilde{\theta})$ of θ can be approximated by

$\tilde{\theta}_{(.)} = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}^b$. The bootstrap variance estimator of $\hat{\theta}$ is given by

$$\tilde{v}_b = E_* (\tilde{\theta} - E_* \tilde{\theta})^2 \tag{3.1.2}$$

with its Monte Carlo approximation

$$\tilde{v}_b(a) = \frac{1}{B-1} \sum_{b=1}^{B} (\tilde{\theta}^b - \tilde{\theta}_{(.)})^2 \tag{3.1.3}$$

In the linear case, $\theta = \bar{Y}$ and $p = 1$, $\tilde{v}_b$ reduces to the usual unbiased variance estimator var $(\bar{y})$, for any choice of m_h, noting that

$$\text{var}_* (\tilde{y}_h) = (1 - f_h) \frac{m_h}{n_h - m_h} \left[\frac{n_h - m_h}{n_h m_h} \right] s_h^2$$

$$= (1 - f_h) \frac{s_h^2}{n_h}$$

Hence

$$\tilde{v}_b = \text{var}_* (\tilde{y}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

$$= \text{var}(\bar{y})$$

Optimum choice of Bootstrap Sample Sizes m_h

The optimum choice of bootstrap sample sizes has been obtained using the rescaling factor of the pseudo values. The rescaling factor is $(1 - f_h)^{1/2} \left(\frac{\sqrt{m_h}}{\sqrt{n_h - m_h}} \right)$.

Putting rescaling factor is equal to one, we get

$$m_h = \frac{n_h}{(2 - n_h/N_h)}$$

It can be seen that for optimum choice m_h, $\tilde{y}_{hi} = y^*_{hi}$ and the proposed method reduces to the naive bootstrap; however, in the latter method's step 1, a simple random sample of size $\frac{n_h}{(2 - n_h/N_h)}$ is selected from $\left(y_{hi}\right)_{i=1}^{n_h}$ in stratum h.

3.2 The Rescaling Bootstrap Without Replacement Technique for Two-Stage Cluster Sampling Without Replacement

The proposed RS BWO technique is now extended to two stage cluster sampling with equal probabilities and without replacement. The technique is justified by showing that in the linear case the variance estimator reduces to the usual variance estimator for the two stage sampling designs.

Suppose that the population comprises N clusters with M_i elements (subunits) in the i-th cluster (i = 1, ..., N). The population size M_o (= $\sum M_i$) is unknown in many situations. 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 the i-th cluster if the latter is selected. The customary unbiased estimator of the population total Y is

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} M_i \bar{y}_i = \frac{N}{n} \sum_{i=1}^{n} \hat{Y}_i \qquad (3.2.1)$$

where $\bar{y}_i$ is the sample mean for the i-th sample cluster.

The corresponding estimator of θ is written as $\hat{\theta} = g(\hat{\bar{Y}})$, where

$$\hat{\bar{Y}} = \frac{\hat{Y}}{M_o} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{Y}_i}{\overline{M}_o} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{M_i y_{ij}}{\overline{M}_o} \qquad (3.2.2)$$

where y_{ij} is the y-value for the j-th sample element in the i-th cluster and $\overline{M}_o = \frac{M_o}{N}$. For instance, if $\theta = \frac{Y}{M_o}$ where M_o is unknown, then $\hat{\theta} = \frac{\hat{Y}}{\hat{X}}$ where

$\hat{\bar{X}} = \frac{\hat{X}}{M_o}$ and $\hat{X} = \frac{N}{n} \sum M_i \bar{x}_i$ with $x_{ij} = 1$ for all elements j in any cluster i. An unbiased estimator of variance of $\hat{\bar{Y}}$ is given by

$$\text{var}(\hat{\bar{Y}}) = \frac{1-f_1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{Y}_i}{\bar{M}_o} - \hat{\bar{Y}} \right)^2 + \frac{f_1}{n^2} \sum_{i=1}^{n} \frac{(1-f_{2i})}{m_i(m_i-1)} \sum_{j=1}^{m_i} \left(\frac{M_i y_{ij}}{\bar{M}_o} - \frac{\hat{Y}_i}{\bar{M}_o} \right)^2 \quad (3.2.3)$$

where $f_1 = \frac{n}{N}$ and $f_{2i} = \frac{m_i}{M_i}$

Now the RS BWO technique for two-stage cluster sampling without replacement can be proposed to obtain $\left(y_{ij}^{**}\right)$ form $\left(y_{ij}\right)$ which is as follows:

1. Select a simple random sample of n′ clusters without replacement from the n sample clusters and then draw a simple random sample of m'_i elements without replacement from the m_i elements in the i-th sample cluster if the latter is chosen. (Independent bootstrap subsampling for the same cluster chosen more than once). We use the following notations:

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 resample cluster.

Calculate

$$\tilde{y}_{ij} = \hat{\bar{Y}} + \lambda'_1 \left(\frac{\hat{Y}_i^*}{\bar{M}_o} - \hat{\bar{Y}} \right) + \lambda_{2i}'^{*} \left(\frac{M_i^* y_{ij}^{**}}{\bar{M}_o} - \frac{\hat{Y}_i^*}{\bar{M}_o} \right) \quad (3.2.4)$$

where $\lambda_1'^2 = \frac{n}{n-1}(1-f) f_1^* = \lambda_1^2 f_1^*$

$$\lambda_{2i}'^{2*} = \frac{m_i^*}{(m_i^*-1)} f_1 (1-f_{2i}^*) f_{2i}^{**} f_1^* \left(\frac{m_i^*-1}{m_i^*-m_i'^*} \right)$$

$$= \lambda_{2i}^{*2} f_{2i}^{**} f_1^* \left(\frac{m_i^*-1}{m_i^*-m_i'^*} \right)$$

where

$$f_1^* = \frac{n'}{n},\ f_{2i}^* = \frac{m_i^*}{M_i^*},\ f_{2i}^{**} = \frac{m_i'^*}{m_i^*}$$

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

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

$$= \hat{\bar{Y}} + \frac{\lambda'_1}{n'}\sum_{i=1}^{n'}\left(\frac{\hat{Y}_i^*}{\bar{\bar{M}}_o} - \hat{\bar{Y}}\right) + \frac{1}{n'}\sum_{i=1}^{n'}\lambda_{2i}'^*\left(\frac{\hat{Y}_i^{**}}{\bar{\bar{M}}_o} - \frac{\hat{Y}_i^*}{\bar{\bar{M}}_o}\right) \tag{3.2.5}$$

where $\hat{Y}_i^{**} = M_i^*\ \bar{y}_i^{**},\ \bar{y}_i^{**} = \frac{1}{m_i'^*}\sum_{j=1}^{m_i'^*} y_{ij}^{**}$

2. Implement steps 2 and 3 of section 3 with $\tilde{\theta} = g(\tilde{\bar{Y}})$.

Let E_{2*} and var_2^* denote the conditional expectation and variance respectively, 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

$$E_{2*}(\tilde{\bar{Y}}) = \hat{\bar{Y}} + \frac{\lambda'_1}{n'}\sum_{i=1}^{n'}\left(\frac{\hat{Y}_i^*}{\bar{\bar{M}}_o} - \hat{\bar{Y}}\right)$$

and

$$E_*(\tilde{\bar{Y}}) = E_{1*}\ E_{2*}(\tilde{\bar{Y}}) = \hat{\bar{Y}} + \frac{\lambda'_1}{n'}\sum_{i=1}^{n'}(\hat{\bar{Y}} - \hat{\bar{Y}}) = \hat{\bar{Y}} \tag{3.2.6}$$

Similarly,

$$var_1^*\ E_{2*}(\tilde{\bar{Y}}) = \frac{\lambda_1'^2}{n'}\ var_1^*\left(\frac{\hat{Y}_i^*}{\bar{\bar{M}}_o} - \hat{\bar{Y}}\right)$$

$$= \frac{\lambda_1^2}{n^2}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{\bar{\bar{M}}_o} - \hat{\bar{Y}}\right) \tag{3.2.7}$$

Also

$$\text{var}_2^* (\tilde{\tilde{Y}}) = \frac{1}{n'^2} \sum_{i=1}^{n'} \lambda'^{*2}_{2i} \, \text{var}_2^* \left(\frac{\hat{y}_i^{**}}{\overline{\overline{M}}_o} - \frac{\hat{Y}_i^*}{\overline{\overline{M}}_o} \right)$$

$$= \frac{1}{n\,n'} \sum_{i=1}^{n'} \lambda^{*2}_{2i} \frac{1}{m_i^{*2}} \sum_{j=1}^{m_i^*} \left(\frac{M_i^* \, y_{ij}^*}{\overline{\overline{M}}_o} - \frac{\hat{Y}_i^*}{\overline{\overline{M}}_o} \right)^2$$

Now

$$E_{1*} \, \text{var}_2^* (\tilde{\tilde{Y}}) = \frac{1}{n^2} \sum_{i=1}^{n'} \lambda^2_{2i} \frac{1}{m_i^2} \sum_{j=1}^{m_i} \left(\frac{M_i \, y_{ij}}{\overline{\overline{M}}_o} - \frac{\hat{Y}_i}{\overline{\overline{M}}_o} \right)^2 \quad (3.2.8)$$

Hence, by combining (3.2.7) and (3.2.8) we get

$$\text{var}^* (\tilde{\tilde{Y}}) = \text{var}_1^* \; E_{2*} (\tilde{\tilde{Y}}) + E_{1*} \, \text{var}_2^* (\tilde{\tilde{Y}}) = \text{var} (\hat{Y})$$

Hence, the bootstrap variance estimator, var* $(\tilde{\tilde{Y}})$, reduces to var $(\hat{Y})$ in the linear case.

4. A Simulation Study

In order to study the performance of the proposed new resampling technique as compared to various existing methods commonly used in variance estimation for stratified random sampling without replacement, a simulation study has been taken. A trivariate normal population of size 5000 units have been generated. The parameters i.e. mean vector and variance-covariance matrix for generating above population were obtained with the help of Census (1981) data of villages from Sujangarh and Ratangarh Tehsils of Churu district from Rajasthan, India. The area of the village under irrigation is the character under study (y), whereas geographical area of the village has been treated as auxiliary character (x). The population is stratified on the basis of number of household in the village into five equal strata each of size 1000 units. Stratified sampling design with proportional allocation has been adopted to select 100 samples each of size 200 units from this population. Further from each of these samples 200 bootstrap samples have been generated using the following resampling methods:

(i) Naive bootstrap method (BWR M1)

(ii) Bootstrap with replacement method (BWR M2)

(iii) Rescaling bootstrap with replacement method (RS BWR)

(iv) Bootstrap without replacement (Gross) method (GS BWO)

(v) Bootstrap without replacement (Sitter) method (SR BWO)

(vi) The proposed Rescaling bootstrap without replacement method (RS BWO)

(vii) Jackknife method (JACK).

The bootstrap estimates and estimates of variance have been obtained from each of the above methods for the parameters of interest i.e. population mean, ratio, correlation coefficient and regression coefficient. Assuming the notation for stratified sampling of Section 2, let y_{hi} denote the characteristic of interest of the i-th observation from the h-th stratum and let x_{hi} denote a related auxiliary variable. Let $y_{1hi} = x_{hi}$, $y_{2hi} = y_{hi}$, $y_{3hi} = y_{hi}\, x_{hi}$, $y_{4hi} = x^2_{hi}$ and $y_{5hi} = y^2_{hi}$. The sample estimates of the corresponding population parameter used are :

(1) the mean, $\bar{y} = \bar{y}_2$, (2) the ratio, $r = \dfrac{\bar{y}_2}{\bar{y}_1}$, (3) the correlation coefficient,

$$C = \frac{\bar{y}_3 - \bar{y}_1\, \bar{y}_2}{[\,(\bar{y}_4 - \bar{y}_1^2)\ (\bar{y}_5 - \bar{y}_2^2)\,]^{1/2}}$$

and (4) the regression coefficient,

$$b = \frac{\bar{y}_3 - \bar{y}_1\, \bar{y}_2}{\bar{y}_4 - \bar{y}_1^2}$$

where

$$\bar{y}_j = \sum_{h=1}^{L} W_h\ \bar{y}_{jh} \quad \text{and} \quad \bar{y}_{jh} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{jhi}$$

for j = 1, 2, ..., 5

The mean square error [MSE $(\hat{\theta})$] for different statistics are obtained from a separate simulation study of 2000 samples each of size 200, with each sample having $n_h = 40$ from each stratum. The MSE $(\hat{\theta})$ obtained from this simulation are

$V(\bar{y}) = 8.9570313$, MSE (r) = 2.53 E–6

MSE (c) = 1.05 E –3, MSE (b) = 4.21 E–6

The percentage relative bias (R.B.) and relative stability (R.S.) of estimates of variance estimators are calculated as follows:

$$\% \text{ R.B.} = \left(\frac{E(\hat{V}(\hat{\theta}))}{MSE(\hat{\theta})} - 1 \right) \times 100$$

and

$$\text{R.S.} = \frac{\sqrt{MSE\{\hat{V}(\hat{\theta})\}}}{MSE(\hat{\theta})}$$

The results for the mean, ratio, correlation coefficient and regression coefficient are presented in the table. The following points can be observed from the results of the study:

1. In case of mean, the proposed bootstrap variance estimator is less biased and more stable than other bootstrap variance estimators but as compared to jackknife the bias and stability are almost of the order of jackknife variance estimator.
2. In the case of ratio, the proposed bootstrap variance estimator is having less % R.B. and more stable than any other variance estimators and hence is preferable to all the available estimators.
3. In the case of estimation of variance of correlation coefficient, the proposed variance estimator is less biased than other variance estimators. Also it is more stable than other bootstrap variance estimators and less stable (but quite close) than jackknife variance estimator.
4. In the case of regression coefficient, the proposed bootstrap variance estimator is less biased and more stable than the other bootstrap variance estimators but is more biased and less stable than jackknife estimator.

5. Conclusions

In this article a new bootstrap technique of variance estimation is proposed for complex survey data known as "Rescaling Bootstrap Without Replacement" which parallels the original sampling scheme. The technique has been extended for some important sampling designs such as (a) Stratified Random Sampling Without Replacement and (b) Two Stage Cluster Sampling Without Replacement. It can be extended to other complex survey designs also, with appropriate modifications. The proposed bootstrap technique yields variance estimators for non-linear statistics $\hat{\theta}$ which can be expressed as functions of means. All of these variance estimators reduce to the standard ones in the special case of linear statistics. The performance of the proposed technique is better than the other similar bootstrap variance estimation techniques in terms of percentage relative bias and relative stablility. Also, its performance is almost comparable with the jackknifed variance estimator except in the case of estimation variance of regression coefficient, where jackknifed variance estimator performs well as far as % relative bias and relative stability are concerned. The proposed bootstrap variance estimator is simple, precise and easy to apply in practical situations as compared to its counterparts.

Table. Monte Carlo estimates, estimates of variance, relative biases, and relative stabilities of mean, ratio, correlation and regression coefficients for different techniques under stratified sampling without replacement ($n_h = 40$)

Method	Estimate $\hat{\theta}$	Estimate of variance $V(\hat{\theta})$	% Relative bias (R.B.)	Relative stability (R.S.)
		Mean		
BWR M1 ($m_h = 40$)	69.4430847	9.4990845	6.0517060	1.125392E–01
BWR M2 ($m_h = 39$)	69.4415894	9.8029509	9.4441961	1.40379E–01
RS BWR ($m_h = 30$)	69.0640411	9.7805214	9.1937839	1.35669E–01
GS BWO ($m_h = 40$)	69.423248	9.158215	2.246094	1.07283E–01
SR BWO ($n'_h = 10$, $k_h = 3.12$)	69.428946	9.3872595	4.8032455	1.159411E–01
RS BWO ($m_h = 20$)	66.6718292	9.1411734	2.0558385	1.007585E–01
Jack	66.2075653	8.779997	–1.9764540	0.98855606E–01
		Ratio		
BWR M1 ($m_h = 40$)	3.509707E–02	2.340243E–06	–7.3634204	1.093665E–01
BWR M2 ($m_h = 39$)	3.509479E–02	2.368041E–06	–6.2920459	1.082131E–01
RS BWR ($m_h = 30$)	3.487158E–02	2.364558E–06	–6.4107637	1.150430E–01
GS BWO ($m_h = 40$)	3.5095E–02	2.2311E–06	–11.693776	1.48390E–01
SR BWO ($n'_h = 10$, $k_h = 3.12$)	3.508889E–02	2.315345E–06	8.3893945	1.202700E–01
RS BWO ($m_h = 20$)	3.510283E–02	2.408318E–06	–4.7091427	1.031972E––1
Jack	3.329721E–02	2.096146E–06	–17.0557974	2.038463E–01

Method	Estimate $\hat{\theta}$	Estimate of variance $V(\hat{\theta})$	% Relative bias (R.B.)	Relative stability (R.S.)
		Correlation		
BWR M1 ($m_h = 40$)	5.022699E–01	3.314248E–03	214.5993951	2.167551
BWR M2 ($m_h = 39$)	5.017058E–01	3.341067E–03	217.1451434	2.198514
RS BWR ($m_h = 30$)	5.075139E–01	3.319107E–03	215.0606275	2.1764514
GS BWO ($m_h = 40$)	5.017112E–01	3.133404E–03	197.433151	1.998460
SR BWO ($n'_h = 10$, $k_h = 3.12$)	5.017856E–01	3.322302E–03	215.3639075	2.1795442
RS BWO ($m_h = 20$)	5.424193E–01	2.870624E–03	128.6055196	2.7464198
Jack	5.525825E–01	2.536659E–03	140.7880723	1.4964436
		Regression		
BWR M1 ($m_h = 40$)	1.731737E–02	5.438919E–06	29.1611719	3.171752E–01
BWR M2 ($m_h = 39$)	1.729015E–02	5.488562E–06	30.3490857	3.346007E–01
RS BWR ($m_h = 30$)	1.755523E–02	5.488300E–06	30.3253384	3.340691E–01
GS BWO ($m_h = 40$)	1.727700E–02	5.1382E–06	22.031340	2.49507E–01
SR BWO ($n'_h = 10$, $k_h = 3.12$)	1.728899E–02	5.441182E–06	29.2092140	3.218570E–01
RS BWO ($m_h = 20$)	1.871268E–02	5.092135E–06	20.9213963	2.411363E–01
Jack	1.964026E–02	4.379153E–06	3.9895512	2.283236E–01

REFERENCES

[1] Bean, J.A., 1975. Distribution and properties of variance estimator for complex multistage probability samples. *Vital and Health Statistics*, **2**, 65, Washington, D.C.: U.S. Government Printing Office.

[2] Bickel, P.J. and Freedman, D.A., 1984. Asymptotic Normality and the Bootstrap in Stratified Sampling. *Ann. Statist.*, **12**, 470-482.

[3] Campbell, C. and Mayor, M., 1978. Some properties of T confidence intervals for survey data. *Proc. Amer. Statist. Assoc.*, 437-442.

[4] Efron , B., 1979. Bootstrap methods : Another look at Jackknife. *Ann. Statist.*, **7**, 1-26.

[5] Efron, B., 1982. Transformation theory : how normal is a one parameter family of distributions? *Ann. Statist.*, **10**, 322-339.

[6] Gross, S., 1980. Median estimation in sample surveys. *Proc. of Survey Research Methods Section, Amer. Statist. Assoc.*, 181-184.

[7] Kish, L. and Frankel, M.R., 1974. Inference from stratified samples (with discussion). *Jour. Roy. Statist. Soc.*, **B 36**, 1-37.

[8] Krewski, D. and Rao, J.N.K., 1981. Inference from stratified sample properties of the linearization, Jackknife and Balanced Repeated Replication Methods. *Ann. Statist.*, **9**, 1010-1019.

[9] McCarthy, P.J. and Snowden, C.B., 1985. The bootstrap and finite population sampling. *Vital and Health Statist.*, **2**, **95**, 85-1369.

[10] Rao, J.N.K. and Wu, C.F.J., 1985. Inference from stratified samples; Second order analysis of three methods for non-linear statistics. *Jour. Amer. Statist. Assoc.*, **80**, 620-630.

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

[12] Sitter, R.R., 1992. A resampling procedure from complex survey data. *Jour. Amer. Statist. Assoc.*, **87**, 755-765.