Using the Bootstrap to Estimate the Variance in a Very Complex Sample Design

Steven Kaufman, National Center for Education Statistics
room 9059,1990 K St. NW, Washington, D.C 2006

Key Words: BHR , NAEP, Super-Population

1.0 Introduction

Three commonly used methodologies for variance estimation are the balanced half-sample replication (BHR), the jackknife and the bootstrap. BHR was initially developed for finite population sampling, but of the three methods, it has the most restrictive assumptions. BHR requires that exactly two primary sampling units (PSUs) be selected with replacement within each stratum.

The jackknife and the bootstrap were initially developed for situations where some type of observations ($\theta_i, i = 1 \text{ to } n$) are independent and identically distributed (iid). For the jackknife, θ_i are the jackknife pseudo-values for an estimate θ. An additional assumption is $V(\theta_i) = V(\sqrt{n}\theta)$. For the bootstrap, θ_i are the actual observations and an additional requirement is an estimate for the cumulative distribution function of θ_i.

Because of the iid assumption, neither the jackknife nor the bootstrap can directly be used with finite population sampling. Instead, modifications are required which adjust the variance estimates for the dependent sampling. The jackknife usually clusters the final stage units, so that the clusters are approximately iid within each stratum. For the bootstrap, the bootstrap estimator is compared with an unbiased sample variance estimator and some parameter(s) (e.g., the bootstrap sample size) is (are) adjusted to eliminate any bias.

Comparing these three approaches, one might surmise the bootstrap would be more flexible than either the BHR or the jackknife, since it has more flexibility in adjusting for the dependent sampling. As such, one might expect the bootstrap to work best in very complex sample designs, assuming an appropriate adjustment can be derived to correct for the dependence in the sampled units.

Given the flexibility of the bootstrap, this paper will describe a bootstrap methodology for variance estimation that can be applied to a variety of complex sample designs. It will then be applied to a very complex sample design. This design will be a two-stage design, where the first stage is a randomized systematic probability proportional to size sample (PPS) (Kaufman, 1999) and the second stage is a simple random sample without replacement (SRS w/o replacement). What makes the design very complex is the requirement that the variance estimation must appropriately reflect the without replacement aspect of the PPS selection, but reflect a with replacement SRS, instead of the actual w/o replacement SRS. (Section 4.0, explains why such a variance estimator is useful.) After the bootstrap variance estimator is derived, a simulation study will measure its performance.

2.0 The Bootstrap Variance

Assume $x_i, i = 1 \text{ to } n$, are iid, where x_i is generated from a distribution function $F(x)$. $\hat{T}_n$ is some total generated from the x_i's. Let an estimate of $V(\hat{T}_n)$ be $v(\hat{T}_n) = \sigma(F, \hat{T}_n)$, where σ represents some function of F and $\hat{T}_n$. The bootstrap variance estimate is defined to be $v^*(\hat{T}_n) = \sigma(\hat{F}, \hat{T}_n)$, where $\hat{F}$ is an estimate of F. If $\hat{F}$ is the empirical distribution function of x_i and $\hat{T}_n$ is $\bar{X}_n$ then $v^*(\bar{X}_n) = \sigma^2 / n$, where $\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 / n$.

The Monte Carlo approximation for $V(\hat{T}_n)$, $v^*(\hat{T}_n^*)$, is $1/(B-1) \cdot \sum_{b=1}^{B} (\hat{T}_{bn}^* - \bar{T}_n^*)^2$, where $\hat{T}_{bn}^*$ is the bootstrap analog of $\hat{T}_n$ using n^* independent selections from $\hat{F}$. This is independently repeated B times to get B $\hat{T}_{bn}^*$'s Since $\sigma(\hat{F}, \hat{T}_n)$ may be unknown or extremely complex, $v^*(\hat{T}_n^*)$ is often used as $v^*(\hat{T}_n)$.

Using stratified SRS w/o replacement, let $n_h^* = n_h$, let T be the sample mean $\bar{X}$ and let $\hat{F}_h(x)$ be the empirical distribution function of x. $E^* v^*(\hat{T}^*) = \sum_{h=1}^{H} 1/n_h (n_h - 1)/n_h s_h^2$. E^* is the expectation with respect to the bootstrap sampling, s_h^2 is the usual unbiased estimate of the stratum population variance and H is the number of stratum h. This estimator is biased because of the $(n_h - 1)/n_h$ term, as well as the missing finite population correction (FPC) $1 - n_h / N_h$.

So, a naïve application of the bootstrap does not work in the finite population setting. However, by choosing an appropriate n_h^*, an unbiased bootstrap estimator can be obtained. See example 3.1.

3.0 Bootstrap Distribution Function

In this discussion, the bootstrap will be defined in terms of the sampling process rather than in terms of a specific variable of interest (i.e., the object is to generate a set of bootstrap samples). The advantage of this is that once the bootstrap samples are generated, there is no need to repeat the resampling process for each variable. In fact, a set of bootstrap replicate weights can be generated similar to BHR replicate weights (Kaufman, 1999).

In this context, let $\mathbf{I}_{n_h^*.} = G(\mathbf{I}_{n_h}, n_h^*)$, where $\mathbf{I}_{n_h^*.}$ is a vector representing a bootstrap PSU sample of size n_h^* and $G(\mathbf{I}_{n_h}, n_h^*)$ is some random mechanism generating $\mathbf{I}_{n_h^*.}$ that's a function of the original PSU sample $\mathbf{I}_{n_h}$ and n_h^*. $G(\mathbf{I}_{n_h}, n_h^*)$ must preserve the first-order expectations for every n_h^* (i.e., $E^*\hat{T}_h^* = \hat{T}_h$ for all n_h^*, $\hat{T}_h$ being the full-sample estimated stratum total).

$G(\cdot)$ can be a function of more variables. In Sitter's mirror-match bootstrap (1992), $(\mathbf{I}^1_{n_h^*.}, \cdots, \mathbf{I}^{k_h}_{n_h^*.}) = G(\mathbf{I}_{n_h}, k_h, n_h^*)$, where k_h represents the number of times $G(\mathbf{I}_{n_h}, n_h^*)$ is repeated.

3.1 Ex. 1 - Stratified SRS w/o Replacement

In this example, it is assumed $G(\mathbf{I}_{n_h}, n_h^*)$ directly selects $\mathbf{I}^*_{n_h^*.}$ with-replacement from $\mathbf{I}_{n_h}$. The object then is to choose n_h^* so the bootstrap variance estimator is unbiased. By comparing the bootstrap expectation of the bootstrap variance estimator with the traditional unbiased variance estimator, one can see that $n_h^* = (n_h - 1)/(1 - n_h / N_h)$ will do the job. See McCarthy and Snowden (1985). This is called the with-replacement bootstrap (BWR).

3.2 Ex. 2 - Rao, Hartley, Cochran Sampling

Rao, Hartley Cochran (Cochran, 1977) proposed a simple way of selecting a without replacement unequal probability sample. To do this, the PSUs on the frame are randomly placed into n_h groups. Each PSU, has a measure of size e_{ih}. Each random group has N_{gh} PSU's and $\sum_g N_{gh} = N_h$. Within each random group, one PSU is independently selected with probability $e_{ih} / \sum_{j=1}^{N_{gh}} e_{jh}$.

Sitter (1990) proposed a bootstrap procedure, similar to the mirror-match procedure to estimate this variance. In this method, $G(\mathbf{I}_{n_h}, n_h^*)$ uses a bootstrap frame developed from the sample, and bootstrap measures of size to select the independent bootstrap samples (e.g., if a selected PSU has weight w_i then w_i bootstrap-PSUs (i^*) are generated just like the original PSU, each with measure of size $1/w_i$, e_i^*). The bootstrap-PSUs are then randomly placed into n_h^* groups. Each group has N_{gh}^* bootstrap-PSUs. Within each bootstrap random group, one bootstrap-PSU is independently selected with probability $e_{ih}^* / \sum_{j=1}^{N_{gh}^*} e_{jh}^*$ In this procedure, where possible, n_h^* and N_{gh}^* are chosen so that the bootstrap variance estimator equals the unbiased estimator in Cochran (1977).

3.3 Ex. 3 - Randomized Systematic PPS Sample

When PSUs are placed in a specific order before sample selection, there is no unbiased variance estimator for systematic PPS samples. One can not apply the methodology described above to generate a bootstrap variance estimator. (A partial solution to the classical systematic sample problem is found in Kaufman, 1998). However, Kaufman (1999) modified the classical systematic PPS sampling procedure (Wolter, 1985) so that it resembled the Rao, Hartley Cochran sampling procedure. An argument similar to Cochran (1977) for this procedure can be made to produce an unbiased variance estimator, taking expectations across all possible orderings. This unbiased variance estimator can be used to determine n_h^* for an unbiased bootstrap variance estimator using a bootstrap-PSU (i^*) frame. Kaufman calls this type of sample, a randomized systematic PPS sample. One drawback with this procedure is that some PSUs can be selected multiple times.

A randomized systematic sample can be chosen in the following way: 1) Order the frame in the desired way for a regular systematic selection. 2) Partition the frame into n_h groups (partition groups), so each group's total measures of size are equal. 3) Consecutively pair the partition groups to form pseudo-strata. 4) Some PSUs may span multiple pseudo-strata. A PSU spanning multiple pseudo-strata must have its respective measures of size proportionally allocated into two pseudo-PSUs, so that the psuedo-PSUs are totally within the respective pseudo-strata. PSUs that span groups within a pseudo-stratum need not be split. 5) The PSUs and psuedo-PSUs within each pseudo-stratum are placed in a random order. Finally, a classical systematic PPS sample is selected within strata.

The randomized systematic sample, as with the classical systematic sample implicitly stratifies the

frame according to the original ordering in 1) above. The randomized systematic sample does not control as well as the classical systematic sample, but the control is still strong. For any contiguous group of frame PSUs, the classical systematic procedure will be within one PSU of the expected sample size for that group, while the randomized systematic sampling will be within two PSUs.

For the randomized systematic sample, Kaufman observes that the unbiased sample variance estimator can be written as a scaling factor times the BHR variance estimator. The scaling factor is always greater than zero, but can be either less than or greater than 1. Therefore, depending on the magnitude of the scaling factor, BHR can be either an extremely large under or over estimate. The simulation results (Kaufman, 1999) show this.

3.4 Summary

These examples demonstrate the versatility of the bootstrap with multiple designs. Replicate weights similar to BHR replicate weights can be generated, so standard software packages can be used. Additionally, by more correctly measuring the variance, the bootstrap can have lower mean square error (MSE) than the BHR, even when applying a simple FPC. See Kaufman (1999).

The technique used in these examples is to develop $G(\mathbf{I}_{n_h}, n_h^*)$ and to use a known unbiased variance estimator to determine an appropriate n_h^* so that $E^* v^*(\hat{T}_h^*) = v(\hat{T}_h)$.

For an arbitrary sample design, the bootstrap technique is: design $G(\mathbf{I}_{n_h}, \mathbf{A}_h^*)$, so first-order expectations are preserved for all $\mathbf{A}_h^*$, where $\mathbf{A}_h^*$ is an appropriate parameter space. $\mathbf{A}_{ho}^* \in \mathbf{A}_h^*$ is then determined so that $E^* v^*(\hat{T}_h^*) = v(\hat{T}_h)$ (i.e., second-order expectations are preserved). The choice of $G(\mathbf{I}_{n_h}, \mathbf{A}_h^*)$ and $\mathbf{A}_h^*$ can be flexible.

4.0 The NAEP Design

The State National Assessment of Educational Progress (state NEAP) measures the cognitive proficiency, in subjects such as reading and science, of our students in 4th and 8th grades. This is done through a two-stage sample design of schools ($\mathbf{I}_{n_h}^{(1)}$) and students within each selected school i ($\mathbf{I}_{m_i}^{(2)}$). Schools are chosen first with a systematic PPS selection process using students as the measure of size. Students are selected using a systematic equal probability procedure within each selected school.

Variances assume a super-population model (i.e., no FPC used). The current replication methods apply no FPC at the school level, and implicitly apply one at the student level (i.e., the super-population model is: a finite number of students go to each of an infinite number of schools).

Chromy (1998) stated "We view the structure of schools and their communities as fundamental contributors to the characteristics of students and the performance measures of the student population rather than viewing all states' students as arising from a single process and then simply being partitioned in some arbitrary manner among the schools in a state." He recommends a more appropriate model would to be assume that an infinite number of students go through a finite number of schools. This means an FPC should be applied at the school level, but not at the student level. This is the opposite of the current method.

Replication methodologies usually assume sampling is done with-replacement, so no FPCs are applied. This is true for the first stage of selection. At subsequent stages of selection, FPCs are implicitly applied. Chromy's recommendation could be implemented by explicitly applying a first stage FPC and subtracting out the implicitly applied second stage FPC. The disadvantage here is the periodic negative variance.

5.0 A Bootstrap Solution

An alternative methodology is to use the bootstrap. In the single stage examples above n_h^* is chosen, so that $G(\mathbf{I}_{n_h}, n_h^*)$ will produce an unbiased variance estimate. In a two-stage sample, where n_h first stage and m_i second stage units are selected, one can develop n_h^* from an appropriate $G_1(\mathbf{I}_{n_h}^{(1)}, n_h^*)$, and $m_{i^*}^*$ from an appropriate $G_2(\mathbf{I}_{m_{i^*}}^{(2)}, m_{i^*}^* \,|\, i^* \in \mathbf{I}_{n_h^*}^{(1)}, n_h^*)$, to produce an unbiased variance estimate (See Sitter (1992)).

For the NAEP problem, there is enough flexibility in determining n_h^* and $m_{i^*}^*$, so that the variance estimator reflects without replacement sampling at the first stage and with replacement sampling at the second stage.

5.1 The Two-Stage Sample Design

A proper simulation requires some modifications to the NAEP sample design. Specifically, NAEP uses a systematic PPS selection procedure to select the schools. Since the school frame is placed in a non-random order before selection, no unbiased variance estimator exists. To get around this problem, we will use the randomized systematic PPS selection procedure proposed by Kaufman.

Likewise, a second modification is that the second stage selection will be done with a SRS

without replacement selection, instead of a systematic equal probability selection.

Since a simulation study requires knowledge of the entire population and we don't know this for the entire student population, a different population is required. Instead, the NCES frame of school districts and the schools in these districts will be used as first and second stage units respectively.

See section 7.0 for the implications of these simplifying assumptions to the NAEP design.

Our sample design is: A stratified randomized systematic PPS sample of school districts will comprise the first stage sample. The measure of size will be number of schools in the district. The second stage sample will be a SRS w/o replacement sample of schools within each selected district. The desired variance reflects w/o replacement sampling at the first stage sample and reflects with-replacement sampling at the second stage.

5.2 The Estimate of Interest

Let $\hat{T} = \sum_{h\in H}\sum_{i=1}^{n_h} w_i \hat{Y}_i = \sum_{h\in H}\sum_{i=1}^{n_h} \hat{T}_i = \sum_{h\in H} \hat{T}_h$, where w_i is the sampling weight associated with the i^{th} school district (i.e., $1/p_i$, p_i being the selection probability for i); and $\hat{Y}_i$ is an unbiased estimator of the school total for i. $\hat{Y}_i = \sum_{j=1}^{m_i} M_i y_{ij} / m_i$, where M_i is the number of schools in district i, y_{ij} is the variable of interest for school j in district i, and m_i is number of schools selected in district i.

5.3 Estimating $v(\hat{T})$

Let $\overline{T}_h = \hat{T}_h / \sum_{i=1}^{n_h} w_i M_i = \hat{T}_h / X_h$.

Using a Taylor series approximation, $v(\overline{T}_h) \cong \sum_{i=1}^{n_h} w_i^2 M_i^2 v(\overline{y}_i) / X_h^2$, with $\overline{y}_i = \sum_{j=1}^{m_i} y_{ij} / m_i$.

From Cochran (1977) theorem 11.2, it follows that an unbiased estimator, within the Taylor Series approximation, for $V(\hat{T})$ is $v(\hat{T}) \cong \sum_{h\in H} X_h^2 \left(v_1(\overline{T}_h) + \sum_{i=1}^{n_h} p_i v_{2\,\text{wor}}(\hat{T}_i) / X_h^2 \right)$

or $v(\hat{T}) \cong \sum_{h\in H} \left(v_1(\hat{T}_h) + \sum_{i=1}^{n_h} \left(p_i v_{2\,wor}(\hat{T}_i) \right) \right)$ (a)

$v_1(\hat{T}_h)$ is an unbiased variance estimator of the first stage sample evaluated at $\hat{T}_h$. See Kaufman (1999) for $v_1(T_h)$. T_h includes all second-stage units.

$v_{2\,\text{wor}}(\hat{T}_i)$ is the unbiased estimate of the second stage variance of $\hat{T}_i$. $v_{2\,\text{wor}}(\hat{T}_i) = w_i^2 M_i^2 (1-f_{2i}) s_{2i}^2 / m_i$, where $f_{2i} = m_i / M_i$ and $s_{2i}^2 = \sum_{j=1}^{m_i} (y_{ij} - \overline{y}_i)^2 / (m_i - 1)$.

A bootstrap variance estimator is generated by:

1. Using $\mathbf{I}_{n_h^*}^b = G(\mathbf{I}_{n_h}, n_h^*)$ and n_h^* from Kaufman (1999), we have B sets of $\mathbf{I}_{n_h^*}^b$'s, as well as B sets of bootstrap-district (i^*) weights, $w_{i^*b}^*$, providing an unbiased $v_1(T_h)$. The b^{th} replicate weight, $w_{ib}^* = \sum_{i^*\in S_{ib}} w_{i^*}^*$, where S_{ib} is the set of i^* selected in the b^{th} replicate which were generated from i.

   As in Sitter's and Kaufman's procedures, a solution for n_h^* may not exist. This can occur in strata where n_h is small and $n_h^*=1$ is not small enough to sufficiently increase the bootstrap variance. A solution is to combine strata, indexed by C, and sort the combined stratum by original stratum first. This increased n_C in the combined stratum should now allow a solution for n_C^*.

2. Given n_h^* and $\mathbf{I}_{n_h^*}^b$ from step 1, define $\mathbf{I}_{m_{i^*}^*}^b = G_{i^*}(\mathbf{I}_{m_{i^*}}, m_{i^*}^* \mid i^* \in \mathbf{I}_{n_h^*}^b, n_h^*)$ as follows:

   For $i^* \in \mathbf{I}_{n_h^*}^b$, independently select $m_{i^*}^*$ schools with-replacement from the m_{i^*} originally sampled in district i which generated i^*. The conditional bootstrap replicate weight for the j^{th} school is $w_j^* = K_j^* M_{i^*} / m_{i^*}^*$, where K_j^* is the number of times the j^{th} school is selected.

3. For the b^{th} district bootstrap sample in step 1, $\mathbf{I}_{n_h^*}^b$, select $\mathbf{I}_{m_{i^*}^*}^b$ for each $i^* \in \mathbf{I}_{n_h^*}^b$, to get a set of conditional school bootstrap weights given the i^*'s, $w_{i^*jb}^* = K_{jb}^* M_{i^*} / m_{i^*}^*$ and a set of overall replicate weights, $w_{ijb}^* = \sum_{i^*\in S_{ib}} w_{i^*}^* w_{i^*jb}^*$.

4. Repeat step 3 B times for each district bootstrap sample, producing B sets of w_{ijb}^*'s.

5. Using the B sets of replicate weights in step 4, compute B estimates $\hat{T}_b^*$. The simple variance of these B estimates is the bootstrap variance, $v^*(\hat{T}^*)$, where $V^*(\hat{T}^*) = E^* v^*(\hat{T}^*)$.

If $m_{i^*}^* = (m_i - 1) w_{i^*}^* / (1 - f_{2i})$ is an integer then

$$E_1^* V_2^*(\hat{T}^*) = \sum_{h\in H} E_1^*\left(\sum_{i^*=1}^{n_h^*} w_{i^*}^{*2} M_i^2 (m_i - 1)/m_i s_{2i}^2 / m_{i^*}^* \right)$$

$$= \sum_{h\in H} \sum_{i=1}^{n_h^*} E_1^*\left((1-f_{2i}) w_{i^*}^* M_i^2 s_{2i}^2 / m_i\right)$$

$$= \sum_{h\in H} \sum_{i=1}^{n_h} \left((1-f_{2i}) w_i M_i^2 s_{2i}^2 / m_i\right)$$

$$= \sum_{h\in H} \sum_{i=1}^{n_h} p_i V_{2\,wor}(\hat{T}_i) \qquad \text{(b)}$$

Also, $V_1^* E_2^*(\hat{T}^*) = \sum_{h\in H} V_1^*\left(\sum_{i=1}^{n_h} \hat{T}_i \right) = \sum_{h\in H} v_1(\hat{T}_h)$, (c) from Kaufman (1999).

Since $V^*(\hat{T}^*) = V_1^* E_2^*(\hat{T}^*) + E_1^* V_2^*(\hat{T}^*)$, it now follows using (b) and (c) that $V^*(\hat{T}^*)$ with n_h^* and m_h^* defined above is an unbiased estimator for $V(\hat{T})$, within the Taylor Series approximation.

If $m_{i^*}^*$ is not an integer then it needs to be bracketed between the integer less than $m_{i^*}^*$ (m^L) and the integer greater than $m_{i^*}^*$ (m^U), where m^L is selected with probability $m^L(m^U - m_{i^*}^*)/m_{i^*}^*$ and m^U otherwise.

Our goal is to reflect the second stage variance as thought the sampling was done with replacement. In this situation, the variance estimate is:

$$\sum_{h\in H} \left(v_1(\hat{T}_h) + \sum_{i=1}^{n_h} \left(p_i v_{2wr}(\hat{T}_i) \right) \right), \qquad \text{(d)}$$

$v_{2wr} = w_i^2 M_i^2 s_{2i}'^2 / m_i$ and $s_{2i}'^2 = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 / m_i$.

Following steps 1-5 with $m_{i^*}^* = m_i w_{i^*}^*$, it follows that $V^*(\hat{T}^*)$ equals the result in (d).

6.0 Simulations

Using the sample design described above (i.e., the first stage strata are states with districts ordered before sample selection by urbanicity and measure of size), the districts are allocated to produce state estimates. For each of the 500 simulation samples, estimates are produced by region, state, and urbanicity. The region and state estimates are number of students, number of teachers and the student teacher ratio. The urbanicity estimates will additionally include number of schools, students per school and teachers per school. Reducing simulation time, only one region will be simulated.

Within the k^{th} simulation, two samples are selected. $\hat{T}_{1k}$ and $\hat{T}_{2k}$ are the respective estimates computed from these samples. $\hat{T}_{1k}$ is based on schools selected SRS without replacement, while $\hat{T}_{2k}$ is based on schools selected with replacement SRS. $\hat{T}_{2k}$ is used to produce an unbiased estimate of the desired variance, which is used to measure the performance of the bootstrap variance.

The relative error (RE) of the bootstrap standard error is used to measure efficiency. $RE = \left(\sqrt{v^*/v_2} - 1\right)\cdot 100$, where v^* is the average of the bootstrap variance estimates and v_2 is an unbiased estimate of the desired variance (i.e., $1/499\sum_k (\hat{T}_{2k} - \bar{T}_2)^2$).

The relative difference (RD) is used to measure how much larger the with-replacement standard error is compared to the without replacement standard error. $RD = \left(\sqrt{v_2/v_1} - 1\right)\cdot 100$, where v_1 is an unbiased estimate of the without replacement variance (i.e., $1/499\sum_k (\hat{T}_{1k} - \bar{T}_1)^2$).

7.0 Results and Conclusions

Table 1 provides the relative errors of the standard error for Urbanicity estimates. 12 errors are less than 5% in absolute value. 6 are greater than or equal to 5% in absolute value. Of these 6, 2 are greater than 10%. The relative difference measures how much the finite population standard error needs to be increased to be equal to the desired standard error. From table 1, all of the relative differences are significant, as one would expect from a state-based design. They range from 3.7% to 27.6%. So, the desired standard errors are much larger than the finite population standard errors, and the bootstrap makes the appropriate adjustments to get to the desired standard errors.

Table 2 provides the relative errors and differences for the West region. Since there is no variation in the school estimates, they are not provided in the table. For the same reason, the student and teacher average relative errors and differences are the same as the respective student and teacher numbers. So they are not repeated. Even with large relative differences from 10.7% to 14.2%, the bootstrap standard errors are close to the desired standard errors with relative errors from -8.6% to -1.9%.

Table 3 provides the relative errors for the State estimates. To conserve space, individual state averages are not provided for the 13 states simulated. Instead, quartiles are provided along with the minimum and maximum for each estimate. The median relative errors are all very close to zero. The inter-quartile ranges are also small with the largest being 5.0%. None of the relative errors are

larger than 10% in absolute value. The state relative differences, in table 4, are very large. They range from 1.5% to 289%. Even with large relative differences, the bootstrap accurately estimates the desired standard error.

These results clearly demonstrate that the proposed bootstrap variance estimator accurately reflects the desired standard errors. This would be difficult with standard BHR or jackknife variance estimators without some negative estimates.

Given the simulation design is not identical to the actual NAEP design, there are a number of ways for NAEP to implement these results. The NAEP model assumption is that an infinite number of students are going through a finite number schools. One way of viewing this, is that as a function of time, the student population, in any specific school, is continuously changing. It now seems reasonable to assume that the number of students in that school is random, as a function of time. Since school enrollment is used to order the frame before sample selection, the school ordering can be considered random. As such, the randomized systematic selection used in the simulation could be a good approximation for the NAEP first stage selection when estimating variances.

An alternative for the first stage is to assume the systematic sample variance can be approximated by some known finite population variance under a super-stratified frame assumption (i.e., two PSUs selected per stratum). Using the same arguments presented in this paper, the $v(T)$ appropriate for the first-stage finite population variance assumed, can be used as an approximation in the first part of (d) to develop the bootstrap variance.

For the second stage selection, the actual design uses a systematic equal probability selection, instead of a SRS without replacement selection. Since the student ordering is unlikely to be correlated with student assessment scores, using SRS in the bootstrap isn't likely to add any significant bias to the variance estimator. An alternative to this is to actually select students using without replacement SRS.

8.0 References

Chromy, J. (1998) "The Effects of Finite Sampling Corrections on State Assessment Sample Requirements", NAEP Validity Studies Report, American Institutes for Research, Palo Alto, CA.

Cochran, W. (1977) *Sampling Techniques*. New York: John Wiley and Sons.

Kaufman, S (1999). "Using the Bootstrap to Estimate the Variance from a Single Systematic PPS Sample," *Proceedings for the Section on Survey Methods, American Statistical Association*, pp. Alexandria, Va.: ASA.

Kaufman, S (1998). "A Bootstrap Variance Estimator for Systematic PPS Sampling," *Proceedings for the Section on Survey Methods, American Statistical Association*, pp. 769-774. Alexandria, Va.: American Statistical Association.

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

Sitter, R. R. (1990). "Comparing Three Bootstraps for Survey Data," *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality.*

Sitter, R. R. (1992). A Resampling Procedure for Complex Survey Data, *J. Amer. Statist Assoc.* , **87**, 755-765.

Wolter, K. M.(1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Table 1 – Relative Error (%) and Relative Distance (%) by Urbanicity and Estimate

Urbanicity	Estimate	Relative Error	Relative Dist.
Urban	Students	2.1	4.8
	Schools	9.4	8.2
	Teachers	2.2	9.0
	Ave. Stud	-1.9	3.7
	Ave Teach	-2.8	8.4
	Stud/Teach	-7.8	6.8
Suburban	Students	3.1	9.5
	Schools	15.0	13.4
	Teachers	2.8	11.7
	Ave. Stud	-0.7	7.0
	Ave Teach	-0.9	8.7
	Stud/Teach	-7.4	23.2
Rural	Students	1.6	15.7
	Schools	12.3	10.2
	Teachers	1.6	15.4
	Ave. Stud	-3.1	19.0
	Ave Teach	-3.4	19.4
	Stud/Teach	-6.8	27.6

Table 2 – Relative Error (%) and Relative Distance (%) by Region and Estimate

Region	Estimate	Relative Error	Relative Distance
West	Students	-3.0	10.7
	Teachers	-1.9	14.2
	Stud/Teach	-8.6	12.6

Table 3 – Quartiles of the Relative Error for States by Estimate

%	Min	1st Quar	Med	3rd Quar	Max
Students	-3.7	-0.6	0.0	2.7	6.3
Teachers	-4.6	-1.5	1.1	3.5	7.5
Stud/Teach	-7.7	-3.6	0.2	1.3	9.8

Table 4 – Quartiles of the Rel. Difference for States by Estimate

%	Min	1st Quar	Med	3rd Quar	Max
Students	2.2	10.6	32.9	42.0	241.
Teachers	1.5	12.5	36.4	42.3	289.
Stud/Teach	5.9	16.9	29.1	36.8	183.