A New Multiple-bootstrap-datasets Presentation Method for Confidentiality Protection

Yan Li[1,2] and Paul D. Williams[1]
National Center for Health Statistics, Hyattsville, MD, 20782[1]
JPSM, University of Maryland, College Park, MD, 20740[2]

Keywords: Confidentiality Protection, Bootstrap, Inverse sampling

1. Introduction[1]

Government agencies routinely release various public use microdata files (PUMF) at the respondent level such that outside researchers can perform various statistical analyses according to their own needs and verify results published by government agencies. There are mainly two equally critical but conflicting goals when releasing PUMF: maximizing the "openness" of its operation and preserving the confidentiality of the survey respondents. "Without adequate access to data, decision making is poorly based, and without adequate assurance of confidentiality, voluntary reporting is likely lessened." (Duncan and Lambert, 1986)

The basic requirement of creating PUMF is to provide adequate information on the sampling design and the variables so that various official estimates and their associated randomization-based variance estimates can be reproduced. Releasing survey weights and the cluster membership of a given respondent is needed to conduct standard statistical inference with complex survey data. However, releasing the key information may put the respondents at the risk of disclosure (Fienberg and Willenborg, 1998).

Commonly used methods to guard the confidentiality of survey respondents include cell suppression, data masking, and data swapping (Willenborg and de Wall, 1996; Duncan and Pearson, 1991; Cox, 1994, Fienberg, 1997). Nevertheless, these methods can distort relationships among variables in the data set. In addition, Rubin (1993) proposed to create multiple, synthetic data sets for public release. The agency selects units from sampling frame and imputes their data using models fits to the original survey data. However, a challenge for this approach is that a specific model must be assumed to reflect the structure of the data with a reasonable accuracy.

An alternative approach, the mean bootstrap method, was proposed by Yung (1997) to reduce the disclosure risk arising from the release of bootstrap weights with PUMF. The mean bootstrap method averages the bootstrap weights over C bootstrap samples to reduce the possibility of zero weights. In order to obtain B mean bootstrap weights, $B\times C$ bootstrap weights are produced. For example, in the General Social Survey, 5000 bootstrap weight variables were produced. These weights were then averaged in groups of size C=25 to obtain the B=200 mean bootstrap weights that accompany the original macrodata (Phillips, 2004). This procedure has successfully reduced the disclosure risk arising from the pattern of "zero" bootstrap weights. However, it has several disadvantages. First, compared to the standard bootstrap approach, $B\times C$ (instead of B) bootstrap weights are needed to obtain B mean bootstrap weights. Second, this method relies on poststratification adjustments to add noise to the mean bootstrap weights so that members from the same cluster do not share the same bootstrap weight. However, poststratification adjustments can be deduced back by smart users, and members from the same cluster still share the same weight, which might help an inquisitive user to recreate stratum and cluster membership from patterns of B mean bootstrap weights.

The main objective of this paper is to propose a new multiple-bootstrap-datasets presentation (MBDP) method. The MBDP method can effectively overcome the disadvantages of the mean bootstrap method: 1) only B bootstrap weights are needed in order to produce B bootstrap datasets, and 2) the MBDP method is capable of reducing the disclosure risk arising from the pattern of bootstrap weights, not relying on the poststratification adjustments.

As mentioned earlier, the original microdata may contain some information that can potentially reveal respondents' identity. For example, if the sample contains an outlier respondent (e.g., the richest person in the country), then the microdata can not only reveal this person's data but can potentially reveal information on some other respondents in the same cluster. Furthermore, combination of different variables for some respondents might be unique, which can be employed to identify the respondents. The circumstances discussed above are interesting research areas and need to be studied in the future. Readers interested in different research issues related to confidentiality and disclosure limitations are referred to Willenborg and de Waal (1996). This work focuses on addressing the disclosure risk arising from the pattern of bootstrap weights.

The outline of this paper is as follows. In Section 2, the proposed MBDP method is presented. In Section 3, this method is compared with another competitive method under two simple but realistic design settings analytically, and the corresponding simulation studies are conducted. Finally, in Section 4 this work is summarized.

[1] The opinions expressed in this paper are those of the authors and not necessarily those of the National Center for Health Statistics.

2. Proposed MBDP Method

The original microdata contain the information on the respondents, sample design and auxiliary variables. Some of the information cannot be released to protect the confidentiality of the respondents. In order to reduce the disclosure risk, the idea of this proposed MBDP method is to release multiple bootstrap datasets (avoiding zero replicate weights) instead of single original microdata to block all design information.

2.1 Difference between the MBDP Method and the Mean Bootstrap Method

The main difference between the MBDP method and the mean bootstrap method is in the manner of presentation in the PUMF. The mean bootstrap method basically provides one PUMF that contains *all B* bootstrap replicate weights and the variables. By investigating the pattern of *all B* bootstrap weights, the stratum and cluster membership might be identified by using the fact that records in the cluster have the same weight. In contrast, we suggest the presentation of multiple PUMFs. For each PUMF, only the clusters that are selected in the bootstrap method are included. For complex survey design, it is possible that members from different clusters share the same bootstrap weights. Therefore, members of the same cluster can not be identified by their bootstrap weights. Among different PUMFs, there is no one-to-one correspondence because different clusters are selected for different PUMFs. It is not possible to combine all *B* PUMFs and investigate the pattern of all *B* bootstrap weights to identify the members of the same cluster. Therefore, disclosure risk arising from the pattern of bootstrap weights is reduced. Furthermore, the bootstrap methods have already been well studied for the last two decades. “Various bootstrap procedures were developed to deal with various complex issues such as complex correlation structure induced by the survey design, weighting, imputation, small-area estimation, among others.” (Lahiri, 2003) Therefore, based on multiple bootstrap datasets valid inferences on variance estimation can be obtained, while the disclosure risk arising from the pattern of bootstrap replicate weights is reduced.

2.2 MBDP Estimates and Their Variances

The population parameter, θ, is estimated using *B* independent bootstrap datasets instead of the original microdata.

Define $\hat{\theta}^* = \theta(\hat{\bar{Z}}^*)$, where $\hat{\bar{Z}}^* = \hat{\bar{Z}}(S^*) = (\hat{\bar{Z}}_1^*, \hat{\bar{Z}}_2^* ..., \hat{\bar{Z}}_p^*)$ is estimated using the bootstrap sample *s**. Based on a bootstrap distribution of $\hat{\theta}^*$, the following two estimators of θ are considered: (i) $E_*[\theta(\hat{\bar{Z}}(S^*))] = \hat{\theta}^B$, the mean of the bootstrap distribution of $\hat{\theta}^* = \theta(\hat{\bar{Z}}(S^*))$, and (ii) $\theta[E_*(\hat{\bar{Z}}(S^*))] = \hat{\hat{\theta}}^B$, the original estimate with $\hat{\bar{Z}}$ replaced by $E_*(\hat{\bar{Z}}(S^*))$. Note if $\theta = \theta(\bar{Z})$ is a linear function of $\bar{Z} = (\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_p)$, the two estimators are same. If $\theta = \theta(\bar{Z})$ is nonlinear, we have the following two theorems:

Theorem 1: Assume regularity conditions to achieve good estimates $E_D(\hat{\bar{Z}}) = \bar{Z}$ and $\hat{V}_D(\hat{\bar{Z}}) - V_D(\hat{\bar{Z}}) = O_p(\frac{1}{n})$, where *n* is the sample size. Then, we have $\hat{\hat{\theta}}^B - \hat{\theta}^B = O_p(\frac{1}{n})$, $\hat{\theta}^B - \hat{\theta} = O_p(\frac{1}{n})$, and $\hat{\hat{\theta}}^B - \hat{\theta} = 0$ (See Appendix for derivation).

Monte Carlo simulation methods can be used to estimate the expectation and variance of the estimates. Suppose one decides to release *B* bootstrap samples. Then,

$$\hat{\hat{\theta}}^B = \theta[E_*(\hat{\bar{Z}}(S^*))] \approx \theta[\frac{1}{B}\sum_{b=1}^{B} \hat{\bar{Z}}(s_b^*)] = \tilde{\tilde{\theta}}^B, \text{ and}$$

$$\hat{\theta}^B = E_*[\theta(\hat{\bar{Z}}(S^*))] \approx \frac{1}{B}\sum_{b=1}^{B} \theta[\hat{\bar{Z}}(s_b^*)] = \tilde{\theta}^B,$$

where { $s_1^*, s_2^*, ..., s_B^*$ } are *B* bootstrap samples.

Theorem 2: Assume regularity conditions to achieve good estimates $E_D(\hat{\bar{Z}}) = \bar{Z}$ and $\hat{V}_D(\hat{\bar{Z}}) - V_D(\hat{\bar{Z}}) = O_p(\frac{1}{n})$, where *n* is the sample size. Then, we have

$E[\tilde{\tilde{\theta}}^B] = \theta(\bar{Z}) + O(\frac{1}{n})$, $E[\tilde{\theta}^B] = \theta(\bar{Z}) + O(\frac{1}{n})$,

$V[\tilde{\tilde{\theta}}^B] = (\frac{1}{B} + 1) \times V_D[\theta(\hat{\bar{Z}})]$, and

$V[\tilde{\theta}^B] = (\frac{1}{B} + 1) \times V_D[\theta(\hat{\bar{Z}})]$ (See Appendix for derivation).

Using a proper bootstrap procedure according to the corresponding sampling design, the variance $V_D[\theta(\hat{\bar{Z}})]$ can be estimated by

$$\text{var}_D[\theta(\hat{\bar{Z}})] = \frac{1}{B}\sum_{b=1}^{B} (\theta[\hat{\bar{Z}}(s_b^*)] - \tilde{\theta}^B)^2 . \qquad (2.1)$$

It can be seen from this theorem that in case sample size is sufficiently large, the bootstrap estimates are approximately unbiased; when *B* is sufficiently large the variances of bootstrap estimates approach the original microdata-based variance $V_D[\theta(\hat{\bar{Z}})]$.

3. Comparison between the Proposed MBDP Method and the Inverse Sampling Method

An inverse sampling algorithm was proposed by Hinkens, Oh, and Scheuren in 1997. This is an innovative technique. The idea is to generate a new sample from the original complex sample using a subsampling mechanism, and the generated new sample has a simpler data structure, like simple random sample (SRS), that is easier to analyze. This technique provides a useful tool to allow the public to access microdata because the design information would not be needed for the analysis.

However, the success of applications heavily depends on the complexity of survey design. Hinkins and Scheuren (2001) attempted to invert the sample for National Health Interview Survey (NHIS) to protect confidentiality. Nevertheless, inverting the NHIS data resulted in microdata only to the secondary sampling unit (SSU) level – data aggregated to clusters of households. For most researchers SSUs are simply not attractive as a unit of analysis. Therefore, this would not solve the problem of releasing useful microdata while protecting confidential data simultaneously for the survey data with complicated sampling design, like NHIS data. On the other hand, inverse sampling is still a valuable confidentiality protection method for surveys with simpler sampling designs.

Next, we will compare the MBDP method with the inverse sampling method analytically under two design settings: stratified SRS and one-stage clustering sampling. Simulation studies are also conducted to validate our analytical work.

3.1 Stratified simple random sampling (SSRS)

The population mean for the survey variable, *y*, under SSRS is $\bar{Y} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{Y}_h$, where N_h is the population size for the stratum *h*, $N = \sum_{h=1}^{L} N_h$, and $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$. The classical estimate for the population mean and its variance are

$$\bar{y}_{str} = \sum_{h=1}^{L} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}\right), \qquad (3.1.1)$$

$$V_D(\bar{y}_{str}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}, \qquad (3.1.2)$$

where n_h is the sample size for the stratum *h* and $S_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$ (Cochran, 1977).

According to Rao et al (2003), in order to obtain an inverse sample for SSRS, one needs to first generate a random number m_h for each stratum from the hypergeometric distribution with the mass function $f(m) = \prod_{h=1}^{L} \binom{N_h}{m_h} \Big/ \binom{N}{m}$, and then draw a SRS of size m_h from n_h sample units in stratum *h*, where $m = \min(n_h) = \sum m_h$, $E(m_h) = \frac{mN_h}{N}$,

$V(m_h) = m \frac{N_h}{N} (1 - \frac{N_h}{N}) \frac{N-m}{N-1}$, and

$Cov(m_h, m_{h'}) = -m \frac{N_h}{N} \frac{N_{h'}}{N} \frac{N-m}{N-1}$. It can be seen from this scheme that there are two steps involved in order to obtain the inverse sample for SSRS:

1. A random number m_h for each stratum is generated by hypergeometric distribution; and
2. m_h sample units are selected with SRSWOR from n_h originally sampled units in stratum *h*.

The estimator for the population mean based on the j^{th} inverse sample is $\bar{y}_j^* = \frac{1}{m} \sum_{h=1}^{L} \sum_{k=1}^{m_h} y_{hk}$, where $k \in s_h$ and s_h represents original sample of size n_h from stratum *h*, and m_h denotes the number of units selected from s_h in the inverse sample. Subsequently, the inverse sampling estimate of the population mean is $\bar{y}_{INV} = \frac{1}{g} \sum_{j=1}^{g} \bar{y}_j^*$ (Rao et. al., 2003), where *g* denotes the number of the inverse samples generated. It can be readily shown that $E(\bar{y}_{INV}) = \bar{Y}$ and by assuming *m/N* → *0*,

$$V_{INV}(\bar{y}_{INV}) = \frac{SSTO}{gmN} - \frac{1}{g} \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{S_h^2}{n_h} + \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{S_h^2}{n_h},$$

where $SSTO = \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y})^2$. Under proportionally stratified sampling, we have

$$V_{INV}(\bar{y}_{INV}) = \frac{1}{gN} \left(\frac{SSTO}{m} - \frac{SSW}{n}\right) + \frac{SSW}{Nn},$$

where $n = \sum n_h$ and $SSW = \sum_{h=1}^{L} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$ (Derivation is available upon request).

We recall that from Theorem 2 the variance for the proposed MBDP estimate is $V_{MBDP} = (\frac{1}{B} + 1) \times V_D(\bar{y}_{str})$. Next, we compare V_{MBDP} with V_{INV} . Under proportionally stratified sampling, we have

$$V_{MBDP} - V_{INV} = (1 + \frac{1}{B}) \frac{SSW}{Nn} - \frac{1}{gN} \left(\frac{SSTO}{m} - \frac{SSW}{n}\right) - \frac{SSW}{Nn}. \qquad (3.1.3)$$

For a fair comparison, let *Bn=gm*, where *n* denotes the bootstrap subsample size and *m* is the inverse sample size. Note that the total sample size in *g* inverse samples and *B* bootstrap samples would be equal. Consequently, we have

$$V_{MBDP} - V_{INV} = \frac{SSW}{gNn} - \frac{SSB}{gNm} = \frac{1}{gN} \left(\frac{SSW}{n} - \frac{SSB}{m}\right), \qquad (3.1.4)$$

where *SSB = SSTO – SSW*. Therefore, if $\frac{SSW}{n} < \frac{SSB}{m}$, then the MBDP method gives smaller variance estimates and is more efficient. Otherwise, the inverse sampling method is more efficient.

3.2 Simulation study 1

To validate our analytical results of Section 3.1, a limited simulation study is conducted. A finite population $\{ y_{hi}; h = 1,..., L; i = 1,...N_h \}$ is generated by the model

$y_{hi} = \mu_h + \varepsilon_{hi}$ for specified *H*, *N*$_h$, μ_h and σ_h^2, where μ_h is the stratum mean and $\varepsilon_{hi} \overset{iid}{\sim} N(0,\sigma_h^2)$.

We create *R*=500 independent SSRS samples with proportionally allocated sample size $\{ n_h = n\frac{N_h}{N} \}$ from each stratum. For the r^{th} SSRS sample, different estimates and their efficiencies are studied as follows.

1) Original micordata-based estimates: (3.1.1) and (3.1.2) with S_h^2 estimated by $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$ and $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$.

2) Inverse sampling estimates

i) Generate a random number m_h for each stratum from the hyper-geometric distribution with the mass function $f(m) = \prod_{h=1}^{L}\binom{N_h}{m_h}\Big/\binom{N}{m}$, where $m = \min(n_h)$;

ii) Draw a SRS of m_h without replacement from n_h sample units in the h^{th} stratum;

iii) Repeat step i) and ii) multiple times, say *g*. Conditioning on the original SSRS sample, *g* independent inverse samples are generated. The population mean can then be estimated along with its variance estimates (Rao et. al., 2003):

$$\bar{y}_{r,INV} = 1/g\sum_{i=1}^{g}\bar{y}_{(i)} \,, \qquad (3.2.1)$$

$$\mathrm{var}(\bar{y}_{r,INV}) = 1/g\sum_{i=1}^{g} v_{(i)} - 1/g\sum_{i=1}^{g}(\bar{y}_{(i)} - \bar{y}_{INV})^2 \,, \qquad (3.2.2)$$

where $\bar{y}_{(i)}$ and $v_{(i)}$ denote the estimate and the variance estimate produced from the i^{th} inverse sample, respectively.

3) MBDP estimates

i) Select a subsample of size n_h from original sample units for each stratum with replacement;

ii) Repeat step i) *B* times and *B=gm/n*;

iii) We have:

$$\bar{y}_{r,MBDP} = \frac{1}{B}\sum_{b=1}^{B}\bar{y}_{(b)} \,, \qquad (3.2.3)$$

$$\mathrm{var}(\bar{y}_{r,MBDP}) = (1+\frac{1}{B})\times\left[\frac{1}{B}\sum_{b=1}^{B}(\bar{y}_{(b)} - \bar{y}_{MBDP})^2\right], \qquad (3.2.4)$$

where $\bar{y}_{(b)}$ denotes the estimate produced from the b^{th} bootstrap sample.

The estimators $\bar{y}_{r,str}$, $\bar{y}_{r,INV}$, $\bar{y}_{r,MBDP}$, $\mathrm{var}(\bar{y}_{r,str})$, $\mathrm{var}(\bar{y}_{r,INV})$, and $\mathrm{var}(\bar{y}_{r,MBDP})$ are calculated for each SSRS sample. The means over *R* SSRS samples corresponding to each estimator are

$\bar{y}_{str} = \frac{1}{R}\sum_{r=1}^{R}\bar{y}_{r,str}$, $\bar{y}_{INV} = \frac{1}{R}\sum_{r=1}^{R}\bar{y}_{r,INV}$, $\bar{y}_{MBDP} = \frac{1}{R}\sum_{r=1}^{R}\bar{y}_{r,MBDP}$, $\mathrm{var}(\bar{y}_{str}) = \frac{1}{R}\sum_{r=1}^{R}\mathrm{var}(\bar{y}_{r,str})$, $\mathrm{var}(\bar{y}_{INV}) = \frac{1}{R}\sum_{r=1}^{R}\mathrm{var}(\bar{y}_{r,INV})$, and $\mathrm{var}(\bar{y}_{MBDP}) = \frac{1}{R}\sum_{r=1}^{R}\mathrm{var}(\bar{y}_{r,MBDP})$, respectively.

Tables 1 and 2 report the true population mean, three different point estimates and corresponding variance estimates produced by the original microdata-based method, the inverse sampling method, and the MBDP method over 500 simulation runs. The following parameters are adopted in the calculation: *N*=(1183761,552909,678371,436023), $n_h = N_h \times 0.005$, σ_h^2 ={1,1,1,1}, *g*=1000, and three different values of μ_h. *N* is chosen based on the statistics of income sample (Hinkins et al, 1997). It can be seen from Table 1 that the three different methods perform equally well in terms of point estimates. However, from Table 2 it can be observed that the MBDP estimates are more efficient than inverse sampling estimates when $\{\mu_h; h = 1,...,4\}$ gets more disperse. This result confirms our expectation. The more disperse of $\{\mu_h; h = 1,...,4\}$, the smaller of the value of $\frac{SSW}{n} - \frac{SSB}{m}$. Therefore, by (3.1.4) the more efficient of the MBDP estimates compared to the inverse sampling estimates. Table 2 also reports the proportion of negative variance estimates produced by the inverse sampling method over 500 SSRS samples. The maximum of this proportion can be as high as 44%, and the more disperse the value of μ_h, the higher the proportion of negative variance estimates produced by the inverse sampling method. Table 3 presents the results when we only consider the SSRS samples with positive variance estimates $\mathrm{var}(\bar{y}_{INV})$. Large difference (54.70 vs. 7.03) between $\mathrm{var}(\bar{y}_{INV})$ and $\mathrm{var}(\bar{y}_{MBDP})$ are found when μ_h = {10, 15, 20, 25} and the MBDP estimates are much more efficient than the inverse sampling estimates.

3.3 One-stage Cluster sampling with equal cluster size *M*

Let *A* and *a* represent the total number of clusters in the population and the number of sampled clusters, respectively. For one-stage cluster sampling the population mean is $\bar{Y} = \frac{1}{A}\sum_{i=1}^{A}\bar{Y}_i$, estimated by

$$\bar{y}_{cl} = \frac{1}{a}\sum_{i\in s_a}\bar{Y}_i \,, \qquad (3.3.1)$$

where $\bar{Y}_i$ denotes the i^{th} cluster mean and s_a represents the set of sampled clusters. Assuming finite population correction can be ignored, the variance

$$Var(\bar{y}_{cl}) = \frac{1}{a} S_A^2 , \qquad (3.3.2)$$

where $S_A^2 = \frac{1}{(A-1)} \sum_{i=1}^{A} (\bar{Y}_i - \bar{Y})^2$, can be estimated by

$$\text{var}(\bar{y}_{cl}) = \frac{1}{a} s_a^2 , \qquad (3.3.3)$$

where $s_a^2 = \frac{1}{(a-1)} \sum_{i \in s_a} (\bar{Y}_i - \bar{y})^2$ (Cochran, 1977).

According to Rao et. al. (2003), the approximate scheme to obtain an inverse sample for one-stage cluster sampling is to select one unit randomly from each cluster, and then treat the inverse sample as SRS. The estimator for the population mean using the j^{th} inverse sample is $\bar{y}_j^* = \frac{1}{a} \sum_{i \in s_a} y_{ik}$, where $k \in s_i$ and s_i denotes the set of units in the i^{th} sampled cluster. The expectation and variance for $\bar{y}_j^*$ are $E(\bar{y}_j^*) = \bar{Y}$ and $V(\bar{y}_j^*) = \frac{SSTO}{aAM}$, where $SSTO = \sum_{i=1}^{A} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2$. Define the inverse sampling estimator as $\bar{y}_{INV} = \frac{1}{g} \sum_{j=1}^{g} \bar{y}_j^*$ (Rao et. al., 2003). It can be derived that the variance for $\bar{y}_{INV}$ is $V_{INV}(\bar{y}_{INV}) = \frac{SSB}{aMA} + \frac{SSW}{gaMA}$, where $SSB = \sum_{i=1}^{A} M (\bar{Y}_i - \bar{Y})^2$, and $SSW = \sum_{i=1}^{A} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_i)^2$ (Derivation is available upon request).

The variance for the proposed MBDP estimate by Theorem 2 given in Section 2 is

$$V_{MBDP} = (\frac{1}{B} + 1) \times V_D[\theta(\hat{\bar{Z}})] = (\frac{1}{B} + 1) \times \frac{SSB}{aMA} . \qquad (3.3.4)$$

Next, we compare the variances by the inverse sampling method and the MBDP method,

$$V_{MBDP} - V_{INV} =$$

$$(1 + \frac{1}{B}) \frac{SSB}{aMA} - \frac{SSB}{aMA} - \frac{SSW}{gaMA} = \frac{1}{aMA} (\frac{SSB}{B} - \frac{SSW}{g}) . \qquad (3.3.5)$$

Let *Bn=gm*. Then, we have

$$V_{MBDP} - V_{INV} = \frac{1}{aMAB} (SSB - \frac{SSW}{M}) . \qquad (3.3.6)$$

This equation indicates that the variance difference depends on the data structure. If $SSB < \frac{SSW}{M}$, then $V_{MBDP} < V_{INV}$ and the MBDP method is more efficient. Otherwise, the inverse sampling method is more efficient.

3.4 Simulation study 2

Again a limited simulated study is conducted to validate our analytical results in Section 3.3. First we generate a finite population { $y_{ij}; i = 1, ..., A; j = 1, ...M$ } using the one-way random effects model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where $\alpha_i \overset{iid}{\sim} N(0, \tau^2)$ and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ for specified *A*, *M*, μ , τ^2 , and σ^2 .

We draw *a* clusters with SRSWOR. Given the sampled clusters, we study different estimates and their efficiencies:

1) Original microdata-based estimates (3.3.1) and (3.3.3).

2) Inverse sampling estimates

   i) Randomly draw one unit from each sampled clusters, and the inverse sample consists of *a* units;

   ii) Repeat step i) multiple times, say *g*. Conditioning on the original sample, *g* independent inverse samples are generated;

   iii) According to Rao *et. al.* (2003), the population mean can be estimated along with its variance estimates by (3.2.1) and (3.2.2).

3) MBDP estimates

   i) Resample *a* clusters from original sampled clusters with replacement;

   ii) Repeat step i) *B* times and *B* independent bootstrap samples are produced. Let *B = g/M*;

   iii) The population mean can be estimated along with its variance estimate by (3.2.3) and (3.2.4).

We draw *R=500* original samples independently from the finite population. For the r^{th} original sample, six estimators $\bar{y}_{r,cl}$, $\bar{y}_{r,INV}$, $\bar{y}_{r,MBDP}$, $\text{var}(\bar{y}_{r,cl})$, $\text{var}(\bar{y}_{r,INV})$, and $\text{var}(\bar{y}_{r,MBDP})$ are calculated. The means over *R* simulation runs corresponding to each estimator are

$$\bar{y}_{cl} = \frac{1}{R} \sum_{r=1}^{R} \bar{y}_{r,cl} , \quad \bar{y}_{INV} = \frac{1}{R} \sum_{r=1}^{R} \bar{y}_{r,INV} ,$$

$$\bar{y}_{MBDP} = \frac{1}{R} \sum_{r=1}^{R} \bar{y}_{r,MBDP} , \quad \text{var}(\bar{y}_{cl}) = \frac{1}{R} \sum_{r=1}^{R} \text{var}(\bar{y}_{r,cl}) ,$$

$$\text{var}(\bar{y}_{INV}) = \frac{1}{R} \sum_{r=1}^{R} \text{var}(\bar{y}_{r,INV}) \text{, and}$$

$$\text{var}(\bar{y}_{MBDP}) = \frac{1}{R} \sum_{r=1}^{R} \text{var}(\bar{y}_{r,MBDP}) \text{, respectively.}$$

Tables 4 and 5 give the relative biases for different point estimates and the corresponding variance estimates produced by the original microdata-based method, the inverse sampling method, and the MBDP method over *R* simulation runs. The following parameters are used: *A*=10,000, *M*=15, *a*=30, $\mu = 100$, *g*=1,500, *B*=100, σ^2 =15, and five different values of τ^2 . It can be observed from Table 4 that the three different methods perform similarly in terms of relative bias for different point estimates. The three variance estimates from Table 5 are also similar. Nevertheless, a slight pattern is showed that the inverse sampling estimates are more efficient than the proposed MBDP estimates as τ^2 increases. This result confirms our expectation. The value of $SSB - \frac{SSW}{M}$ increases with τ^2, and accordingly by (3.3.6) the more efficient are the inverse sampling estimates. Table 5 also reports the proportion of negative variance estimates produced by the inverse sampling method over 500 simulation runs. The results show that the maximum of the proportion of negative variance estimates by inverse sampling method is 5.6%, and the smaller the value of τ^2, the higher the proportion of negative variance estimates.

3. Summary and Future Research

This paper proposed a new method to reduce the disclosure risk arising from the pattern of bootstrap replicate weights while valid inferences on the variance estimation for complex surveys can be obtained. This new method was compared with the inverse sampling method under two simple but realistic survey designs. Both our analytical work and simulation work showed that whether the MBDP estimates have smaller variances depends very much on the data structure. One advantage of the MBDP method over the inverse sampling method is shown via simulation: variance estimates produced by the MBDP method are all positive, however, *negative* variance estimates can be produced by the inverse sampling method, especially under proportionally stratified sampling when the stratum means, μ_h's, are more disperse.

In the future research, we will compare the MBDP method and the inverse sampling method under more complex design settings, like two-stage cluster sampling, probability proportional to size sampling, etc. Furthermore, applications of the MBDP method to the real complex survey data would be useful. The inverse sampling method was applied to NHIS data (Hinkins and Scheuren, 2001) and this approach was not judged feasible for NHIS data. It would be interesting to see the performance of the MBDP method in the application to NHIS data. This paper also discussed a few potential situations when this new method might fail to protect the confidentiality of respondents in the Section 1. Further research is needed to understand the disclosure risk for the proposed method and to provide effective solutions in these circumstances.

References

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

Cox, L. H. (1994), "Matrix masking methods for disclosure limitation in microdata," *Survey methodology,* 20, 165-169.

Duncan, G. T. and Lambert, D. (1986), "Disclosure-limited data dissemination," *Journal of the American statistical association,* 81, 10-18.

Duncan, G. T. and Pearson, R. W. (1991), "Enhancing access to microdata while protecting confidentiality," *Statistical science*, 6, 219-239.

Fienberg, S. E. (1997), "Confidentiality and disclosure limitation methodology: challenges for national statistics and statistical research," *Background paper commissioned by committee on national statistics,* Available electronically at http://lib.stat.cmu.edu/www/cmu-stats.

Fienberg, S. E. and Willenborg, L. (1998), "Introduction to the special issue: disclosure limitation methods for protecting the confidentiality of statistical data," *Journal of official statistics,* 14, 337-345.

Hansen, M. H., Hurwitz, W. H., and Madow, W.G. (1953), *Sample survey methods and theory,* Volume I, New York: Wiley.

Hinkins, S., Oh, H. L., and Scheuren, F. (1997), "Inverse sampling design algorithms," *Survey Methodology*, 23, 11-21.

Hinkins, S., and Scheuren, F. (2001), "Increasing public accessibility to national health interview survey data (NHIS) using inverse sampling," Report prepared for NCHS under a professional services contract.

Lahiri, P. (2003), "On the impact of bootstrap in survey sampling and small-area estimation," *Statistical science,* 18, 199-210.

Leonard, M., Russell, B., and Scheuren, F. (1999), *Focal child public use file codebook,* Methodology Report No. 2, 1997 National survey of America's Families.

Phillips, O. (2004), "Using bootstrap weights with Wes Var and SUDDAN," *Statistics Cananda,* No. 12-002-XIE.

Rao, J. N. K., Scott, A. J., and Benhin, E. (2003), "Undoing complex survey data structures: some theory and applications of inverse sampling," *Survey methodology*, 29, 107-128.

Rubin, D. B. (1993), "Discussion: Statistical disclosure limitation," *Journal of official statistics*, 9, 462-468.

Shao, J. and Tu, D. (1995), *The jackknife and bootstrap*, Springer, New York.

Willenborg, L. and De Waal, T. (1996), *Statistical disclosure control in practice*, Spinger-Verlag.

Wolter, K. (1985), *Introduction to variance estimation,* New York: Springer-Verlag.

Yung, W. (1997), "Variance estimation for public use microdata files," *Proceedings of statistics Canada symposium*, 91-95.

Table 1. Means for different point estimates over 500 simulation runs.
(N=(1183761,552909,678371,436023); $n_h = N_h \times 0.005$; σ_h^2 ={1,1,1,1}; g=1000)

	$\overline{Y}$	$\overline{y}_{str}$	$\overline{y}_{INV}$	$\overline{y}_{MBDP}$
μ_h = {10, 10.5, 11, 11.5}	10.56	10.56	10.56	10.56
μ_h = {10, 11, 12, 13}	11.13	11.13	11.13	11.13
μ_h = {10, 15, 20, 25}	15.64	15.64	15.64	15.64

Table 2. Means for different variance estimates over 500 simulation runs.
(N=(1183761,552909,678371,436023); $n_h = N_h \times 0.005$; σ_h^2 ={1,1,1,1}; g=1000)

	var($\overline{y}_{str}$) (×10^{-5})	var($\overline{y}_{INV}$) (×10^{-5})	var($\overline{y}_{MBDP}$) (×10^{-5})	Prop (Neg#)
μ_h = {10, 10.5, 11, 11.5}	7.02	6.95	7.06	0.2%
μ_h = {10, 11, 12, 13}	7.03	7.27	6.98	3.8%
μ_h = {10, 15, 20, 25}	7.02	9.70	7.04	44%

Table 3. Means for different variance estimates over simulation runs with only positive var($\overline{y}_{inv}$) .
(N=(1183761,552909,678371,436023); n_h =0.005 N_h ; σ_h^2 ={1,1,1,1}; g=1000)

	var($\overline{y}_{str}$) (×10^{-5})	var($\overline{y}_{INV}$) (×10^{-5})	var($\overline{y}_{MBDP}$) (×10^{-5})	Prop (Neg#)
μ_h = {10, 10.5, 11, 11.5}	7.02	6.97	7.06	0.2%
μ_h = {10, 11, 12, 13}	7.03	7.65	6.97	3.8%
μ_h = {10, 15, 20, 25}	7.02	54.7	7.03	44%

Table 4. Relative biases for different point estimates over 500 simulation runs.
(A=10,000, M=15, a=30, $\mu = 100$ **,** g=1,500, B=100, and σ^2 =15)

	$\frac{\overline{y}_{cl} - \overline{Y}}{\overline{Y}}$ (×10^{-4})	$\frac{\overline{y}_{INV} - \overline{Y}}{\overline{Y}}$ (×10^{-4})	$\frac{\overline{y}_{MBDP} - \overline{Y}}{\overline{Y}}$ (×10^{-4})
τ^2 =0.01	-0.9	-1.0	-0.9
τ^2 =0.1	0.7	0.6	0.7
τ^2 =1	-1.4	-1.3	-1.4
τ^2 =5	2.3	2.4	2.4
τ^2 =10	-3.5	-3.6	-3.4

Table 5. Means for different variance estimates over 500 simulation runs.
(A=10,000, M=15, a=30, μ =100, g=1,500, B=100, and σ^2 =15)

	var($\overline{y}_{cl}$)	var($\overline{y}_{INV}$)	var($\overline{y}_{MBDP}$)	Prop (Neg#)
τ^2 =0.01	0.035	0.036	0.035	5.6%
τ^2 =0.1	0.037	0.037	0.037	4.2%
τ^2 =1	0.067	0.066	0.067	0%
τ^2 =5	0.199	0.197	0.201	0%
τ^2 =10	0.372	0.372	0.377	0%

Appendix

In this appendix, we prove Theorems 1 and 2.

Theorem 1 (Proof):

$$\hat{\theta}^B - \hat{\hat{\theta}}^B = E_*[\theta(\hat{\bar{Z}}(S^*))] - \theta[E_*\{\hat{\bar{Z}}(S^*)\}]$$

$$= E_*[\theta(\hat{\bar{Z}}(S^*))] - \theta[\hat{\bar{Z}}(S)]$$

$$\approx \theta[\hat{\bar{Z}}(S)] + \nabla[\theta(\hat{\bar{Z}}(S))]' E_*[\hat{\bar{Z}}(S^*) - \hat{\bar{Z}}(S)] +$$

$$\frac{1}{2} trH_\theta[\hat{\bar{Z}}(S)] V_*[\hat{\bar{Z}}(S^*)] - \theta[\hat{\bar{Z}}(S)]$$

$$= \frac{1}{2} trH_\theta[\hat{\bar{Z}}(S)] V_*[\hat{\bar{Z}}(S^*)]$$

$$= O_p(\frac{1}{n}).$$

where

$$\nabla[\theta(\hat{\bar{Z}}(S))] = \begin{bmatrix} \frac{\partial}{\partial \bar{Z}_1}\theta(\bar{Z}) \\ \vdots \\ \frac{\partial}{\partial \bar{Z}_p}\theta(\bar{Z}) \end{bmatrix}_{\bar{Z}=\hat{\bar{Z}}(s)} \quad \text{, and}$$

$$H_\theta[\hat{\bar{Z}}(S)] = \left(\left(\frac{\partial^2}{\partial \bar{Z}_i \partial \bar{Z}_j}\theta(\bar{Z})\right)\right)_{\bar{Z}=\hat{\bar{Z}}(s)} .$$

Since $\hat{\hat{\theta}}^B - \hat{\theta} = 0$, $Var(\hat{\hat{\theta}}^B) = Var(\hat{\theta})$. Therefore, inference does not change as far as bias and variance are concerned.

Theorem 2 (proof):

(i) For estimate $\tilde{\tilde{\theta}}^B$

Define $U_b^* = \hat{\bar{Z}}(s_b^*)$. Then U_1^*, U_2^*,…, U_B^* are *iid* with respect to *p(s*)*.

$E_*[U_b^*] = \hat{\bar{Z}}$.

$V_*[U_b^*] = V_*[\hat{\bar{Z}}(S^*)] = \hat{V}_{boot}$ for *b=1,2,…,B.*

$\tilde{\tilde{\theta}}^B = \theta[\bar{U}^*]$, where $\bar{U}^* = \frac{1}{B}\sum_{b=1}^{B} U_b^*$

$$E_*[\theta\{\bar{U}^*\}] = \theta\{\hat{\bar{Z}}\} + \nabla(\theta\{\hat{\bar{Z}}\})' E_*(\bar{U}^* - \hat{\bar{Z}}) +$$

$$\frac{1}{2} trH_{\bar{\theta}}(\hat{\bar{Z}}) E_*(\bar{U}^* - \hat{\bar{Z}})^2$$

$$= \theta\{\hat{\bar{Z}}\} + \nabla(\theta\{\hat{\bar{Z}}\})' E_*(\bar{U}^* - \hat{\bar{Z}}) +$$

$$\frac{1}{2} trH_{\bar{\theta}}(\hat{\bar{Z}}) \frac{E_*(U_b^* - \hat{\bar{Z}})^2}{B}$$

$$= \theta\{\hat{\bar{Z}}\} + \nabla(\theta\{\hat{\bar{Z}}\})' E_*(\bar{U}^* - \hat{\bar{Z}}) + \frac{1}{2} trH_{\bar{\theta}}(\hat{\bar{Z}}) \frac{\hat{V}_{boot}}{B}$$

$$= \theta\{\hat{\bar{Z}}\} + O_p(\frac{1}{nB})$$

$$V_*[\theta\{\bar{U}^*\}] = V_*[\theta\{\hat{\bar{Z}}\} + \nabla(\theta\{\hat{\bar{Z}}\})'(\bar{U}^* - \hat{\bar{Z}})]$$

$$\approx \frac{1}{B}\{\nabla(\theta\{\hat{\bar{Z}}\})' \hat{V}_{boot} \nabla(\theta\{\hat{\bar{Z}}\})\}$$

Therefore,

$$E[\tilde{\tilde{\theta}}^B\}] = E_D E_*[\theta\{\bar{U}^*\}] \approx E_D[\theta\{\hat{\bar{Z}}\}]$$

$$\approx \theta\{\bar{Z}\} + \nabla(\theta\{\bar{Z}\})' E_D(\hat{\bar{Z}} - \bar{Z}) + \frac{1}{2} trH_{\bar{\theta}}(\bar{Z}) E_D(\hat{\bar{Z}} - \bar{Z})^2$$

$$= \theta\{\bar{Z}\} + O_p(\frac{1}{n})$$

$$V[\tilde{\tilde{\theta}}^B] = E_D V_*[\theta\{\bar{U}^*\}] + V_D E_*[\theta\{\bar{U}^*\}]$$

$$\approx \frac{1}{B} E_D\{\nabla(\theta\{\hat{\bar{Z}}\})' \hat{V}_{boot} \nabla(\theta\{\hat{\bar{Z}}\})\} + V_D[\theta\{\hat{\bar{Z}}\}] \quad .$$

$$= \frac{1}{B} V_D[\theta\{\hat{\bar{Z}}\}] + V_D[\theta\{\hat{\bar{Z}}\}]$$

$$= (\frac{1}{B} + 1) \times V_D[\theta\{\hat{\bar{Z}}\}]$$

(ii) For estimate $\tilde{\theta}^B$

$$\tilde{\theta}^B = \frac{1}{B}\sum_{b=1}^{B} \theta[U_b^*].$$

$$E(\tilde{\theta}^B) = E_D E_*[\frac{1}{B}\sum_{b=1}^{B} \theta[U_b^*]] = E_D E_*[\theta[U_1^*]]$$

$$\approx E_D E_*[\theta(\hat{\bar{Z}}) + \{\nabla\theta(\hat{\bar{Z}})\}'(U_1^* - \hat{\bar{Z}}) +$$

$$\frac{1}{2} trH_\theta(\hat{\bar{Z}})(U_1^* - \hat{\bar{Z}})(U_1^* - \hat{\bar{Z}})'] \quad ,$$

$$= E_D[\theta(\hat{\bar{Z}})] + \frac{1}{2} trE_D[H_\theta(\hat{\bar{Z}})\hat{V}_{boot}]$$

$$= \theta(\bar{Z}) + O_p(\frac{1}{n})$$

where $E_D[\theta(\hat{\bar{Z}})] = \theta(\bar{Z}) + O_p(\frac{1}{n})$.

$$V(\tilde{\theta}^B) = E_D V_*[\frac{1}{B}\sum_{b=1}^{B} \theta[U_b^*]] + V_D E_*[\frac{1}{B}\sum_{b=1}^{B} \theta[U_b^*]]$$

$$= E_D\{\frac{V_*[\theta(U_1^*)]}{B}\} + V_D E_*[\theta(U_1^*)]$$

$$\approx E_D\{\frac{V_*[\theta(U_1^*)]}{B}\} + V_D[\theta(\hat{\bar{Z}}) + \frac{1}{2} trH_\theta(\hat{\bar{Z}})\hat{V}_{boot}],$$

$$\approx E_D\{\frac{V_*[\theta(U_1^*)]}{B}\} + V_D[\theta(\hat{\bar{Z}})]$$

$$\approx \frac{V_D[\theta(\hat{\bar{Z}})]}{B} + V_D[\theta(\hat{\bar{Z}})]$$

$$= (\frac{1}{B} + 1) \times V_D[\theta(\hat{\bar{Z}})]$$

where $V_D[\theta(\hat{\bar{Z}})]$ is estimated by

$$\hat{V}_{boot}[\theta(\hat{\bar{Z}})] = \frac{1}{B}\sum_{b=1}^{B}[\theta(\hat{\bar{Z}}_b^*) - \theta(\hat{\bar{Z}})]^2 .$$