Research Report

Large-Sample Properties of Minimum Discriminant Information Adjustment Estimates Under Complex Sampling Designs

ETS RR–20-13

Lili Yao
Shelby Haberman
Daniel F. McCaffrey
J. R. Lockwood

December 2020

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Large-Sample Properties of Minimum Discriminant Information Adjustment Estimates Under Complex Sampling Designs

Lili Yao,[1] Shelby Haberman,[2] Daniel F. McCaffrey,[1] & J. R. Lockwood[1]

1 Educational Testing Service, Princeton, NJ
2 Consultant, Jerusalem, Israel

Minimum discriminant information adjustment (MDIA), an approach to weighting samples to conform to known population information, provides a generalization of raking and poststratification. In the case of simple random sampling with replacement with uniform sampling weights, large-sample properties are available for MDIA estimates of population means and related functions of such means. This research report provides large-sample properties of MDIA estimates under complex sampling designs, such as stratified and two-stage sampling. Cases are considered for both sampling with replacement and sampling without replacement. MDIA is one case of calibration weighting, and this report includes results showing that sample calibration weights can exist only if MDIA weights exist, and MDIA weights can exist in situations where other calibration weights do not. Similarly, results in the report show that calibration weighting does properly generalize MDIA for populations. To illustrate results and explore the use of large-sample approximations in samples of moderate size, an application from Florida middle schools is examined for several sampling procedures to evaluate MDIA estimates for the prevalence of literacy coaches in those schools.

Keywords MDIA; large-sample properties; complex sampling

doi:10.1002/ets2.12297

Minimum discriminant information adjustment (MDIA) is a statistical method to provide sample weights that satisfy linear constraints (Csiszár, 1975; Haberman, 1984) typically specified in terms of sample means of auxiliary variables matching known targets, such as population means or the means of samples from populations of interest. MDIA has been applied in a variety of disciplines. For example, in survey sampling, raking (Deming & Stephan, 1940) is a special case of MDIA, and calibration estimation using exponential tilting (Kim, 2010) is another case of MDIA. A sizable literature (Haberman, 1974; Ireland et al., 1969; Ireland & Kullback, 1968; Kullback, 1959, 1971; Mosteller, 1968) has explored application of MDIA outside survey sampling. More general versions of MDIA have been employed in educational applications, such as linking without the use of anchor tests or equivalent groups (Haberman, 2015) and for adjustment of repeater data for biases resulting from when test takers repeat tests (Haberman et al., 2015; Haberman & Yao, 2015). Applications to causal inference also exist (Graham & de Xavier Pinto, 2012; Hainmueller, 2011). However, relatively few general large-sample results for MDIA are available in the literature, except in the case of simple random sampling with replacement (Haberman, 1984). To fill this gap, this research report concerns large-sample properties associated with estimation procedures for MDIA, especially in the case of complex sampling. The complex sampling designs considered here include simple random sampling without replacement, stratified sampling with replacement, stratified sampling without replacement, two-stage sampling with replacement, and rejective sampling. In particular, we provide the asymptotic variances and the estimated asymptotic variances of MDIA means for each complex sampling design. In addition, asymptotic results for sample MDIA means are generalized to sample calibration means (Deville et al., 1993; Deville & Särndal, 1992), including sample calibration means based on empirical likelihood (Hartley & Rao, 1968; Owen, 2001).

A basic description of MDIA and a review of known large-sample results appear in the section Background of Minimum Discriminant Information Adjustment. The section Large-Sample Results Under Complex Sampling Designs describes new results introduced in this report. In the section Calibration Weighting, MDIA is considered in the context of calibration weighting (Deville et al., 1993; Deville & Särndal, 1992). The section Application provides an application to estimate the rate of reading coaching in a statewide study of Florida middle schools by using the MDIA method. This application

Corresponding author: L. Yao, E-mail: lili.yao@gmail.com

permits an examination of the accuracy of large-sample approximations in the case of moderate sample sizes. Implications of results are considered in the Conclusion section.

Background of Minimum Discriminant Information Adjustment

MDIA weighting is applied to make inferences about outcomes in a target population from samples from a study population that may not be representative of the target population. The sample from the study population is weighted by the MDIA weights so that the distribution of the weighted sample more closely resembles the corresponding distribution for the target population.

To describe the general framework used, let the positive integer m be the dimension of the observations, and let the positive integer $q \geq m$ be the dimension of the underlying sample space R^q of q-dimensional vectors with associated expectation E and probability measure P. Let the positive integer n be the sample size. Let the random vectors $\mathbf{U}_i$, $1 \leq i \leq n$, be the m-dimensional sample vectors defined on R^q, and let the random vector $\mathbf{U}_*$ of dimension m be the population vector of interest. Throughout the paper, all random variables and random functions are defined on R^q. Let the $m \times n$ random matrix $\widetilde{\mathbf{U}}$ have columns $\mathbf{U}_i$ for $1 \leq i \leq n$. For each sampled unit i, $1 \leq i \leq n$, $\mathbf{U}_i$ includes the fundamental data observed for sampled unit i. Thus $\mathbf{U}_i$ includes both variables of direct interest and auxiliary variables. To avoid technical complications, let $\mathcal{B}$, the set of real Baire functions on R^m, be the smallest set of real functions on R^m that includes all continuous functions on R^m and has the property that Y is in $\mathcal{B}$ if Y_t, $t \geq 1$, is a sequence of functions in $\mathcal{B}$ that converges pointwise to Y. If Y is a real Baire function on R^m, then $Y_* = Y(\mathbf{U}_*)$ and $Y_i = Y(\mathbf{U}_i)$, $1 \leq i \leq n$, are random variables. If the expectation of Y_* is defined, then $E_*(Y) = E(Y_*)$.

For any positive integer d, a d-dimensional function $\mathbf{Y}$ on R^m is a d-dimensional Baire function on R^m if each element of $\mathbf{Y}$ is a real Baire function on R^m. If $\mathbf{Y}$ is a d-dimensional Baire function on R^m, then $\mathbf{Y}_* = \mathbf{Y}(\mathbf{U}_*)$ and $\mathbf{Y}(\mathbf{U}_i) = \mathbf{Y}_i$, $1 \leq i \leq n$, are random vectors. If the expectation of $\mathbf{Y}_*$ is defined, then $E_*(\mathbf{Y}) = E(\mathbf{Y}_*)$.

To simplify discussion, the assumption is made throughout the report that any real function on R^m ever mentioned is a real Baire function and any d-dimensional vector function on R^m that is mentioned is a d-dimensional vector Baire function.

The sample weight function v is a positive bounded real function on R^m, and sampled unit i, $1 \leq i \leq n$, has weight $v_i = v(\mathbf{U}_i)$. For any real function Y on R^m, let the sample mean $M(Y) = n^{-1} \sum_{i=1}^n Y_i$, and let the weighted sample mean $\overline{E}(Y)$ of Y be $M(vY)/M(v)$. In this report, the sample vectors $\mathbf{U}_i$, $1 \leq i \leq n$, and the population vector $\mathbf{U}_*$ are related by the unbiasedness requirement that the expectation $E(M(vY))$ of $M(vY)$ is $E_*(Y)$ whenever Y is a real function on R^m and $E_*(|Y|) < \infty$. The added requirement is that $M(vY)$ has finite variance whenever the variance $\sigma_*^2(Y) = E_*\left(\left[Y - E_*(Y)\right]^2\right)$ is finite. The case of $Y = 1_{Rm}$, the real function on R^m with constant value 1, then implies that $1 = E(M(v)) = n^{-1} \sum_{i=1}^n E(v_i)$; however, it is not assumed that $M(v)$ is always 1, so that the ratio estimate $\overline{E}(Y)$ need not have expectation $E_*(Y)$ even if $E_*(|Y|) < \infty$. Let $\mathbf{Y}$ be a k-dimensional function on R^m for some positive integer k, and let $|\mathbf{Y}|$ be the maximum absolute value of an element of $\mathbf{Y}$. Let the sample mean $M(\mathbf{Y}) = n^{-1} \sum_{i=1}^n \mathbf{Y}_i$, and let the weighted sample mean $\overline{E}(\mathbf{Y}) = [M(v)]^{-1} M(v\mathbf{Y})$, so that $E(M(v\mathbf{Y}))$ is $E_*(\mathbf{Y})$ whenever $E_*(|\mathbf{Y}|) < \infty$.

Adjustment is based on a d-dimensional function $\mathbf{Z}$ on R^m with elements Z_j, $1 \leq j \leq d$, for a positive integer d. The Z_j are auxiliary variables. For each sampled unit i from 1 to n, $\mathbf{Z}_i = \mathbf{Z}(\mathbf{U}_i)$ has elements Z_{ij}, $1 \leq j \leq d$. At the population level, $\mathbf{Z}_*$ has elements Z_{j*}, $1 \leq j \leq d$. It is convenient to assume that a vector $\mathbf{a}$ of dimension d with elements a_j, $1 \leq j \leq d$, exists such that $\mathbf{a}'\mathbf{Z} = \sum_{j=1}^d a_j Z_j = 1_{Rm}$. This assumption certainly holds if $Z_1 = 1_{Rm}$. Arguments are also simplified without essential loss of generality if $\mathbf{b}'\mathbf{Z}_* = 0$ with probability 1 for a d-dimensional vector $\mathbf{b}$ if, and only if, $\mathbf{b} = \mathbf{0}_d$, the d-dimensional vector with all elements 0.

A sample MDIA weight function (Haberman, 1984) is a positive real function $\widehat{w}$ on R^m with value $\widehat{w}_i = \widehat{w}(\mathbf{U}_i)$ for $1 \leq i \leq n$. Let the d-dimensional target vector $\mathbf{z}$ with elements z_j for $1 \leq j \leq d$ satisfy the constraint that $\mathbf{a}'\mathbf{z} = \sum_{j=1}^d a_j z_j = 1$. Subject to the constraint that the weighted average

$$\overline{E}(\widehat{w}\mathbf{Z}) = \mathbf{z}, \tag{1}$$

the sample discriminant information (Kullback & Leibler, 1951)

$$\overline{K}(\widehat{w}) = \overline{E}(\widehat{w}\log\widehat{w}) \geq 0 \tag{2}$$

is minimized. This sample discrimination compares the weight function $\widehat{w}$ to the uniform weight function 1_{Rm} on R^m. Let $\mathscr{X}$ be the set of $\mathbf{Z}_i$, $1 \le i \le n$. For any nonempty set $\mathscr{T}$ in R^m, the convex hull of $\mathscr{T}$ is the set of weighted averages $\sum_{j=1}^{k} c_j \mathbf{t}_j$, where k is a positive integer, c_j is a positive real number for $1 \le j \le k$, $\mathbf{t}_j$ is in $\mathscr{T}$ for $1 \le j \le k$, and $\sum_{j=1}^{k} c_j = 1$. A sample MDIA weight function exists if, and only if, $\mathbf{z}$ is in the interior of the convex hull of $\mathscr{X}$. Because $\mathbf{a}'\mathbf{Z}_i$ is 1 for $1 \le i \le n$, the weighted average $\overline{E}(\widehat{w})$ must be 1. If a sample MDIA weight function $\widehat{w}$ exists, then the sample MDIA weights $\widehat{w}_i = \widehat{w}(\mathbf{U}_i)$, $1 \le i \le n$, are uniquely determined, and $\log(\widehat{w}_i) = \widehat{\beta}'\mathbf{Z}_i$ for $1 \le i \le n$ for some $\widehat{\beta}$ in R^d. Conversely, if some $\widehat{\beta}$ in R^d exists such that $\widehat{w} = \exp\left(\widehat{\beta}'\mathbf{Z}\right)$ and Equation 1 holds, then $\widehat{w}$ is a sample MDIA weight function. If a sample MDIA weight exists and $\mathbf{b}'\mathbf{Z}_i = 0$, $1 \le i \le n$, only if the d-dimensional vector $\mathbf{b} = \mathbf{0}_d$, then $\widehat{w}$ and $\widehat{\beta}$ are uniquely defined. More generally, the condition that $\widehat{\beta} = \sum_{i=1}^{n} b_i \mathbf{Z}_i$ for some real b_i, $1 \le i \le n$, identifies $\widehat{w}$ and $\widehat{\beta}$ if a sample MDIA weight function exists. If no sample MDIA weight function exists, then the convention is adopted that $\widehat{w} = 1_{Rm}$ and $\widehat{\beta} = \mathbf{0}_d$.

The sample MDIA weight function $\widehat{w}$ is related to the positive population MDIA weight function w. This relationship relies on basic results concerning exponential families (Berk, 1972) and adjustment by minimum discrimination information (Csiszár, 1975). Let $w_* = \exp(\beta'\mathbf{Z})$ satisfy the condition that $E_*(w_*\mathbf{Z}) = \mathbf{z}$. Let f be a positive real function on R^m, and let $E_*(f\mathbf{Z}) = \mathbf{z}$. Then the nonnegative discriminant information $K_*(w) = E_*(w \log w) = \beta'\mathbf{z} \le K_*(f)$, with equality if, and only if, $f_* = w_*$ with probability 1. If $f_* = \exp(\mathbf{b}'\mathbf{Z}_*)$ with probability 1, then $\mathbf{b} = \beta$.

For computation and for the study of large-sample approximations, define the function $\overline{\ell}$ on R^d so that, for $\mathbf{b}$ in R^d,

$$\overline{\ell}(\mathbf{b}) = \mathbf{b}'\mathbf{z} - \overline{E}\left(\exp\left(\mathbf{b}'\mathbf{Z}\right)\right), \tag{3}$$

and let $\sup\left(\overline{\ell}\right)$ be the supremum of $\overline{\ell}(\mathbf{b})$ for $\mathbf{b}$ in R^d. Because $\overline{\ell}$ is concave and has gradient $\nabla\overline{\ell}(\mathbf{b}) = \mathbf{z} - \overline{E}\left(\exp\left(\mathbf{b}'\mathbf{Z}\right)\mathbf{Z}\right)$ at $\mathbf{b}$ in R^d (Berk, 1972), if a sample MDIA weight function $\widehat{w}$ exists, then $\nabla\overline{\ell}\left(\widehat{\beta}\right) = \mathbf{0}_d$ and $\overline{\ell}\left(\widehat{\beta}\right) = \sup\left(\overline{\ell}\right)$. Conversely, if $\widehat{\beta}$ is in R^d and $\overline{\ell}\left(\widehat{\beta}\right) = \sup\left(\overline{\ell}\right)$, then $\nabla\overline{\ell}\left(\widehat{\beta}\right) = \mathbf{0}_d$ and $\widehat{w} = \exp\left(\widehat{\beta}'\mathbf{Z}\right)$ is a sample MDIA weight function. The relationship of $\overline{\ell}$ to the sample MDIA weight functions implies that, if a sample MDIA weight function exists, then it can be computed by the Newton–Raphson algorithm used in maximum-likelihood estimation for log-linear models. In addition, iterative proportional fitting (raking) may also be employed if all elements of the vectors $\mathbf{Z}_i$, $1 \le i \le n$, are either 0 or 1 (Darroch & Ratcliff, 1972; Deming & Stephan, 1940; Haberman, 1974). A license for noncommercial use is available from the authors for software for the Newton–Raphson algorithm (Haberman, 2014).

Corresponding to $\overline{\ell}$ is the function ℓ on R^d defined for $\mathbf{b}$ in R^d by

$$\ell(\mathbf{b}) = \mathbf{b}'\mathbf{z} - E_*\left(\exp\left(\mathbf{b}'\mathbf{Z}\right)\right). \tag{4}$$

Let $\sup(\ell)$ be the supremum of $\ell(\mathbf{b})$ for $\mathbf{b}$ in R^d. Then ℓ is a concave function that is finite and strictly concave on the nonempty subset Ω of $\mathbf{b}$ in R^d such that $E_*(\exp(\mathbf{b}'\mathbf{Z})) < \infty$. To simplify arguments, assume that Ω has a nonempty interior Ω_0. Then ℓ is infinitely differentiable on Ω_0. For $\mathbf{b}$ in Ω_0, $E_*(\exp(\mathbf{b}'\mathbf{Z})|\mathbf{Z}|) < \infty$ and the gradient $\nabla\ell(\mathbf{b})$ of ℓ at $\mathbf{b}$ is $\mathbf{z} - E_*(\exp(\mathbf{b}'\mathbf{Z})\mathbf{Z})$. If $E_*(\exp(\mathbf{b}'\mathbf{Z})|\mathbf{Z}|) = \infty$ for any $\mathbf{b}$ in Ω but not in Ω_0, then w exists if, and only if, $\mathbf{z}$ is the interior of any convex set that contains $\mathbf{Z}_*$ with probability 1. In addition, $w = \exp(\beta'\mathbf{Z})$ if, and only if, β is in Ω_0 and $\ell(\beta) = \sup(\ell)$.

MDIA weight functions are associated with a variety of MDIA means and probabilities for both samples and populations. If Y is a real function on R^m, then the sample MDIA mean of Y is $\widehat{E}_w(Y) = [M(v)]^{-1} M\left(v\widehat{w}Y\right)$. If $E_*(w|Y|) < \infty$, then the population MDIA mean of Y is $E_w(Y) = E_*(wY)$. A similar definition holds for vectors of interest. If k is a positive integer and $\mathbf{Y}$ is a k-dimensional function on R^m, then the sample MDIA mean of $\mathbf{Y}$ is $\widehat{E}_w(\mathbf{Y}) = [M(v)]^{-1} M\left(v\widehat{w}\mathbf{Y}\right)$. By definition, $\widehat{E}_w(\mathbf{Z}) = \mathbf{z}$. In the population case, if $E_*(w|\mathbf{Y}|) < \infty$, then $E_w(\mathbf{Y}) = E_*(w\mathbf{Y})$. To study weighted fractions of sample units in a specified set B in R^m, define the indicator function χ_B to be the real function on R^m such that $\chi_B(\mathbf{u}) = 1$ for $\mathbf{u}$ in B and $\chi_B(\mathbf{u}) = 0$ for $\mathbf{u}$ in R^m but not in B. The set B is a Baire set if χ_B is a real Baire function. If B is a Baire set, then the weighted sample probability $\overline{P}(B)$ of B is $\overline{P}(B) = \overline{E}\left(\chi_B\right)$, the probability $P_*(B)$ that $\mathbf{U}_*$ is in B is $E_*(\chi_B)$, the sample MDIA probability of B is $\widehat{P}_w(B) = \widehat{E}_w\left(\chi_B\right)$, and the MDIA population probability is $P_w(B) = E_w(\chi_B)$. It is always assumed in this report that a set that is mentioned is a Baire set. Much more complex MDIA parameters may be defined (Haberman & Yao, 2015).

Example 1. In the most trivial use of MDIA, $d = 1$, $Z_1 = 1_{Rm}$, and $z_1 = 1$. Then $\widehat{w} = w = 1_{Rm}$. If Y is a real function, then $\widehat{E}_w(Y) = \overline{E}(Y)$. In addition, if Y_* has a finite expectation, then $E_w(Y) = E_*(Y)$.

Table 1 Notation Used

Symbol	Description
$\mathbf{U}_i$	Sample unit for observation i
$\mathbf{Z}$	Auxiliary variables used to construct MDIA weights
$\mathbf{z}$	Target vector for MDIA weight constraints
$\widehat{w}$	Sample MDIA weight function
w_*	Population MDIA weight function
Y	Real variable of interest
$\mathbf{Y}$	Vector of interest
β	Vector of coefficients in the MDIA constraints
H	Stratifying or grouping variable
$\widehat{E}_w(Y)$	Sample MDIA mean of Y
$\widehat{P}_w(B)$	Sample MDIA probability of B
$E_w(Y)$	Population MDIA mean of Y
$P_w(B)$	Population MDIA probability of B
$E_*(.)$	Expectation on the random vector $\mathbf{U}_*$

Note. MDIA = minimum discriminant information adjustment.

Example 2. For a more general but still elementary case, consider poststratification. Let the real function H on R^m be the stratifying or grouping function. This function has positive integer values from 1 to $J \geq 1$. Consider the case of $d = J$. Let $\mathbf{z}$ be the d-dimensional vector with elements $z_j > 0$, $1 \leq j \leq d$, with sum 1, and let $\mathbf{Z}$ be the d-dimensional vector with elements $Z_j = \delta_j(H)$, $1 \leq j \leq d$, where, for real x, δ_x is the real function on the real line such that, for y real, $\delta_x(y) = 1$ if $y = x$ and $\delta_x(y) = 0$ if $y \neq x$. Then $Z_{j^*} = 1$ is equivalent to $H_* = j$, and $E_*(Z_j)$ is the probability $p_{H^*}(j)$ that $H_* = j$. Assume that $p_{H^*}(j) > 0$ for $1 \leq j \leq J$. Then $w = z_j / p_{H^*}(j)$ if $H = j$ and $1 \leq j \leq J$. If Y is a real function on R^m and $E_*(|Y|) < \infty$, then $E_w(Y) = \sum_{j=1}^{J} z_j E_* \left(Y | H = j\right)$, where $E_*(Y | H = j) = E_*(Y\delta_j(H))/p_{H^*}(j)$ is the conditional expectation of Y_* given $H_* = j$.

If $\overline{p}_H\left(j\right) = \overline{E}\left(\delta_j(H)\right) > 0$ for $1 \leq j \leq J$, then $\widehat{w}_i = z_j / \overline{p}_H\left(j\right)$ if $1 \leq i \leq n$, $1 \leq j \leq J$, and $H_i = j$, and, for a real function Y on R^m, $\widehat{E}_w(Y) = \sum_{j=1}^{J} z_j \overline{E}\left(Y | H = j\right)$, where $\overline{E}\left(Y | H = j\right) = \overline{E}\left(Y\delta_j(H)\right) / \overline{p}_H\left(j\right)$.

For a summary of notation introduced in this section, see Table 1.

Large-Sample Results Under Complex Sampling Designs

Large-sample results under complex sampling designs (Cochran, 1977) are based on large-sample results for simple random sampling with replacement (Haberman, 1984) and large-sample results for contingency tables (Haberman, 1974, chapter 9). Consistency and asymptotic normality are examined. Specific examples include simple random sampling with or without replacement, stratified simple random sampling with or without replacement, and a simple form of two-stage sampling with replacement. In all cases, the sample size n becomes large. Cases vary in terms of the definition of the sampling vectors $\mathbf{U}_i$, $1 \leq i \leq n$, population vector $\mathbf{U}_*$, and sampling weight function v for different values of n. Arguments presented rely on results commonly found in books on mathematical statistics (e.g., Rao, 1973). In all cases, positive real numbers V_t and V_b exist such that $V_b \leq v(\mathbf{u}) \leq V_t$ for all sample sizes n. Two basic cases, sampling with replacement and sampling without replacement from a finite population, must be distinguished.

Sampling With Replacement

In sampling with replacement, the following theorem applies.

Theorem 1. *Let* $\mathbf{U}_*$ *be independent of the sample size n. Let M(vX) converge to* $E_*(X)$ *with probability 1 whenever X is a real function on* R^m *and* $E_*(|X|) < \infty$. *For a real function Y on* R^m *and a nonempty open subset B of* R^d *that contains* $\mathbf{0}_d$, *let* $E_w([|Y|+1]\exp(\mathbf{b}'\mathbf{Z}))$ *be finite for* $\mathbf{b}$ *in B. Then* $\widehat{E}_w(Y)$ *converges to* $E_w(Y)$ *with probability 1.*

Proof. The case of $X = 1_{Rm}$ implies that $M(v)$ converges to 1 with probability 1, so that $\overline{E}(X)$ converges to $E_*(X)$ with probability 1 whenever X is a real function on R^m and $E_*(|X|) < \infty$. Examination of the proof of strong consistency results in Haberman (1984) shows that substitution of $\overline{E}(X)$ for $M(X)$ in all arguments that involve a real function X on R^m such that $E_*(|X|) < \infty$ implies that arguments apply without material change. □.

The following cases are examples where the theorem applies.

Example 3. Let $\mathbf{U}_i$, $1 \le i \le n$, be mutually independent with the same distribution as the m-dimensional random vector $\mathbf{T}$ on R^m; let $\mathbf{U}_i$ remain the same for a given integer $i \ge 1$ for all sample sizes $n \ge i$. For any real function X on R^m, let $X_0 = X(\mathbf{T})$. If $|X_0|$ has a finite expectation, let $E_0(X) = E(X_0)$. Assume that positive constants V_b and V_t exist such that, if X is a real function on R^m such that $|X_*|$ has a finite expectation, then $V_b E_0(|X|) \le E_*(|X|) \le V_t E_0(|X|)$. Then a real positive function g exists such that $V_b \le g \le V_t$ and $E_0(gX) = E_*(X)$ if X is a real function on R^m and $E_*(|X|)$ is finite (Halmos, 1950). Let $v = g$. Then $E(M(vX)) = E_*(X)$ if X is a real function on R^m such that $|X_*|$ has a finite expectation. The variance $\sigma^2(M(vX)) = \sigma_0^2(vX)/n$ whenever the variance $\sigma_0^2(X)$ of $X(\mathbf{T})$ is finite. In terms of E_*, $\sigma_0^2(vX) = E_*\left(v\left[X - E_*(X)\right]^2\right)$. The strong law of large numbers implies that $M(vX)$ converges to $E_*(X)$ with probability 1. In simple random sampling with replacement (Haberman, 1984), $\mathbf{T} = \mathbf{U}_*$ and $v = g = 1_{Rm}$, so that $M(vX) = M(X)$ and $\sigma^2(M(vX)) = \sigma_*^2(X)/n$.

Example 4. To treat stratified sampling with replacement, as in Example 2, let H be a real function on R^m with positive integer values no greater than the positive integer J, and let $H_* = j$ with positive probability $p_{H^*}(j)$ for positive integers $j \le J$. Define $\mathbf{T}$ and g as in Example 3. Let $p_{H0}(j) = E_0(\delta_j(H)) > 0$, $1 \le j \le J$, denote the probability that $H_0 = H(\mathbf{T}) = j$. For any real function X on R^m and positive integer $j \le J$ such that $E(|X_0|\,\delta_j(H_0))$ is finite, let $E_0(X|H = j) = E_0(X\delta_j(X))/p_{H0}(j)$ denote the conditional expectation of X_0 given $H_0 = j$. Let $\mathbf{U}_i$, $i \ge 1$, be mutually independent, and let $j(i)$, $i \ge 1$, be a positive integer not greater than J. Let $\mathbf{U}_i$, $i \ge 1$, have the same distribution as the conditional distribution of $\mathbf{T}$ given $H_0 = j(i)$, so that, if X is a real function on R^m such that $E_0(|X|\,\delta_{j(i)}(H))$ is finite, then $E(X(\mathbf{U}_i)) = E_0(X|H_0 = j(i))$. For each positive integer $j \le J$, let n_j be the number of integers $i \le n$ such that $j(i) = j$. Assume that n_j/n converges to $p_H(j) > 0$ as the sample size n increases. Assume that n is sufficiently large that n_j is positive for $1 \le j \le J$. Let $v = g\sum_{j=1}^{J}\left(n/n_j\right)p_{H0}\left(j\right)\delta_j(H)$, so that $v(\mathbf{u}) = (n/n_j)p_{H0}(j)g(\mathbf{u})$ if $H(\mathbf{u}) = j$, $\mathbf{u}$ is in R^m, and $1 \le j \le J$. Then $E(M(vX)) = E_*(X)$ whenever X is a real function on R^m such that $E_*(|X|)$ is finite. The variance

$$\sigma^2(M(vX)) = \sum_{j=1}^{J}\left[p_{H*}\left(j\right)\right]^2 E_*\left(v\left[X - E_*\left(X|H = j\right)\right]^2|H = j\right)/n_j.$$

For positive integers $j \le J$ and X a real function on R^m, let $M(X|H = j)$ be the average of X_i for positive integers $i \le n$ such that $H_i = j$. Then

$$M(vX) = \sum_{j=1}^{J} p_{H0}\left(j\right) M\left(gX|H = j\right).$$

If $E_*(|X||H = j)$ is finite and $1 \le j \le J$, then the strong law of large numbers implies that $M(gX|H = j)$ converges to $E_0(gX|H = j)$ with probability 1 for $1 \le j \le J$. If the real function X on R^m satisfies the condition that $E_*(|X|)$ is finite, then $M(vX)$ converges with probability 1 to $E_0(gX) = E_*(X)$.

If $\mathbf{T} = \mathbf{U}_*$ and $g = 1_{Rm}$, then $E_*(v[X - E_*(X|H = j)]^2)$ is the conditional variance $\sigma_*^2\left(X|H = j\right)$ of X_* given $H_* = j$.

A variation on Theorem 1 is important in sampling without replacement but applies more generally. The proof is in the appendix.

Theorem 2. *Let $\mathbf{U}_*$ converge in distribution to $\mathbf{U}_\infty$ as the sample size n goes to ∞. Let h be a nonnegative continuous real function on R^m such that $E(h(\mathbf{U}_\infty))$ is finite. Let $M(vX)$ converge in probability to $E_\infty(X) = E(X_\infty)$, where $X_\infty = X(\mathbf{U}_\infty)$, if X is a real function on R^m that is continuous with probability 1 at $\mathbf{U}_\infty$ and that satisfies the condition that $|X| \le h$. Let $\mathbf{Z}$ be continuous with probability 1 at $\mathbf{U}_\infty$, and let Y be a real function on R^m such that Y is continuous with probability 1 at $\mathbf{U}_\infty$. Assume that $\mathbf{b}'\mathbf{Z}_\infty = 0$ with probability 1 for a $\mathbf{b}$ in R^d if, and only if, $\mathbf{b} = \mathbf{0}_d$. Let β_∞ in R^d and $W = \exp\left(\beta_\infty'\mathbf{Z}\right)$ satisfy $E_\infty(W\mathbf{Z}) = \mathbf{z}$. For some nonempty open subset B of R^d such that $\mathbf{0}_d$ is in B, let $W[|Y| + 1]\exp(\mathbf{b}'\mathbf{Z}) \le h$ for all $\mathbf{b}$ in B. Then $\widehat{E}_w(Y)$ converges to $E_W(Y) = E_\infty(WY)$ with probability 1, and $E_w(Y)$ converges to $E_W(Y)$.*

In the statement of Theorem 2, the condition that $M(vX)$ converges in probability to $E_\infty(X)$ if X is a real function on R^m such that $|X| \le h$ and X is continuous at $\mathbf{U}_\infty$ with probability 1 can often be verified in the following manner, as shown in the appendix.

Corollary 1. *Let $\mathbf{U}_*$ converge in distribution to $\mathbf{U}_\infty$ as the sample size n goes to ∞. Let h be a nonnegative continuous real function on R^m. Let the variance $\sigma^2(M(vX))$ of $M(vX)$ converge to 0 if X is a real bounded function on R^m. Let $\mathbf{Z}$ be continuous*

with probability 1 at $\mathbf{U}_\infty$*, and let Y be a real function on* R^m *such that Y is continuous with probability 1 at* $\mathbf{U}_\infty$*. Assume that* $\mathbf{b}_\infty{}' \mathbf{Z}_\infty = 0$ *with probability 1 for a* $\mathbf{b}$ *in* R^d *if, and only if,* $\mathbf{b} = \mathbf{0}_d$*. Let* β_∞ *in* R^d *and* $W = \exp\left(\beta'_\infty \mathbf{Z}\right)$ *satisfy* $E_\infty(W\mathbf{Z}) = \mathbf{z}$*. For some nonempty open subset B of* R^d *such that* $\mathbf{0}_d$ *is in B, let* $W[|Y|+1]\exp(\mathbf{b}'\mathbf{Z}) \le h$ *for all* $\mathbf{b}$ *in B. Then* $\widehat{E}_w(Y)$ *converges to* $E_W(Y) = E_\infty(WY)$ *with probability 1, and* $E_w(Y)$ *converges to* $E_W(Y)$*.*

Sampling Without Replacement

In sampling without replacement from a finite population, results are affected by the need for $\mathbf{U}_\star$ to vary with n. To see this issue, let $\mathscr{U}$ be a finite set with $N \ge n$ elements and assume that $N \ge 2$. Let $\mathbf{U}_\star$ only have values in $\mathscr{U}$. For X a real function on R^m, let $M_{\mathscr{U}}(X)$ be the average of $X(\mathbf{u})$ for $\mathbf{u}$ in $\mathscr{U}$. Let $\mathscr{U}_n$ be the set consisting of the $N!/(N-n)!$ matrices $\widetilde{\mathbf{u}}$ with n distinct columns $\mathbf{u}_i$ in $\mathscr{U}$, $1 \le i \le n$, and let $M_{\mathscr{U}n}(Y)$ be the average of $Y\left(\widetilde{\mathbf{u}}\right)$ for $\widetilde{\mathbf{u}}$ in $\mathscr{U}_n$, where Y is a real function on the set $R^{m\times n}$ of m by n real matrices. In sampling from $\mathscr{U}$ without replacement, $\widetilde{\mathbf{U}}$ is in $\mathscr{U}_n$, so that the $\mathbf{U}_i$ are distinct members of $\mathscr{U}$ for $1 \le i \le n$. This requirement on sampling without replacement from $\mathscr{U}$ can only be met if $n \le N$.

In this section, two elementary cases are considered.

Example 5. In simple random sampling without replacement, $\widetilde{\mathbf{U}}$ is uniformly distributed on $\mathscr{U}_n$, so that if Y is a real function on $R^{m\times n}$, then $E\left(Y\left(\widetilde{\mathbf{U}}\right)\right) = M_{\mathscr{U}n}(Y)$ and $\mathbf{U}_i$, $1 \le i \le n$, is uniformly distributed on $\mathbf{U}$. Let $\mathbf{T}$ be uniformly distributed on $\mathscr{U}$, so that $\mathbf{U}_i$ and $\mathbf{T}$ have the same distribution for $1 \le i \le n$ and $E_0(X) = E(X(\mathbf{T})) = M_{\mathscr{U}}(X)$ for any real function X on R^m. Let $\mathbf{T}$ converge in distribution to $\mathbf{T}_\infty$, and let $\mathbf{U}_\star$ converge in distribution to $\mathbf{U}_\infty$. If X is a real function on R^m and $X_{0\infty} = X(\mathbf{T}_\infty)$ satisfies the condition that $|X_{0\infty}|$ has a finite expectation, then let $E_{0\infty}(X) = E(X_{0\infty})$. As in Example 3, let positive $V_{b\infty}$ and $V_{t\infty}$ exist such that $V_{b\infty}E_{0\infty}(|X|) \le E_\infty(|X|) \le V_{t\infty}E_{0\infty}(|X|)$ if $|X_\infty|$ has a finite expectation. Then a positive real function g on R^m exists such that $V_{b\infty} \le g \le V_{t\infty}$ and $E_\infty(X) = E_{0\infty}(gX)$ if $|X_\infty|$ has a finite expectation. Assume that g is continuous at $\mathbf{U}_\infty$ with probability 1. Let $v = g/M_{\mathscr{U}}(g)$, so that $V_{b\infty}/V_{t\infty} \le v \le V_{t\infty}/V_{b\infty}$ and v converges to g. Let $E_*(X) = M_{\mathscr{U}}(vX) = E_0(vX)$ for any real function X on R^m, so that the probability that $X_\star = \mathbf{u}$ in $\mathscr{U}$ is $v(\mathbf{u})/N$. The expectation $E(M(vX)) = E_*(X)$ for all real functions X on R^m. The variance $\sigma^2(M(vX)) = (1-f)E_*(v[X-E_*(X)]^2)/n$, where $f = (n-1)/(N-1)$. The finite upper bound of v implies that $\sigma^2(M(vX))$ converges to 0 whenever X is bounded. Thus Corollary 1 applies.

Example 6. Simple stratified random sampling without replacement is closely related to stratified random sampling with replacement. Define $\mathbf{U}_\star$, $\mathbf{T}_\star$, $\mathbf{T}_\infty$, and $\mathbf{U}_\infty$ as in Example 5. Define H as in Example 2 to be a real function on R^m with positive integer values no greater than the positive integer J. For positive integer $i \le n$, let $j(i)$ be a positive integer no greater than J. For $1 \le j \le J$, let n_j be the number of integers $i \le n$ such that $j(i) = j$, and let $N_j \ge \max(2, n_j)$ be the number of $\mathbf{u}$ in $\mathscr{U}$ such that $H(\mathbf{u}) = j$. Then $p_{H0}(j) = N_j/N$ for $1 \le j \le J$. Let $\mathbf{j}$ be the n-dimensional vector with elements $j(i)$, $1 \le i \le n$. Let n_j/n approach $p_H(j) > 0$, $1 \le j \le J$, as the sample size n increases, and assume that n is large enough so that each n_j is positive. Let $\mathbf{H}_n$ be the n-dimensional function on $R^{m\times n}$ such that, for $\widetilde{\mathbf{u}}$ in $R^{m\times n}$ with columns $\mathbf{u}_i$, $1 \le i \le n$, $\mathbf{H}\left(\widetilde{\mathbf{u}}\right)$ has elements $H(\mathbf{u}_i)$ for $1 \le i \le n$. For any real function Y on $R^{m\times n}$, let $M_{\mathscr{U}n}\left(Y|\mathbf{H} = \mathbf{j}\right)$ be the average of $Y\left(\widetilde{\mathbf{u}}\right)$ for $\widetilde{\mathbf{u}}$ in $\mathscr{U}_n$ such that $\mathbf{H}\left(\widetilde{\mathbf{u}}\right) = \mathbf{j}$, and let $E\left(Y\left(\widetilde{\mathbf{U}}\right)\right) = M_{\mathscr{U}n}\left(Y|\mathbf{H} = \mathbf{j}\right)$. For a real function X on R^m and a positive integer $j \le J$, let $M_{\mathscr{U}}\left(X|H = j\right)$ be the average of $X(\mathbf{u})$ for $\mathbf{u}$ in $\mathscr{U}$ such that $H(\mathbf{u}) = j$. Then $E\left(X\left(\mathbf{U}_i\right)\right) = M_{\mathscr{U}}\left(X|H = j(i)\right)$ if X is a real function on R^m. Let H be continuous with probability 1 at $\mathbf{T}_\infty$, and let the probability $p_{H\infty}(j)$ that $H_\infty = j$ be positive for $1 \le j \le J$.

Let $v_r = g/M_{\mathscr{U}}(g)$ and

$$v = nv_r \sum_{j=1}^{J} p_{H0}(j)\,\delta_j(H)/n_j.$$

Then $E(M(vX)) = E_\star(X)$ whenever X is a real function on R^m such that $E_\star(|X|)$ is finite. To consider variances, let $f_j = (n_j - 1)/(N_j - 1)$ for $1 \le j \le J$. Then

$$\sigma^2(M(vX)) = \sum_{j=1}^{J}\left(1 - f_j\right)\left[p_{H*}(j)\right]^2 E_*\left(v\left[X - E_*\left(X|H = j\right)\right]^2\right)/n_j.$$

If X is bounded, then $\sigma^2(M(vX))$ approaches 0 as the sample size n approaches ∞. Thus Corollary 1 applies.

As in Example 4, $E_*\left(v\left[X-E_*\left(X|H=j\right)\right]^2|H=j\right)=\sigma_*^2\left(X|H=j\right)$ for $1\le j\le J$ if $\mathbf{T}=\mathbf{U}_*$ and $g=1_{Rm}$.

Normal Approximations

Normal approximations are available for both sampling with replacement and sampling without replacement, although arguments for sampling with replacement are somewhat simpler. Let $E_w(|\mathbf{Z}|^2)$ be finite, and let $\mathbf{b}'\,\mathbf{Z}_*=0$ only if $\mathbf{b}$ in R^d is $\mathbf{0}_d$. For a real function Y on R^m such that $E_w(|\,Y||\mathbf{Z}|\,)$ is finite, for $\mathbf{b}$ in R^d, let

$$S_w(\mathbf{b};Y)=E_w\left(\left[Y-\mathbf{b}'\mathbf{Z}\right]^2-Y^2\right) \tag{5}$$

for each d-dimensional vector $\mathbf{b}$. Let

$$\mathbf{c}_w(Y)=\left[E_w\left(\mathbf{Z}\mathbf{Z}'\right)\right]^{-1}\left[E_w(Y\mathbf{Z})\right]. \tag{6}$$

Then $S_w(\mathbf{b};\,Y)\ge S_w(\mathbf{c}_w(Y);Y)$ with equality if, and only if, $\mathbf{b}=\mathbf{c}_w(Y)$. Define the residual $r_w(Y)$ to be $Y-[\mathbf{c}_w(Y)]'\,\mathbf{Z}$. This residual satisfies $E_w(r_w(Y)\mathbf{Z})=\mathbf{0}_d$. Because $\mathbf{a}$ is assumed to exist such that $\mathbf{a}'\,\mathbf{Z}=1_{Rm}$, $E_w(r_w(Y))=0$. If, in addition, $E_w(|Y|^2)$ is finite, then the MDIA variance of Y is $\sigma_w^2(Y)=E_w\left(\left[Y-E_w(Y)\right]^2\right)$. The MDIA residual variance is then $\sigma_w^2\left(r_w(Y)\right)=E_w\left(\left[r_w(Y)\right]^2\right)$.

The following result applies to sampling with replacement.

Theorem 3. *Assume that the conditions of Theorem* 1 *hold. Let* $E_w(w|Y|^2)$ *and* $E_w(w\,|\mathbf{Z}|^2)$ *be finite. Let* τ^2 *be a nonnegative real function on the set of real functions X on* R^m *such that* $E_*(X^2)<\infty$. *Assume that* $n^{1/2}[M(vX)-E_*(X)]$ *converges in law to the normal distribution* $N(0,\tau^2(X))$ *with mean 0 and variance* $\tau^2(X)$ *if X is a real function on* R^m *such that* $E_*(X^2)$ *is finite. Let* $\sigma_a^2\left(\widehat{E}_w(Y)\right)=\sigma^2\left(M\left(vwr_w(Y)\right)\right)$. *Then* $n\sigma_a^2\left(\widehat{E}_w(Y)\right)$ *converges to* $\tau^2(wr_w(Y))$, *and* $n^{1/2}\left[\widehat{E}_w(Y)-E_w(Y)\right]$ *converges in law to* $N(0,\tau^2(wr_w(Y)))$.

Remark. The statement on convergence in law to $N(0,\tau^2(X))$ is equivalent to a statement on convergence in distribution to a real random variable with distribution $N(0,\tau^2(X))$. The asymptotic variance of $\widehat{E}_w(Y)$ is $\sigma_a^2\left(\widehat{E}_w(Y)\right)$, and the square root $\sigma_a\left(\widehat{E}_w(Y)\right)$ of $\sigma_a^2\left(\widehat{E}_w(Y)\right)$ is the asymptotic standard deviation of $\widehat{E}_w(Y)$.

Proof. As in the proof of Theorem 1, the arguments in Haberman (1984) apply almost without change once $\overline{E}(X)$ is used instead of $M(X)$ for real function X on R^m such that $E_*(X)$ is finite. Consider the case of $E_*(X^2)$ finite, so that $M(vX)$ converges to $E_*(X)$ with probability 1. Because $E_*(v)=1$, $M(v)$ converges to $E_*(v)=1$ with probability 1. Standard properties of ratio estimation imply that $n^{1/2}\left[\overline{E}(X)-E_*(X)\right]$ converges in law to $N(0,\tau^2(X-E_*(X)))$. Because $E_*\left(wr_w(Y)\right)=0$, $n^{1/2}\overline{E}\left(wr_w(Y)\right)$ converges in law to $N(0,\tau^2(wr_w(Y))$. In the proof in Haberman (1984), $\tau^2(wr_w(Y))$ replaces $E_*([wr_w(Y)]^2)$.

Several examples related to Theorem 1 are addressed in the appendix.

The following result is needed for sampling without replacement. As in the case of Theorem 2, the result applies more generally. The required arguments based on convergence in probability rather than convergence with probability 1 proceed as in the proof of Theorem 2.

Theorem 4. *Assume that the conditions of Theorem* 2 *hold. In addition, let* $n^{1/2}[M(vX)-E_*(X)]$ *converge in distribution to* $N(0,\tau^2(X))$ *if X also satisfies the inequality* $X^2\le h$, *and let* $(WY)^2$ *and* $(W|\,\mathbf{Z}|\,)^2$ *both be less than h. Let* $\mathbf{c}_W(Y)=[E_W(\mathbf{Z}\mathbf{Z}')]^{-1}E_W(Y\mathbf{Z})$ *and* $r_W(Y)=Y-[\mathbf{c}_W(Y)]'\,\mathbf{Z}$. *Then* $n\sigma_a^2\left(\widehat{E}_w(Y)\right)$ *converges to* $\tau^2(Wr_W(Y))$, *and* $n^{1/2}\left[\widehat{E}_w(Y)-E_w(Y)\right]$ *converges in law to* $N(0,\tau^2(Wr_W(Y)))$.

Example 7. In Example 5, let $n\,/\,N$ converge to $f_\infty<1$, and let g be continuous at $\mathbf{U}_\infty$ with probability 1. Then $M_{\mathcal{U}}\left(g\right)$ converges to 1. In Theorem 4, let X be continuous with probability 1 at $\mathbf{U}_*$, let $\tau^2(X)$ be positive, and let $|X\,|\le h$ and $X^2\le h$. Let $\tau^2(X)=(1-f_\infty)E_\infty(g[X-E_\infty(X)]^2)$. Because $X^2g/\sup(g)\le h$, it follows that $n\sigma^2(M(vX))$ converges to $\tau^2(X)$. For any real number $\epsilon>0$, let $t_\epsilon(y)$ be y for real y such that $y>n\sigma^2(M(vX))$ and 0 for other real y. Consider $q_\epsilon=M_{\mathcal{U}}\left(t_\epsilon\left(v^2\left[X-E_*(X)\right]^2\right)\right)/\left[n\sigma^2\left(M\left(vX\right)\right)\right]$. As n increases, $t_\epsilon(v^2[X-E_*(X)]^2)$ converges to 0 and is no greater

than $v^2[X - E_*(X)]^2$, and $E_0([v^2[X - E_*(X)]^2)$ converges to $\tau^2(X)$. It follows that q_ϵ converges to 0. Because ϵ is arbitrary, it follows that $n^{1/2}[M(vX) - E_*(X)]$ converges in law to $N(0, \tau^2(X))$ (Hájek, 1960, 1964). As in Example A2, $E_*(wr_w(Y))$ and $E_\infty(Wr_W(Y))$ are both 0, so that $\sigma_a^2\left(\widehat{E}_w(Y)\right) = \left(1-f\right)E_*\left(v\left[wr_w(Y)\right]^2\right)/n$ and $\tau^2(Wr_W(Y)) = E_W(gW[r_W(Y)]^2)$.

Example 8. In Example 6, let n_j/N_j converge to $f_{j\infty} < 1$ for positive integers $j \le J$. Define g as in Example 5. Let X be a real function on R^m such that X is continuous at $\mathscr{U}_\infty$ with probability 1, $\tau^2(X)$ is positive, $|X| \le h$, and $X^2 \le h$. For a real function Y on R^m such that $E_\infty(|Y|)$ is finite and for a positive integer $j \le J$, let $E_\infty(Y|H = j)$ be the conditional expectation of Y_∞ given $H_\infty = j$. As in Example 7, $n_j^{1/2}\left[M\left(gX|H=j\right) - E_0\left(gX|H=j\right)\right]$ converges in law to $N\left(0, \left(1-f_{j\infty}\right)\sigma_{0\infty}^2\left(gX|H=j\right)\right)$, where $\sigma_{0\infty}^2\left(gX|H=j\right)$ is the conditional variance of $g_{0\infty}X_{0\infty}$ given $H_{0\infty} = j$. The $M(gX|H = j)$ are mutually independent for $1 \le j \le J$. As in Example A3, $n^{1/2}[M(vX) - E_*(X)]$ converges in law to $N(0, \tau^2(X))$, where

$$\tau^2(X) = \sum_{j=1}^{J}\left(1-f_{j\infty}\right)\left[p_H(j)\right]^{-1}\left[p_{H\infty}(j)\right]^2 E_\infty\left(g\left[X - E_\infty\left(X|H=j\right)\right]^2\right) \mid H = j, \tag{7}$$

and $n\sigma^2(M(vX))$ converges to $\tau^2(X)$.

If $\delta_j(H)$ is a linear function of $\mathbf{Z}$ for $1 \le j \le J$, then some simplification occurs. The asymptotic variance $\sigma_a^2\left(\widehat{E}_w(Y)\right) = \sum_{j=1}^{J}\left(1-f_j\right)\left[p_{H*}(j)\right]^2 E_*\left(v\left[wr_w(Y)\right]^2|H=j\right)$ if Y is a real function on R^m.

Estimated Asymptotic Variances

Asymptotic variances can be estimated for both sampling with replacement and sampling without replacement. Results are obtained by essentially the same arguments used in Haberman (1984). Slight additions are needed to the conditions for asymptotic normality. The estimates for each case are based on a nonnegative-definite symmetric $n \times n$ matrix $\mathbf{Q}$ with elements Q_{ij}, $1 \le i \le n$, $1 \le j \le n$, and a corresponding estimate $s^2(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij}X_iX_j$ defined for real functions X on R^m. In this report, the matrix $\mathbf{Q}$ is selected so that, if X is a real function on R^m such that $E_*(X^2)$ is finite, then $E(s^2(X)) = \sigma^2(M(vX))$ for n sufficiently large. Because $\mathbf{Q}$ is nonnegative definite, $s(X)$, the square root of $s^2(X)$, has the seminorm properties that $s(cX) = |c|s(X)$ and $s(X+Y) \le s(X) + s(Y)$ for real functions X and Y on R^m and a real constant c. It follows that $|s(X) - s(Y)| \le s(X-Y)$. It is assumed that a real $\gamma > 0$ exists such that $ns^2(X) \le \gamma M(vX^2)$ for all real functions X on R^m and all sample sizes n such that the unbiasedness property holds. Let

$$\widehat{\mathbf{c}}(Y) = \left\{\widehat{E}_w\left(\mathbf{Z}\mathbf{Z}'\right)\right\}^{-1}\widehat{E}_w(Y\mathbf{Z}) \tag{8}$$

and $\widehat{r}_w(Y) = Y - \left[\widehat{\mathbf{c}}(Y)\right]'\mathbf{Z}$. The basic practice is to estimate $\sigma_a^2\left(\widehat{E}_w(Y)\right)$ by the estimated asymptotic variance $s_a^2\left(\widehat{E}_w(Y)\right) = s^2\left(\widehat{w}\widehat{r}_w(Y)\right)$. The square root $s_a\left(\widehat{E}_w(Y)\right)$ of $s_a^2\left(\widehat{E}_w(Y)\right)$ is the estimated asymptotic standard deviation of $\widehat{E}_w(Y)$. For an asymptotic confidence interval, let $0 < \alpha < 1$, and define the constant $z_{\alpha/2}$ so that a standard normal random variable exceeds $z_{\alpha/2}$ with probability $\alpha/2$. Then the asymptotic confidence bounds for $E_w(Y)$ of approximate level $1-\alpha$ are given by the following inequality:

$$\widehat{E}_w(Y) - z_{\alpha/2}s_a\left(\widehat{E}_w(Y)\right) \le E_w(Y) \le \widehat{E}_w(Y) + z_{\alpha/2}s_A\left(\widehat{E}_w(Y)\right). \tag{9}$$

The probability that Equation 9 holds approaches $1-\alpha$ if $\tau^2\left(wr_w(Y)\right) > 0$, $n^{1/2}\left[\widehat{E}_w(Y) - E_w(Y)\right]$ converges in law to $N(0, \tau^2(wr_w(Y)))$, and $s_a^2\left(\widehat{E}_w(Y)\right)$ converges in probability to $\tau^2(wr_w(Y))$. Justification of the practice is based on convergence results for $n\sigma_a^2\left(\widehat{E}_w(Y)\right)$ and the observation that $\left[s_a\left(\widehat{E}_w(Y)\right) - \sigma_a\left(\widehat{E}_w(Y)\right)\right]^2 \le \gamma M\left(v\left[\widehat{w}\widehat{r}_w(Y) - wr_w(Y)\right]^2\right)$.

In sampling with replacement, the added requirements are imposed in Theorem 3 that $\tau^2(wr_w(Y)) > 0$, $E_w([Y^2+1]w\exp(\mathbf{b}'\mathbf{Z})) < \infty$ for all d-dimensional vectors $\mathbf{b}$ in B, and $ns^2(X)$ converges with probability 1 to $\tau^2(X)$ whenever $E_*(X^2)$ is finite. It then follows that $ns_a^2\left(\widehat{E}_w(Y)\right)$ converges with probability 1 to $\tau^2(wr_w(Y))$. If $\tau^2(wr_w(Y))$ is positive, then $s_a^2\left(\widehat{E}_w(Y)\right)/\sigma_a^2\left(\widehat{E}_w(Y)\right)$ converges in probability to 1. Proofs of these assertions are essentially the same as in Haberman (1984).

In sampling without replacement, the added requirements are imposed in Theorem 3 that $\tau^2(Wr_W(Y)) > 0$, $W^2[Y^2+1]\exp(\mathbf{b}'\mathbf{Z})) \le h$ for all d-dimensional vectors $\mathbf{b}$ in B, and $ns^2(X)$ converges in probability to $\tau^2(X)$ whenever $X^2 \le h$ and X is continuous at $\mathbf{U}_\infty$ with probability 1.

Example 9. In Example 7, if $n > 1$, then $s^2(X) = (1-n/N)M([\nu X - M(\nu X)]^2)/(n-1) \le \sup(\nu)(1-n/N)[n/(n-1)]$ $M(\nu X^2)/n$, and $ns^2(X)$ converges with probability 1 to $\tau^2(X)$ if $X^2 \le h$; g and X are continuous at $\mathbf{U}_\infty$ with probability 1; and $s^2\left(\nu\widehat{w}\widehat{r}_w(Y)\right) = (1-n/N)\,M\left(\left[\nu\widehat{w}\widehat{r}_w(Y)\right]^2\right)/(n-1)$.

Example 10. In Example 8, if $n_j > 1$ for $1 \le j \le J$, then $s^2(X) = \sum_{j=1}^{J}\left(N_j/N\right)^2\left(1-n_j/N_j\right)M\left(\left[X - M\left(X|H=j\right)\right]^2 |H=j\right)/\left(n_j-1\right)$ converges with probability 1 to $\tau^2(X)$ if $X^2 \le h$ and X is continuous at $\mathbf{U}_\infty$ with probability 1. If $\delta_j(H)$ is a linear function of $\mathbf{Z}$ for $1 \le j \le J$, then the formula for $s^2\left(\nu\widehat{w}\widehat{r}_w(Y)\right)$ simplifies because $M\left(\nu\widehat{w}\widehat{r}_w(Y)|H=j\right) = 0$ for $1 \le j \le J$.

Calibration Weighting

Calibration weighting (Deville et al., 1993; Deville & Särndal, 1992) provides an approach to weighting that, in principle, generalizes MDIA; however, as shown in this section, the usefulness of the generalization is much more limited than it may appear at first to be. To describe the weighting procedure, let g be a nonnegative and twice-differentiable real function on the set $(0,\infty)$ of positive real numbers such that g has a nonnegative derivative g_1 and a positive second derivative g_2, $g(1) = g_1(1) = 0$, and $g_2(1) = 1$. A positive sample calibration weight function $\widehat{w}$ on R^m minimizes the weighted mean $\widehat{K}_c = \overline{E}\left(g\left(\widehat{w}\right)\right)$ subject to Equation 1. A positive population calibration weight function w on R^m minimizes $K_{c^*} = E_*(g(w))$ subject to the requirement that $w_*|\mathbf{Z}_*|$ has a finite expectation, $E_*(w\mathbf{Z}) = \mathbf{z}$, and $E_*(g(w))$ is finite. The function g, although not a true distance measure, does measure the discrepancy between a positive real number x and the number 1 to the extent that g is strictly convex; $g(x) = 0$ if, and only if, $x = 1$; $g(x)$ is increasing for $x > 1$; and $g(x)$ is decreasing for $x < 1$. In MDIA, $g(x) = x\log(x) - x + 1$, $g_1(x) = \log(x)$, and $g_2(x) = x^{-1}$ for $x > 0$. The equation $\overline{K}_c = \overline{K}$ holds because Equation 1 implies $\overline{E}\left(\widehat{w}\right) = 1$. In the same way, $K_{c^*}(w) = K_*(w)$. In the least squares approach (Deming & Stephan, 1940), $g(x) = (x-1)^2/2$, $g_1(x) = x - 1$, and $g_2(x) = 1$ for $x > 1$. In empirical likelihood (Hartley & Rao, 1968; Owen, 2001), $g(x) = -\log(x) + x - 1$, $g_1(x) = 1 - x^{-1}$, and $g_2(x) = x^{-2}$ for $x > 0$. Additional choices of the discrepancy measure g can be found in Deville and Särndal (1992).

The difficulty encountered with calibration weighting involves problems of generalizability, especially when treating infinite populations. Let A be the nonempty open interval such that y is in A if, and only if, $y = g_1(x)$ for some positive real x. Problems quickly arise if A is not the real line R. Indeed, all is straightforward for MDIA because $A = R$; however, in least squares, A is the set $(-1,\infty)$ of real numbers greater than -1, and for empirical likelihood, A is the set $(-\infty, 1)$ of real numbers less than 1. Let $\widehat{C}$ be the set of $\mathbf{b}$ in R^d such that $\mathbf{b}'\mathbf{Z}_i$ is in A for $1 \le i \le n$. If $\widehat{w}$ is a positive sample calibration weight function with value $\widehat{w}_i$ at $\mathbf{U}_i$, then $g_1\left(\widehat{w}_i\right) = \widehat{\beta}'\mathbf{Z}_i$, $1 \le i \le n$, for some $\widehat{\beta}$ in $\widehat{C}$. The $\widehat{w}_i$, $1 \le i \le n$, are uniquely defined, so that the sample calibration mean $\widehat{E}_w(Y) = \overline{E}\left(\widehat{w}Y\right) = [M(\nu)]^{-1}\sum_{i=1}^{n}\nu_i\widehat{w}_iY_i$ is well defined for any real function Y on R^m. If F_1 is the function on A defined by $g_1(F_1(y)) = y$ for y in A, then $\widehat{w}_i = F_1\left(\widehat{\beta}'\mathbf{Z}_i\right)$ for $1 \le i \le n$. Unless $A = R$, one cannot simply write $\widehat{w} = F_1\left(\widehat{\beta}'\mathbf{Z}\right)$, although the arbitrary convention may be used that $\widehat{w} = F_1\left(\widehat{\beta}'\mathbf{Z}\right)$ for $\widehat{\beta}'\mathbf{Z}$ in A and $\widehat{w} = 1$ otherwise. If $\mathbf{Z}_i'\mathbf{b} = 0$ for $1 \le i \le n$ for $\mathbf{b}$ in R^d only if $\mathbf{b} = \mathbf{0}_d$, then $\widehat{\beta}$ is uniquely defined. If no sample calibration weight function exists, then $\widehat{\beta} = \mathbf{0}_d$ and $\widehat{w} = 1_{Rd}$. A sample calibration weight function $\widehat{w}$ can exist only if a sample MDIA weight function exists. If a MDIA weight function exists, then the sample calibration weight function exists for the case of empirical likelihood (Chen & Qin, 1993); however, this result does not hold in general for least squares.

The population case is much more difficult. To apply the argument of Csiszár (1975), assume that positive real $c_0 < 1$ and positive real c_1, c_2, and c_3 exist such that $g(xy) \le c_1g(y) + c_2y + c_3$ if x and y are positive real numbers such that $|x-1| \le c_0$. This assumption holds for MDIA, least squares, and empirical likelihood. Assume that a population calibration weight w exists. It guarantees that $K_{c^*}(fw)$ is finite if f is a positive real function on R^m such that $1 - c_0 \le f \le 1 + c_0$. With this assumption, if $E_*(fw) = \mathbf{z}$, then $E_*(g_1(w)f) = 0$. It then follows that $g_1\left(w_*\right) = \beta\prime\mathbf{Z}_*$ with probability 1 for some d-dimensional vector β. Unfortunately, this condition cannot hold if A is not R, $E_*(\mathbf{Z})$ is not $\mathbf{z}$, and no d-dimensional vector $\mathbf{b}$ and positive real numbers ν and η exist such that $\mathbf{b}'\mathbf{Z}_* = \eta$ with probability 1 and $|\mathbf{b}'\mathbf{Z}| \le \nu$ with probability 1. Failure

of these conditions is hardly unusual. If A is not R, $E_*(Z)$ is not $\mathbf{z}$, $Z_1 = 1_{Rm}$, and the joint distribution of Z_j, $2 \le j \le d$, is a nonsingular multivariate normal distribution, then no population calibration estimate exists. As a consequence, calibration estimation is not an adequate generalization of MDIA.

Application

Data from a statewide literacy coaching program in Florida public schools (Lockwood et al., 2010) were used to illustrate the MDIA weighting method and to evaluate the accuracy of the large-sample variance approximations under four sampling designs: simple random sampling with replacement, simple random sampling without replacement, stratified simple random sampling with replacement, and stratified simple random sampling without replacement. For each design, cases considered involve both a standard case and a case with added weights. This example provides a framework for evaluating the accuracy of large-sample approximations. By itself, one example is inherently not a representative of all possible sampling designs, all possible data, and all possible sample sizes.

A brief description of literacy coaching and of its possible effect provides some context for the data analyzed and their origin. The main role of a literacy coach is to work with teachers to help them develop effective methods for teaching students reading skills. The study by Lockwood et al. (2010) used longitudinal school-level data to evaluate whether literacy coaches in Florida schools improved measures of student reading achievement. The study data included a census of schools and data on whether each school had a literacy coach. However, in other settings, the information of whether a school employs a literacy coach may be costly to obtain because it is not always coded consistently in administrative databases, and thus discussion with school personnel may be required to measure it accurately. Moreover, the presence of a coach may be related to more easily measured school attributes, such as the size and characteristics of the student population. Thus it is plausible that sampling of schools, along with weighting using auxiliary information, would be used in some circumstances when population characteristics regarding literacy coaching are of interest.

Our application considers use of sampling with MDIA weighting to estimate both the fraction of schools in the state with a literacy coach and the fraction of students in the state who are in a school with such a coach. The true value of each of these target parameters is known in this illustration from the available population-level data, so that the effects of random sampling and estimation can be evaluated with data from a real application.

For each sampling method explored, MDIA adjustment uses variables related to coaching status to weight the study samples to conform to known population attributes of the schools. In addition, in stratified sampling, dummy variables associated with the strata are also employed. Each sampling procedure was replicated 1,000 times. Each sampling procedure uses random samples of size $n = 200$ from the population of $N = 1{,}813$ schools. The sample size of 200 is selected to be large enough that large-sample approximations are plausible but small enough to be meaningful for a population of 1,813 schools. The following questions are of particular interest:

1. What is the value of MDIA adjustment for reducing estimation variance under different sampling designs?
2. How well do the asymptotic standard error estimators for MDIA weighted estimates described in the previous sections perform under different designs with samples of modest size?

Data Description

The application here employs data for students in Grade 5 in the 2005 – 2006 school year. After screening to remove records from four schools due to questionable accuracy of those data, the data contain information from 1,813 schools in Florida that serve Grade 5. For each school in the population, the vector $\mathbf{U}$ includes Y, the indicator function for presence of a school literacy coach; the number S of Grade 5 students who participated in end-of-year testing in reading; and the auxiliary vector $\mathbf{Z}$ used to construct MDIA weights. The variable S is a surrogate for the number of students in Grade 5. The vector $\mathbf{Z}$ includes the average score Z_1 for the school for the state accountability assessments in reading for Grade 5, the average score Z_2 for the school for the state accountability assessments in mathematics for Grade 5, the percentage Z_3 of students in the school who are Black, the percentage Z_4 of students in the school who are Hispanic, and the percentage Z_5 of students in the school who participate in the federal free and reduced-price lunch (FRL) program. In use of MDIA in this example, quadratic terms $Z_6 = Z_1^2$ and $Z_7 = Z_2^2$ are added. In addition, when simple random sampling is used with or without replacement, $d = 8$ and $Z_8 = 1_{Rm}$. When stratified simple random sampling is used with or without replacement,

Table 2 Summary Statistics for the School Population

	Even weights		Weighted by school size	
Variable	Mean	*SD*	Mean	*SD*
Coach[a]	0.64	0.48	0.63	0.48
Number of Grade 5 students	107.85	49.21	130.29	50.71
Average Grade 5 reading score	301.90	22.41	304.73	21.08
Average Grade 5 math score	326.04	20.62	329.08	19.38
School percentage Black students	26.72	27.42	22.91	24.74
School percentage Hispanic students	22.61	23.65	25.23	24.85
School percentage FRL students	54.41	25.12	51.24	25.02

Note. FRL = free and reduced-price lunch.
[a]1 = yes; 0 = no.

$d = 12$, $Z_8 = 1_{Rm}$, and $Z_j = \chi_{B(j)}$ for sets $B(j)$, $9 \le j \le 12$, defined for some real numbers $c(j)$, $8 \le j \le 12$, such that $c(8) = 0$, $c(j) < c(j+1)$, for $8 \le j \le 12$ and $B(j)$ is the set of $\mathbf{u}$ in R^m such that $c(j-1) < S(\mathbf{u}) \le c(j)$. The choice of $c(j)$ is discussed in the section Sampling Designs.

For summary statistics for these variables for the 1,813 schools, see Table 2. The fraction of schools with a literacy coach in the population is .64. Alternatively, the percentage of students in the Grade 5 population who are in schools with literacy coaches is 63%. Estimates $\overline{E}(Y)$ and $\widehat{E}_w(Y)$ are examined for $E_*(Y)$ in this section, where $E_*(Y)$ is either $M_{\mathcal{U}}(Y)$ or $M_{\mathcal{U}}(SY)/M_{\mathcal{U}}(S)$. In the first case, schools are equally weighted. In the second case, schools are weighted by school size. The elementary case considered here has $\mathbf{z} = E_*(\mathbf{Z})$.

Several of the variables in this application are related to coaching status, and thus their use in MDIA weighting may help to reduce error in estimates based on samples. For example, the correlation between the average reading score and coaching status is −.28, and the corresponding correlation for average math scores is −.26. The correlations are negative because schools were more likely to receive funding from the Florida coaching program to hire a literacy coach when their students had lower performance on the state accountability tests. The correlations between coaching status and the percentage of Black, Hispanic, and FRL students are .12, .25, and .39, respectively. Collectively, the average test scores and student demographic characteristics explain about 18% of the variance in coaching status in a linear regression, so that there is at least some opportunity for weighting adjustments with these variables to improve precision of sample estimates.

Sampling Designs

As previously noted, the simulation study evaluated four different sampling designs: simple random sampling of 200 schools with replacement, simple random sampling of 200 schools without replacement, stratified simple random sampling of 200 schools with replacement, and stratified simple random sampling of 200 schools without replacement. No further specifications are needed for the first two cases; however, in the latter two cases, strata must be defined. In the example, five strata are based on school size. Define $c(j)$, $9 \le j \le 12$, so that, for $8 \le j \le 12$, $M_{\mathcal{U}}\left(\chi_{B(j)}S\right)$ is close to $0.2M_{\mathrm{U}}(S)$. Thus approximately one-fifth of all students in all schools in the population are in schools in $B(j)$. One may let H have positive integer values no greater than 5; let $H = j$ if $c(j+8) < S \le c(j+9)$ and $1 \le j \le 4$, and let $H = 5$ if $S > c(13)$. Let $n_j = 40$ for $1 \le j \le 5$. This procedure leads to a weighting variable v relatively close to 1_{Rm} in the case of weighting based on school size S but leads to a variable v somewhat further from 1_{Rm} in the case of uniform weighting of schools.

For each sampling design and approach to weighting schools, the simulations yielded 1,000 observed estimates $\overline{E}(Y)$, $\widehat{E}_w(Y)$, $s_a\left(\overline{E}(Y)\right)$, and $s_a\left(\widehat{E}_w(Y)\right)$. Here $s_a\left(\overline{E}(Y)\right)$ is $s_a\left(\widehat{E}_w(Y)\right)$ for the trivial case in Example 2 with $d = 1$ and $Z_1 = 1_{Rm}$. The sample mean, sample standard error, and standard error of the sample mean of these estimates were then computed. (Not all are displayed.) In the case of the sample MDIA mean $\widehat{E}_w(Y)$, the resulting sample mean is denoted by $M_{mc}\left(\widehat{E}_w(Y)\right)$, the sample standard deviation is $s_{mc}\left(\widehat{E}_w(Y)\right)$, and the standard error of the sample standard deviation is $s_{mce}\left(\widehat{E}_w(Y)\right) = s_{mc}\left(\widehat{E}_w(Y)\right)/(1000)^{1/2}$. Similar notation is employed for $\overline{E}(Y)$, $s_a\left(\overline{E}(Y)\right)$, and $s_a\left(\widehat{E}_w(Y)\right)$.

Table 3 Sample Means of Estimated Fraction of Schools With Coach

Sampling design	Replacement	$E_*(Y)$	$M_{mc}\left(\overline{E}(Y)\right)$	$M_{mc}\left(\widehat{E}_w(Y)\right)$
Simple random	No	.6404	.6408	.6409
Simple random	Yes	.6404	.6406	.6405
Stratified	No	.6404	.6417	.6423
Stratified	Yes	.6404	.6411	.6414

Table 4 Sample Means of Estimated Fraction of Students in Schools With Coach

Sampling design	Replacement	$E_*(Y)$	$M_{mc}\left(\overline{E}(Y)\right)$	$M_{mc}\left(\widehat{E}_w(Y)\right)$
Simple random	No	.6251	.6258	.6234
Simple random	Yes	.6251	.6253	.6252
Stratified	No	.6251	.6262	.6255
Stratified	Yes	.6251	.6268	.6259

Table 5 Standard Error Estimates for Sample Means for Fraction of Schools With Coach

Sampling	Replacement	$\sigma_a\left(\overline{E}(Y)\right)$	$M_{mc}\left(s_a\left(\overline{E}(Y)\right)\right)$	$s_{mce}\left(s_a\left(\overline{E}(Y)\right)\right)$	$s_{mc}\left(\overline{E}(Y)\right)$
Simple random	No	.0320	.0320	.0006	.0323
Simple random	Yes	.0339	.0339	.0007	.0340
Stratified	No	.0347	.0346	.0013	.0334
Stratified	Yes	.0364	.0364	.0014	.0367

Results

Table 3 summarizes the means of the estimates of the fraction of schools with a coach, and Table 4 provides the analogous information for the estimates of the fraction of students in schools with a coach. In Table 3, $E_{\mathcal{U}}(Y) = M_{\mathcal{U}}(Y)$ and $\overline{E}(Y) = M(Y)$. In Table 4, $E_{\mathcal{U}}(Y) = M_{\mathcal{U}}(SY)/M_{\mathcal{U}}(S)$ and $\overline{E}(Y) = M(SY)/M(S)$. Recall the definition of $M_{\mathcal{U}}$ in the section Sampling Without Replacement, the definition of M is in the section Background of Minimum Discriminant Information Adjustment, and the definition of the school size S in the section Data Description. The differences between population values and sample means are quite small, so that bias appears to be minimal.

Tables 5–8 summarize the results for estimation of standard errors. The tables permit comparisons of $\sigma_{a\mathrm{U}}\left(\overline{E}(Y)\right)$, $M_{mc}\left(s_a\left(\overline{E}(Y)\right)\right)$, $s_{mc}\left(\overline{E}(Y)\right)$, $s_{mce}\left(\overline{E}(Y)\right)$, $\sigma_{a\mathrm{U}}\left(\widehat{E}_w(Y)\right)$, $M_{mc}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$, $s_{mce}\left(\widehat{E}_w(Y)\right)$, and $s_{mc}\left(\widehat{E}_w(Y)\right)$ for the different sampling procedures and for uniform weighting of schools versus weighting of schools in proportion to the number S of students, where $\sigma_{a\mathrm{U}}\left(\overline{E}(Y)\right)$, $s_a\left(\overline{E}(Y)\right)$, $\sigma_{a\mathrm{U}}\left(\widehat{E}_w(Y)\right)$, and $s_a\left(\widehat{E}_w(Y)\right)$ are the asymptotic variances for $\overline{E}(Y)$ and $\widehat{E}_w(Y)$ and the estimated asymptotic variances, while $s_{mce}\left(\overline{E}(Y)\right)$ and $s_{mce}\left(\widehat{E}_w(Y)\right)$ are the standard errors of the estimated asymptotic variance. Two issues are to be considered in these tables: One is the accuracy of approximations for asymptotic standard errors; the other is the accuracy of estimates of the mean $E_*(Y)$. Accuracy of approximations for asymptotic standard errors involves comparison of $\sigma_a\left(\overline{E}(Y)\right)$, $M_{mc}\left(s_a\left(\overline{E}(Y)\right)\right)$, and $M_{mc}\left(\overline{E}(Y)\right)$ in the case of the standard estimate $\overline{E}(Y)$ and comparison of $\sigma_a\left(\widehat{E}_w(Y)\right)$, $M_{mc}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$, and $M_{mc}\left(\widehat{E}_w(Y)\right)$ in the case of the sample MDIA mean. Reasonable expectations for accuracy are informed by the estimates $s_{mce}\left(\overline{E}(Y)\right)$ and $s_{mce}\left(\widehat{E}_w(Y)\right)$. Accuracy of estimates involves comparison of $\sigma_a\left(\overline{E}(Y)\right)$ and $\sigma_a\left(\widehat{E}_w(Y)\right)$ for different sampling procedures.

In terms of accuracy of approximations, results are relatively satisfactory. The largest contrast observed involved the sample MDIA mean for stratified simple random sampling with replacement for estimation of $M_{\mathcal{U}}(Y)$ and for simple random sampling without replacement for estimation of $M_{\mathcal{U}}(SY)/M_{\mathcal{U}}(S)$. In each case, the mean of the replicates for

Table 6 Standard Error Estimates for Sample Minimum Discriminant Information Adjustment Means for Fraction of Schools With Coach

Sampling	Replacement	$\sigma_a\left(\widehat{E}_w(Y)\right)$	$M_{mc}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$	$s_{mce}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$	$s_{mc}\left(\widehat{E}_w(Y)\right)$
Simple random	No	.0284	.0287	.0010	.0297
Simple random	Yes	.0307	.0304	.0012	.0312
Stratified	No	.0318	.0307	.0018	.0317
Stratified	Yes	.0334	.0323	.0020	.0344

Table 7 Standard Error Estimates for Fraction of Students in Schools With Coach

Sampling	Replacement	$\sigma_a\left(\overline{E}(Y)\right)$	$M_{mc}\left(s_a\left(\overline{E}(Y)\right)\right)$	$s_{mce}\left(s_a\left(\overline{E}(Y)\right)\right)$	$s_{mc}\left(\overline{E}(Y)\right)$
Simple random	No	.0360	.0358	.0011	.0362
Simple random	Yes	.0381	.0381	.0010	.0383
Stratified	No	.0321	.0320	.0006	.0308
Stratified	Yes	.0340	.0344	.0007	.0336

Table 8 Standard Error Estimates for Sample Minimum Discriminant Information Adjustment Means for Fraction of Students in Schools With Coach

Sampling	Replacement	$\sigma_a\left(\widehat{E}_w(Y)\right)$	$M_{mc}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$	$s_{mce}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$	$s_{mc}\left(\widehat{E}_w(Y)\right)$
Simple random	No	.0315	.0311	.0013	.0331
Simple random	Yes	.0334	.0329	.0015	.0341
Stratified	No	.0286	.0283	.0011	.0284
Stratified	Yes	.0306	.0302	.0020	.0310

the estimated asymptotic standard deviation $M_{mc}\left(s_a\left(\widehat{E}_w(Y)\right)\right)$ or $M_{mc}\left(s_a\left(\overline{E}(Y)\right)\right)$ differs from the corresponding sample standard deviation $s_{mc}\left(\widehat{E}_w(Y)\right)$ or $s_{mc}\left(\overline{E}(Y)\right)$) of the replicates of the sample MDIA mean by approximately .002, as shown in Tables 5 – 8. These deviations are not surprising given the corresponding standard errors of the estimated asymptotic standard deviations for these two cases.

Results for accuracy of the estimation of $E_*(Y)$ present several basic patterns. Sampling without replacement obviously yields greater accuracy than sampling with replacement. The square root of $1 - n/N$ is approximately .94, so that sampling without replacement yields standard errors approximately 6% smaller than corresponding values for sampling with replacement. The precise percentage reduction for stratified sampling is slightly different because of variations in the number of population members in the strata. In this example, stratification is not helpful in estimating the average $M_{\mathcal{U}}(Y)$. As suggested previously, this result reflects the variability in the function v for this case. In contrast, stratification is rather helpful in estimating the weighted average $M_{\mathcal{U}}(SY)/M_{\mathcal{U}}(S)$. For comparable sampling designs, sample MDIA means $\widehat{E}_w(Y)$ provide better estimates than do sample estimates $\overline{E}(Y)$. Thus the prevalence of coaching is more accurately estimated by use of MDIA. Ratios of corresponding asymptotic standard deviations are approximately .9. These ratios are consistent with the previously noted coefficient of determination of .18 from prediction of Y by 1_{Rm} and Z_j, $1 \le j \le 5$, for the square root of $1 - 0.18$ is approximately .9. Alternatively, the ratios of corresponding asymptotic standard deviations of the ratio estimates are approximately .85, indicating that the school sizes have some impact on the precision of the estimates.

The example indicates that MDIA procedures can be used effectively with samples of modest size even when the vector **Z** is not strongly related to the variable Y under study. The example illustrates how to use realistic cases to study accuracy of large-sample approximations for the distributions of MDIA estimates. More examples and more sampling procedures can provide further insight.

Conclusion

Results presented indicate that MDIA can be applied effectively with complex sampling, even with samples of moderate size, in observational studies. Thus the example illustrates gains from use of MDIA to estimate the prevalence of reading coaches in schools. These results extend the known MDIA results for simple random sampling with replacement and for polytomous variables. Effectiveness involves straightforward computations and available large-sample approximations. As noted in the section Calibration Weighting, the MDIA approach can in principle be generalized to treat calibration weights; however, MDIA has distinct advantages in terms of generality of application.

The large-sample results for sample MDIA means are readily generalized to estimation of population parameters much more complex than expectations (Haberman & Yao, 2015). Standard asymptotic arguments (Rao, 1973, pp. 38 – 389) may be applied. For example, let k be a positive integer, and let h be a real Baire function on R^k. Let k-dimensional vector function $\mathbf{Y}$ have elements Y_j for $1 \le j \le k$, and let $w_*\mathbf{Y}_*$ have a finite expectation. If h is continuous at $E_w(\mathbf{Y})$ and $\widehat{E}_w(\mathbf{Y})$ converges to $E_w(\mathbf{Y})$ with probability 1, then $h\left(\widehat{E}_w(\mathbf{Y})\right)$ converges to $h(E_w(\mathbf{Y}))$ with probability 1. In sampling without replacement, if h is continuous in a neighborhood of $E_w(\mathbf{Y})$, $E_w(\mathbf{Y})$ converges in probability to $E_W(\mathbf{Y})$, and $\widehat{E}_w(\mathbf{Y}) - E_w(\mathbf{Y})$ converges in probability to 0, then $h\left(\widehat{E}_w(\mathbf{Y})\right) - h\left(E_w(\mathbf{Y})\right)$ converges in probability to 0.

In the case of asymptotic normality, let h be continuously differentiable on an open subset O of R^k. For $\mathbf{x}$ in O, let $\nabla h(\mathbf{x})$ be the gradient of h at $\mathbf{x}$, and let $\nabla h(\mathbf{x})$ be $\mathbf{0}_k$ for $\mathbf{x}$ in R^k not in O. Let $X_w = [\nabla h(E_w(\mathbf{Y}))]'\mathbf{Y}$, and let $\widehat{\mathbf{X}}_w = \left[\nabla h\left(\widehat{E}_w(\mathbf{Y})\right)\right]'\mathbf{Y}$. Let consistency and asymptotic normality results apply to $\mathbf{c}'\widehat{E}_w(\mathbf{Y})$ for any $\mathbf{c}$ in R^k such that $\mathbf{c}$ is not $\mathbf{0}_k$; let $E_w(\mathbf{Y})$ be in O; and let $\nabla h(E_w(\mathbf{Y}))$ not be $\mathbf{0}_k$. In sampling without replacement, let $E_W(\mathbf{Y})$ be in O, and let $\nabla h(E_W(\mathbf{Y}))$ not be $\mathbf{0}_k$. Let $\sigma_a^2\left(h\left(\widehat{E}_w(\mathbf{Y})\right)\right)$ be $\sigma_a^2\left(\widehat{E}_w\left(X_w\right)\right)$, and let $s_a^2\left(h\left(\widehat{E}_w(\mathbf{Y})\right)\right) = s_a^2\left(\widehat{E}_w\left(\widehat{X}_w\right)\right)$. Then $\left[h\left(\widehat{E}_w(\mathbf{Y})\right) - h\left(E_w(\mathbf{Y})\right)\right]/\sigma_a\left(\widehat{E}_w\left(X_w\right)\right)$ converges in law to $N(0, 1)$. If $s_a^2\left(\widehat{E}_(\mathbf{c}'\mathbf{Y})\right)/\sigma_a^2\left(\widehat{E}_w(\mathbf{c}'\mathbf{Y})\right)$ converges in probability to 1 for any $\mathbf{c}$ in R^k not equal to $\mathbf{0}_k$, then, for $0 < \alpha < 1$, the probability approaches $1 - \alpha$ that $h\left(\widehat{E}_w(\mathbf{Y})\right) - \Phi^{-1}(1 - \alpha/2)\, s_a(h\left(\widehat{E}_w(\mathbf{Y})\right) \le h\left(\mathbf{E}_{w\mathrm{U}}(\mathbf{Y})\right) \le h\left(\widehat{E}_w(\mathbf{Y})\right) + \Phi^{-1}(1 - \alpha/2)\, s_a(h\left(\widehat{E}_w(\mathbf{Y})\right)$.

For instance, if $k = 2$ and $Y_2 = Y_1^2$, then these results are easily applied to estimation of the variance $\sigma_w^2\left(Y_1\right) = E_w\left(Y_2\right) - \left[E_w\left(Y_1\right)\right]^2$. One may let $h(\mathbf{x}) = x_2 - x_1^2$ for $\mathbf{x}$ in R^2 with elements x_1 and x_2. In this case, $O = R^2$, and the gradient $\nabla h(\mathbf{x})$ has elements $-2x_1$ and 1.

In addition to estimation of population parameters that are not expectations, further generalizations may be made to sampling procedures not considered in this report and to less trivial target vectors $\mathbf{z}$. Even for estimation of population means and for sampling methods in this report, more can be studied simply by examining other applications related to causal inference or to linking of educational tests.

References

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, *10*(4), 1100 – 1120. https://doi.org/10.1214/aos/1176345976

Berk, R. H. (1972). Consistency and asymptotic normality of MLEs for exponential models. *Annals of Mathematical Statistics*, *43*(1), 193 – 204. https://doi.org/10.1214/aoms/1177692713

Chen, J., & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, *80*(1), 107 – 116. https://doi.org/10.2307/2336761

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

Csiszár, I. (1975). *i*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*(1), 146 – 158. https://doi.org/10.1214/aop/1176996454

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, *43*(5), 1470 – 1480. https://doi.org/10.1214/aoms/1177692379

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*(4), 427 – 444. https://doi.org/10.1214/aoms/1177731829

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376 – 382. https://doi.org/10.2307/2290268

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013 – 1020. https://doi.org/10.1080/01621459.1993.10476369

Graham, B. S., & de Xavier Pinto, C. C. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, *79*(3), 1053 – 1079. https://doi.org/10.1093/restud/rdr047

Haberman, S. J. (1974). *The analysis of frequency data*. University of Chicago Press.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*(3), 971 – 988. https://doi.org/10.1214/aos/1176346715

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No. No. RM-14-01). Educational Testing Service.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254 – 273. https://doi.org/10.3102/1076998615574772

Haberman, S. J., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, *52*(2), 223 – 251. https://doi.org/10.1111/jedm.12075

Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 363 – 385. https://doi.org/10.1111/bmsp.12052.

Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25 – 46. https://doi.org/10.1093/pan/mpr025

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, *5*, 361 – 374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, *35*(4), 1491 – 1523. https://doi.org/10.1214/aoms/1177700375

Halmos, P. R. (1950). *Measure theory*. Van Nostrand.

Hartley, H. O., & Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, *55*(3), 547 – 557. https://doi.org/10.1093/biomet/55.3.547

Ireland, C. T., Ku, H. H., & Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *Journal of the American Statistical Association*, *64*(32), 1323 – 1341. https://doi.org/10.2307/2286071

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, *55*(1), 179 – 188. https://doi.org/10.1093/biomet/55.1.179

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology*, *36*(2), 145 – 155. https://doi.org/10.1111/j.1751-5823.2010.00099.x

Kullback, S. (1959). *Information theory and statistics*. John Wiley.

Kullback, S. (1971). Marginal homogeneity of multidimensional contingency tables. *Annals of Mathematical Statistics*, *42*(2), 594 – 606. https://doi.org/10.1214/aoms/1177693409

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79 – 86. https://doi.org/10.1214/aoms/1177729694

Lockwood, J. R., McCombs, J. S., & Marsh, J. (2010). Linking reading coaches and student achievement: Evidence from Florida middle schools. *Educational Evaluation and Policy Analysis*, *32*(3), 372 – 388. https://doi.org/10.3102/0162373710373388

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *63*(321), 1 – 28. https://doi.org/10.1080/01621459.1968.11009219

Owen, A. (2001). *Empirical likelihood*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420036152

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). John Wiley. https://doi.org/10.1002/9780470316436

Appendix

Proofs

Proof of Theorem 2

The case of $Y = 1_m$ implies that $M(v)$ converges in probability to 1, so that $\overline{E}(Y)$ converges in probability to $E_\infty(Y)$ whenever Y is a real function on R^m and $M(vY)$ converges in probability to $E_\infty(Y)$. The strong consistency results in Haberman (1984) can be used for weak consistency given substitution of $\overline{E}(Y)$ for $M(Y)$ in all arguments that involve a real function Y on R^m such that Y is continuous at $\mathbf{U}_\infty$ with probability 1 and $|Y| \leq h$. The one significant issue is that β_∞ is in the open set C of vectors $\beta_\infty + \mathbf{b}$ for $\mathbf{b}$ in B. Define the function ℓ_∞ for a d-dimensional vector $\mathbf{b}$ by

$$\ell_\infty(\mathbf{b}) = \mathbf{b}'\mathbf{z} - E_\infty\left(\exp\left(\mathbf{b}'\mathbf{Z}\right)\right). \tag{A1}$$

Because $\ell(\mathbf{b})$ converges to $\ell_\infty(\mathbf{b})$ for $\mathbf{b}$ in C, ℓ converges uniformly to ℓ_∞ on any nonempty closed and bounded subset D of C. Because ℓ_∞ is strictly concave on C and has a unique minimum at β_∞, β converges to β_∞, so that $E_w(Y)$ converges to $E_\infty(WY)$. For $\mathbf{b}$ in C, $\overline{\ell}(\mathbf{b})$ converges in probability to $\ell_\infty(\mathbf{b})$. At this point, it follows that for any closed and bounded subset D of C, $\sup|\overline{\ell}(\mathbf{b}) - \ell_\infty(\mathbf{b})|$ converges in probability to 0 (Andersen & Gill, 1982). Therefore $\widehat{\beta}$ converges in probability to β_∞. Remaining arguments require no special comment.

Proof of Corollary 1

If Y is a real function and Y is bounded, then $M(vY) - E_*(Y)$ converges in probability to 0. If Y is continuous at $\mathbf{U}_\infty$ with probability 1 and $|Y| \leq g$, then consider $Y_t = \max(-t, \min(Y, t))$ for positive integers t. Then $\sigma^2(M(Y_t v))$ converges to 0 and $M(vY_t) - E_*(Y_t)$ converges in probability to 1. Let ϵ and $\delta > 0$ be positive real numbers. The difference $Y - Y_t$ satisfies

$$(\epsilon/2)\,P\left(|M\left(v\left(Y - Y_t\right)\right)| > \epsilon/2\right) \leq E(|\,M\left(v\left(Y - Y_t\right)|\right) \leq E\left(M\left(v|Y - Y_t|\right)\right) = E_*\left(|Y - Y_t|\right),$$

and $E_*(|Y - Y_t|)$ converges to $E_\infty(|Y - Y_t|)$. Because $|Y - Y_t| \leq h$ and $|Y - Y_t|$ converges to 0 as t approaches ∞, a positive integer t exists such that $E_\infty(|Y - Y_t|) < \epsilon\delta/4$. A positive integer n' exists such that, for $n \geq n'$, $P(|M(v(Y - Y_t)) - E_*(Y - Y_t)| > \epsilon/2) < \delta/2$, $|E_*(|Y - Y_t|) - E_\infty(|Y - Y_t|)| < \epsilon\delta/4$, $P(|M(v(Y - Y_t))| > \epsilon/2) < \delta/2$, and $P(|M(vY) - E_\infty(Y)| > \epsilon) < \delta$. It follows that $M(vY)$ converges in probability to $E_\infty(Y)$.

Other Examples

Example A1. For a slightly more complex case, let n_1 and n_2 be positive integers. Let n_1 be constant, let n_2 approach ∞, and let $n = n_1 n_2$. Define $\mathbf{T}$ and $v = g$ as in Example 3. Let $\mathbf{U}_i$, $i \geq 1$, all have the same distribution as $\mathbf{T}$. For each positive integer j, let $\mathbf{U}_{Aj}$ be the $m \times n_1$ matrix with columns $\mathbf{U}_i$, $i = n_1(j-1) + k$, for $1 \leq k \leq n_1$. Let $\mathbf{U}_{Aj}$ be mutually independent and identically distributed for $j \geq 1$. For any real function Y on R^m and any positive integer j, let $L_j(Y)$ be the average of $Y_{n_1(j-1)+k}$ for $1 \leq k \leq n_1$, so that $M(Y)$ is the average of $L_j(Y)$ for $1 \leq j \leq n_2$. By assumption, the $L_j(Y)$ are identically distributed and mutually independent for $j \geq 1$. Let $v = g$. let X be a real function on R^m such that $E_*(|X|) < \infty$. Then $E(M(vX)) = E_*(X)$. In the simplest cases, $v = g = 1_{Rm}$ and $\mathbf{T}$ may be set equal to $\mathbf{U}_*$ as in Example 3. In general, for each positive integer j, $E(L_j(vX)) = E_*(X)$. If X_* has a finite variance, then each $L_j(vX)$, $j \geq 1$, has a finite variance $\sigma^2(L_1(vX))$, and $\sigma^2(M(vX)) = n_2^{-1}\sigma^2\left(L_1(vX)\right)$. The strong law of large numbers implies that $M(vX)$ converges to $E_*(X)$ with probability 1 if X is a real function on R^m such that $E_*(|X|)$ is finite.

This case applies to two-stage sampling with replacement with n_2 primary sampling units and n_1 secondary sampling units per primary sampling unit. One model for this case has $1 \leq m' < m$, with $\mathbf{U}_{i1}$, $i \geq 1$, the random vector of dimension m' with elements the initial m' elements of $\mathbf{U}_i$ and $\mathbf{U}_{i2}$ the random vector of dimension $m - m'$ with elements the last $m - m'$ elements of $\mathbf{U}_i$. $\mathbf{U}_{i1}$ correspond to the primary sampling units, and $\mathbf{U}_{i2}$ correspond to the secondary sampling units. Let $\mathbf{T}_1$ be the random vector of dimension m' with elements equal to initial m' elements of $\mathbf{T}$, and let $\mathbf{T}_2$ be the random vector of dimension $m - m'$ with elements equal to the last $m - m'$ elements of $\mathbf{T}$. The assumption that $\mathbf{U}_i$ and $\mathbf{T}$

have the same distribution implies that $\mathbf{U}_{i1}$ and $\mathbf{T}_1$ have the same distribution, and the conditional distribution of $\mathbf{U}_{i2}$ given $\mathbf{U}_{i1}$ is the same with probability 1 as the conditional distribution of $\mathbf{T}_2$ given $\mathbf{T}_1$. Let $\mathbf{U}_{i1}$ be constant for $n_1(j-1) < i \leq n_1 j$ and $1 \leq j \leq n_1$; let $\mathbf{U}_{i2}$, $(n-1)j < i \leq nj$, be conditionally independent given the common value of $\mathbf{U}_{i1}$ for $(n-1)j < i \leq nj$; and let the conditional distribution of $\mathbf{U}_2$ given $\mathbf{U}_1$ be the same as the conditional distribution of $\mathbf{T}_2$ given $\mathbf{T}_1$. Then all requirements for $\mathbf{U}_i$, $i \geq 1$, are satisfied.

Example A2. In Example 3, let X be a real function on R^m such that $E_*(X^2)$ is finite. Let $\tau^2(X) = E_*(v[X - E_*(X)]^2)$. Then $n\sigma^2(M(vX)) = \tau^2(X)$. The central limit theorem applies to $n^{1/2}[M(vX) - E_*(X)]$. It follows that the condition on convergence in law holds. Because $E_*(wr_w(Y)) = 0$, it also follows that $\tau^2\left(wr_w(Y)\right) = n\sigma_a^2\left(\widehat{E}_w(Y)\right) = E_w\left(vw\left[r_w(Y)\right]^2\right)$. In Haberman (1984), this case applies with $v = 1_{Rm}$. In this instance, $\tau^2(X) = \sigma_*^2(X)$ and $\tau^2\left(wr_w(Y)\right) = \sigma_*^2\left(wr_w(Y)\right) = E_w\left(w\left[r_w(Y)\right]^2\right)$.

Example A3. In Example 4, $n\sigma^2(M(vX))$ converges to

$$\tau^2(X) = \sum_{j=1}^{J} \left[p_{H*}(j)\right]^2 E_*\left(g\left[X - E_*\left(X|H=j\right)\right]^2 |H=j\right)/p_H(j).$$

The central limit theorem applies to $n_j^{1/2}\left[M\left(gX|H=j\right) - E_0\left(gX|H=j\right)\right]$. As n_j becomes large, $n_j^{1/2}\left[M\left(gX|H=j\right) - E_0\left(gX|H=j\right)\right]$ converges in law to $N\left(0, \sigma_0^2\left(gX|H=j\right)\right)$. Because the $M(gX|H=j)$, $1 \leq j \leq J$, are mutually independent, $n^{1/2}[M(vX) - E_*(X)]$ converges in law to $N(0, \tau^2(X))$.

If $\delta_j(H)$, $1 \leq j \leq J \leq d$, is a linear function of $\mathbf{Z}$, then the formula for $\sigma_a^2\left(\widehat{E}_w(Y)\right)$ simplifies because $E_w(\delta_j(H)r_w(Y)) = 0$ for $1 \leq j \leq J$, so that $\sigma_a^2\left(\widehat{E}(Y)\right) = \sum_{j=1}^{J}\left[p_{H*}(j)\right]^2 E_*\left(v\left[wr_w(Y)\right]^2 |H=j\right)$ for $1 \leq j \leq J$.

Example A4. In Example A1, let X be a real function on R^m such that $E_*(X^2) < \infty$. Then $n\sigma^2(M(vX)) = \tau^2(X) = n_1\sigma^2(L_1(vX))$, and the central limit theorem implies that $n^{1/2}[M(vX) - E_*(X)] = (n_1 n_2)^{1/2}[M(vX) - E_*(X)]$ converges in law to $N(0, \tau^2(X))$. In this case, $\sigma_a^2\left(\widehat{E}_w(Y)\right) = E\left(\left[L_1\left(vwr_w(Y)\right)\right]^2\right)/n_2$.

In the two-stage case, $n\sigma^2(M(vX)) = n_1\sigma^2(E(v(\mathbf{T})X(\mathbf{T})|\,\mathbf{T}_1)) + E(\sigma^2(v(\mathbf{T})X(\mathbf{T})|\,\mathbf{T}_1))$. Here the random variable $E(v(\mathbf{T}X(\mathbf{T})|\,\mathbf{T}_1)$ is the expected value of $v(\mathbf{T})X(\mathbf{T})$ given $\mathbf{T}_1$ and the random variable $\sigma^2(v(\mathbf{T})X(\mathbf{T})|\,\mathbf{T}_1)$ is the variance of $v(\mathbf{T})X(\mathbf{T})$ given $\mathbf{T}_1$. Both variables can be defined to be real.

Example A5. In Example A2, $s^2(X) = (n-1)^{-1}M([vX - M(vX)]^2) = (n-1)^{-1}\{M(v^2X^2) - [M(vX)]^2\} \leq [n/(n-1)]\sup(v)M(vX^2)/n$ if $n > 1$. As is well known, $E(s^2(X)) = \sigma^2(M(vX))$ if X has a finite variance. Because $M(v^2X^2)$ converges with probability 1 to $E_*(vX^2)$ and $M(vX)$ converges with probability 1 to $E_*(X)$, $ns^2(X)$ converges with probability 1 to $\tau^2(X)$. In this case, $M\left(v\widehat{wr}_w(Y)\right) = 0$, so that $s_a^2\left(\widehat{E}_w(Y)\right) = (n-1)^{-1}M\left(\left[v\widehat{E}_w(Y)\right]^2\right)$. This case is considered by Haberman (1984) for $v = 1_{Rm}$. The divisor $n - 1$ is replaced there by n without any change in the basic results.

Example A6. In Example A3, if $n_j > 1$ for $1 \leq j \leq J$, then

$$s^2(X) = \sum_{j=1}^{J}\left(n_j - 1\right)^{-1}\left[p_{H*}(j)\right]^2 M\left([vX - M(vX)]^2 |H=j\right). \tag{A2}$$

If X_*^2 has a finite expectation, then $E(s^2(X)) = \sigma^2(M(vX))$. If γ is the maximum of $[n_j/(n-1)][np_{H^*}(j)/n_j]^2\sup(v)$ for $1 \leq j \leq J$ and for n sufficiently large that each $n_j > 1$, then $s^2(X) \leq \gamma M(vX^2)/n$. Because $M([vX - M(vX)]^2|\,H=j) = M([vX]^2|\,H=j) - [M(vX|\,H=j)]^2$ for $1 \leq j \leq J$, $ns^2(X)$ converges with probability 1 to $\tau^2(X)$.

If $\delta_j(H)$ is a linear function of $\mathbf{Z}$ for $1 \leq j \leq J \leq d$, then

$$s_a^2\left(\widehat{E}_w(Y)\right) = \sum_{j=1}^{J}\left(n_j - 1\right)^{-1}\left[p_{H*}(j)\right]^2 M\left(\left[v\widehat{wr}_w(Y)\right]^2 |H=j\right). \tag{A3}$$

Example A7. In Example A4, if $n_2 > 1$, then

$$s^2(X) = \left[n_2\left(n_2-1\right)\right]^{-1}\sum_{j=1}^{n_2}\left[L_j(vX) - M(vX)\right]^2 \le n_1^3\left[n_2/\left(n_2-1\right)\right]\sup(v)\,M\left(vX^2\right)/n. \tag{A4}$$

If X_*^2 has a finite expectation, then the expectation of $s^2(X)$ is $\sigma^2(M(vX))$, and $ns^2(X)$ converges to $\tau^2(X)$ with probability 1. In this case, $s_a^2\left(\widehat{E}_w(Y)\right) = \left[n_2\left(n_2-1\right)\right]^{-1}\sum_{j=1}^{n_2}\left[L_j(v\widehat{w}\widehat{r}_w(Y)\right]^2$.

Suggested citation:

Yao, L., Haberman, S., McCaffrey, D. F. & Lockwood, J. R. (2020). *Large-sample properties of minimum discriminant information adjustment estimates under complex sampling designs* (Research Report No. RR-20-13). Educational Testing Service. https://doi.org/10.1002/ets2.12297

Action Editor: Gautam Puhan

Reviewers: Katherine Castellano and Hongwen Guo