The Effect of Cluster Sampling on the Covariance and Correlation Matrices of Sample Distribution Functions

I. Park, Westat; J. L. Eltinge, Bureau of Labor Statistics and Texas A&M University
I. Park, Westat, 1650 Research Boulevard, Rockville, Maryland 20850; parki@westat.com

Key Words: Superpopulation; Model-based inference; Intra-cluster correlation; Sample distribution; Sample quantile; U.S. Third National Health and Nutrition Examination Survey (NHANES III).

1. Introduction

One often needs to estimate distribution functions and quantiles of a study characteristic Y for the analysis of complex survey data. A customary design-based estimator of a distribution function $F(x)$ is defined as a weighted proportion of sampled values y that are not greater than a given value $x \in \mathbb{R}$, and its variance estimator can be derived using the linearization method. Also, point estimators of a quantile and associated variance estimators can be obtained through the inverse relationship between the quantile and the distribution function and the Bahadur representation. See, e.g., Woodruff (1952), Rao et al. (1990), Francisco and Fuller (1991) and references cited therein.

Due to the sparseness of data in the tail region of a distribution, design-based estimation may perform relatively poorly for estimation of either distribution functions or quantiles at extreme points. Accordingly, one may seek to obtain better quantile point estimation and inference methods by fitting an appropriate parametric model to data from these tail regions. Park and Eltinge (1999), for example, discussed generalized least squares (GLS) estimators of the quantiles in the tail region. These GLS estimators required the use of estimators of the covariance matrices of the approximate distribution of the initial design-based distribution function estimator vector. Direct design-based estimators of these covariance matrices may be unstable in some cases. Consequently, it is of interest to consider approximation methods that may lead to more stable covariance matrix estimators.

This paper develops one such approximation by examining the effect of cluster sampling on the covariance and correlation matrix of sample distribution functions based on a superpopulation model. Under a simplied two stage sampling design, the classical two-way nested random effects model leads to a covariance matrix approximation that depends on an intra-cluster correlation term. The derived result are then compared with the empirical result from medical examination data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

The authors thank Drs. D. Brody, A. Looker and V. L. Parsons for providing the NHANES III data discussed here, and for helpful comments on related statistical and substantive issues. This research was supported in part by the U.S. National Center for Health Statistics. The views expressed here are those of the authors and do not necessarily reflect the policies of the U.S. National Center for Health Statistics nor the U.S. Bureau of Labor Statistics.

2. Framework and Assumptions

Suppose that the sample consists of I_0 clusters and that each cluster has the same number of units J_0. Thus there are $n_0 = I_0 J_0$ units in the sample. Let $y_{ij}, i = 1, \ldots, I_0, j = 1, \ldots, J_0$, be observed values associated with the jth unit in the ith cluster. Using a model-based approach, let y_{ij} be realizations of a customary two-way nested random effects model:

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \tag{1}$$

where μ is an unknown constant and α_i and e_{ij} are uncorrelated normal random variables with zero means and variances $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$ and $\mathrm{Var}(e_{ij}) = \sigma_e^2$, respectively. The total variation of Y_{ij} is given as $\mathrm{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_e^2 \equiv \sigma^2$. Here, α_i and e_{ij} are the two sets of model random variables associated with primary sampling units (i.e., clusters) and ultimate sampling units of the chosen clusters, respectively. Thus their respective variance components σ_α^2 and σ_e^2 represent the variation of Y between clusters and within clusters. Under the model (1), routine calculations (e.g., Graybill, 1976, Theorem 15.1.1) show that

$$E(Y_{ij}) = \mu$$

and

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_\alpha^2 + \sigma_e^2 & \text{if } i = i', j = j', \\ \sigma_\alpha^2 & \text{if } i = i', j \neq j', \\ 0 & \text{if } i \neq i'. \end{cases}$$

The distribution function of Y under model (1) is given as $F(y) = \Phi(\frac{y-\mu}{\sigma}), y \in \mathbb{R}$, where $\Phi(\cdot)$ is the standard normal distribution function. Since $\Phi(x)$ is a strictly increasing continuous function of x, the pth quantile is uniquely expressed as

$$q_p = \mu + \sigma z_p, \tag{2}$$

where $z_p = \Phi^{-1}(p)$ is the upper pth percentile of the standard normal distribution. Suppose that we consider a set of k distinct prespecified probability values $0 < \pi_1 < \ldots < \pi_k < 1$. From the above expression for a normal quantile, those k probability values determine k non-overlapping cells of the form $I_l = (q_{\pi_{l-1}}, q_{\pi_l}], l = 1, \ldots, k$, where $\pi_0 = 0$ or equivalently $q_{\pi_0} = -\infty$. Note that the open interval (q_{π_k}, ∞) is excluded from consideration. In addition, the corresponding $k \times 1$ vector of cell probabilities, $\tau = (\tau_1, \ldots, \tau_k)'$ leads to the $k \times 1$ vector of distribution functions evaluated at $y = q_{\pi_l}, l = 1, \ldots, k$:

$$\pi = F(q_\pi) = A\tau, \tag{3}$$

where $\tau_l = \pi_l - \pi_{l-1}, \pi_0 = 0$ and A is a $k \times k$ lower triangular matrix of ones.

3. Estimation of Distribution Functions and Related Covariance and Correlation Matrices from Clustered Observations

Let $\delta_{(ij,l)}$ denote indicator variables which equal 1 if $Y_{ij} \in I_l$ or 0 otherwise, where $i = 1, \ldots, I_0, j = 1, \ldots, J_0$ and $l = 1, \ldots, k$. Throughout the remainder of this paper, we will use the term "IID-based" to describe point estimators and variance estimators developed under the (incorrect) assumption that our $I_0 J_0$ observations y_{ij} are independently and identically distributed. A customary IID-based estimator of the $k \times 1$ vector τ computed from the observations y_{ij} is given by

$$\hat{\tau}_C = (\hat{\tau}_{C1}, \ldots, \hat{\tau}_{Ck})',$$

where

$$\hat{\tau}_{Cl} = n_0^{-1} \sum_{i=1}^{I_0} \sum_{j=1}^{J_0} \delta_{(ij,l)}$$

are proportions of sample units with $y \in I_l$. Letting $\hat{\tau}_{Cl:i} = J_0^{-1} \sum_{j=1}^{J_0} \delta_{(ij,l)}$ denote within-cluster estimators of τ_l for cell l, $\hat{\tau}_{Cl}$ can also be written as

$$\hat{\tau}_{Cl} = I_0^{-1} \sum_{i=1}^{I_0} \hat{\tau}_{Cl:i}. \tag{4}$$

From expression (3), an IID-based estimator of the $k \times 1$ vector $F(q_\pi)$ is then given as

$$\hat{F}_C(q_\pi) = A\hat{\tau}_C. \tag{5}$$

Now we investigate the mean and covariance of both estimators $\hat{\tau}_C$ and $\hat{F}_C(q_\pi)$. For a fixed l, we see that $\hat{\tau}_{Cl}$ is a simple sample mean of n_0 identically distributed zero-one random variables $\delta_{(ij,l)}$ with $E[\delta_{(ij,l)}] = \Pr(Y_{ij} \in I_l) = \tau_l$ for any i and j. Thus it follows that

$$E(\hat{\tau}_{Cl}) = \tau_l$$

and $E(\hat{\tau}_C) = \tau$, which in turn gives $E[\hat{F}_C(q_\pi)] = F(q_\pi)$ from expression (5).

A double expectation argument (e.g., Cochran, 1977, Section 10.2) indicates

$$E(\hat{\tau}_l) = E_\alpha[E(\hat{\tau}_l|\alpha)]$$

where $E(\cdot|\alpha)$ denotes the conditional expectation given $\alpha = (\alpha_1, \ldots, \alpha_{I_0})'$ and $E_\alpha(\cdot)$ denotes the expectation evaluated with respect to the marginal distribution of a random vector α. Under model (1), the random variables $Y_{ij}, j = 1, \ldots, J_0$ given α_i are IID $N(\mu + \alpha_i, \sigma_e^2)$. Thus, for a given α_i and l, the random variables $\delta_{(ij,l)}$ are IID with

$$\begin{aligned} E[\delta_{(ij,l)}|\alpha] &= \Pr(Y_{ij} \in I_l|\alpha_i) \\ &= \tau_l(\alpha_i), \end{aligned}$$

where $\Pr(\cdot|\alpha_i)$ denotes the probability conditional upon α_i under the model (1) and

$$\tau_l(\alpha_i) = \Phi\left(\frac{q_{\pi_l} - \mu - \alpha_i}{\sigma_e}\right) - \Phi\left(\frac{q_{\pi_{l-1}} - \mu - \alpha_i}{\sigma_e}\right). \tag{6}$$

Also, for a fixed i, $\hat{\tau}_{Cl:i}$ is a sample mean of J_0 random variables $\delta_{(ij,l)}$ by definition. Thus it follows that

$$E(\hat{\tau}_{Cl:i}|\alpha) = \tau_l(\alpha_i).$$

Consequently, it follows from expression (4) that

$$E(\hat{\tau}_{Cl}|\alpha) = I_0^{-1} \sum_{i=1}^{I_0} \tau_l(\alpha_i). \tag{7}$$

Note that for a given cell l, the random variables $\tau_l(\alpha_i)$ are functions of IID random variables $\alpha_i, i = 1, \ldots, I_0$. Thus, the $\tau_l(\alpha_i)$ are also IID random variables. Using the fact that $E(\hat{\tau}_{Cl}|\alpha)$ is a simple sample mean of I_0 IID random variables $\tau_l(\alpha_i)$ as described in expression (7), we may conclude from the unbiasedness of $\hat{\tau}_{Cl}$ and the double expectation argument that

$$E_\alpha[\tau_l(\alpha_i)] = \tau_l \tag{8}$$

for any l and i.

To derive covariance matrices of $\hat{\tau}_C$ and $\hat{F}_C(q_\pi)$, we use a multivariate extension of a standard expression (e.g., Tucker, 1998, Theorem 7, p. 76) for a covariance as the sum of the covariance of conditional expectations and the expected values of a conditional covariance. That is, for any three random vectors X, Y and Z,

$$\begin{aligned}\mathrm{Cov}(X,Y) = \; & \mathrm{Cov}_Z[E(X|Z), E(Y|Z)] \\ & + E_Z[\mathrm{Cov}(X,Y|Z)],\end{aligned} \tag{9}$$

where $E_Z(\cdot)$ and $\mathrm{Cov}_Z(\cdot)$ denote the mean and covariance evaluated with respect to the marginal distribution of Z.

Define $\Sigma_\tau = \mathrm{diag}(\tau) - \tau\tau'$ and $\tau(\alpha_1) = [\tau_1(\alpha_1), \ldots, \tau_k(\alpha_1)]'$. An application of decomposition (9) under the model (1) yields

$$\mathrm{Cov}(\hat{\tau}_C) = n_0^{-1} C_\tau, \tag{10}$$

where

$$C_\tau = \Sigma_\tau + (J_0 - 1)\mathrm{Cov}_\alpha[\tau(\alpha_1)]$$

and $\mathrm{Cov}_\alpha(\cdot)$ denotes covariance evaluated with respect to the distribution of α. From expression (5), it follows that

$$\mathrm{Cov}[\hat{F}_C(q_\pi)] = n_0^{-1} A C_\tau A'. \tag{11}$$

To investigate the effect of cluster sampling on the covariance matrices of $\hat{\tau}_C$ and $\hat{F}_C(q_\pi)$, we consider estimators of τ and $F(q_\pi)$ and their covariance matrices under a reference model which contains no clustering effect and has the same marginal distribution of $N(\mu, \sigma^2)$. Thus, the observation y_{ij} can be modeled as a realization of

$$Y_{0ij} = \mu + e_{0ij}, \tag{12}$$

where e_{0ij} is a $N(0, \sigma^2)$ random variate. In parallel with the discussion for the model (1), the IID-based estimators under the reference model (12) are given by $\hat{\tau}_0 = (\hat{\tau}_{01}, \ldots, \hat{\tau}_{0k})'$ and $\hat{F}_0(q_\pi) = A\hat{\tau}_0$, where $\hat{\tau}_{0l} = n_0^{-1} \sum_{i=1}^{I_0} \sum_{j=1}^{J_0} \delta_{(0ij,l)}$ and $\delta_{0ij,l} = 1$ if $Y_{0ij} \in I_l$ and 0 otherwise. Let $E_0(\cdot)$ and $\mathrm{Cov}_0(\cdot)$ denote the mean and covariance evaluated with respect to the distribution induced by the reference model. Standard arguments (e.g., Agresti 1990, Section 12.1.5) show that $E_0(\hat{\tau}_0) = \tau$ and

$$\mathrm{Cov}_0(\hat{\tau}_0) = n_0^{-1}\Sigma_\tau.$$

Since $\hat{F}_0(q_\pi) = A\hat{\tau}_0$, it follows that $E_0[\hat{F}_0(q_\pi)] = F(q_\pi)$ and

$$\mathrm{Cov}_0[\hat{F}_0(q_\pi)] = n_0^{-1} A \Sigma_\tau A'.$$

In this context, results (10) and (11) under the model (1) can be expressed in the following forms reflecting the effect of clustering on the covariance matrices of $\hat{\tau}_C$ and $\hat{F}_C(q_\pi)$:

$$\mathrm{Cov}(\hat{\tau}_C) = \mathrm{Cov}_0(\hat{\tau}_0) + n_0^{-1}(J_0 - 1)\mathrm{Cov}_\alpha[\tau(\alpha_1)] \tag{13}$$

and

$$\begin{aligned}\mathrm{Cov}[\hat{F}_C(q_\pi)] \; = \; & \mathrm{Cov}_0[\hat{F}_0(q_\pi)] \\ & + n_0^{-1}(J_0 - 1) A \mathrm{Cov}_\alpha[\tau(\alpha_1)] A',\end{aligned} \tag{14}$$

which are extensions of well known expressions for the variance of a sample mean under cluster sampling (e.g., Cochran, 1977, p. 242; and Skinner, 1989, pp. 36-38). The above results show that clustering in the sample design under the model (1) leads to an augmentation of the associated IID-based covariance matrices by a multiple of the covariance matrix of a $k \times 1$ random vector $\tau(\alpha_1)$, i.e., the conditional expections of cell-membership indicator variables $\delta_{(ij,l)}$ within the ith cluster.

Note that both expressions (13) and (14) contain the term $\mathrm{Cov}_\alpha[\tau(\alpha_1)]$, which is not readily evaluated in closed form. From expression (8), the (l, l')th element of $\mathrm{Cov}_\alpha[\tau(\alpha_1)]$ can be written as

$$\mathrm{Cov}_\alpha[\tau_l(\alpha_1), \tau_{l'}(\alpha_1)] = E_\alpha[\tau_l(\alpha_1)\tau_{l'}(\alpha_1)] - \tau_l\tau_{l'}.$$

Note also from expression (6) that $\tau_l(\alpha_1)$ is the difference of two functions of a random variable α_1 in the form $\Phi(\frac{x-\mu-\alpha_1}{\sigma_e}), x \in \mathbb{R}$. Thus the explicit evaluation of the expectation of products $\tau_l(\alpha_1)\tau_{l'}(\alpha_1)$ may be approximated by a linear function in powers of the ratio (σ_α/σ_e) under the following assumption.

(C) The ratio (σ_α/σ_e) converges to zero.

Under condition (C), it can be shown that, for any l and l',

$$\mathrm{Cov}_\alpha[\tau_l(\alpha_1), \tau_{l'}(\alpha_1)] - C_{\alpha ll'} = O\left[\left(\frac{\sigma_\alpha}{\sigma_e}\right)^4\right],$$

where

$$C_{\alpha ll'} = -(\Phi_{0l} - \tau_l)(\Phi_{0l'} - \tau_{l'}) + \Phi_{1l}\Phi_{1l'}\left(\frac{\sigma_\alpha}{\sigma_e}\right)^2,$$
$$\Phi_{01} = \Phi\left(\frac{q_{\pi_l} - \mu}{\sigma_e}\right) - \Phi\left(\frac{q_{\pi_{l-1}} - \mu}{\sigma_e}\right),$$
$$\Phi_{1l} = \phi\left(\frac{q_{\pi_l} - \mu}{\sigma_e}\right) - \phi\left(\frac{q_{\pi_{l-1}} - \mu}{\sigma_e}\right),$$

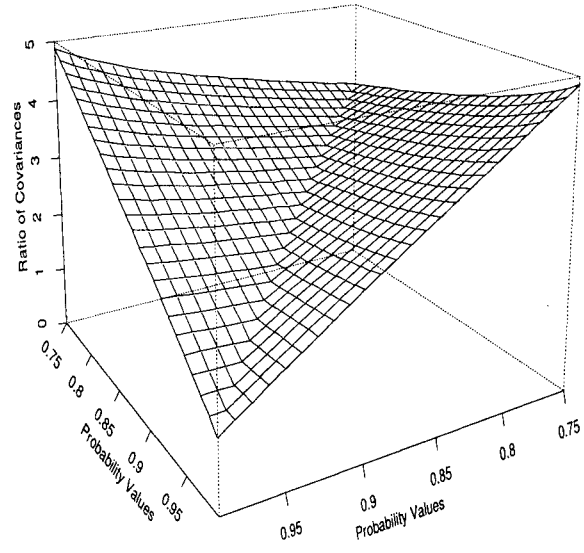
and $\Phi(\cdot)$ and $\phi(\cdot)$ are the distribution and density functions of the standard normal distribution, respectively. Furthermore, the approximate expressions $C_{\alpha ll'} = C_{\alpha ll'}(\gamma)$ are functions of the unknown parameter vector $\gamma = (\mu, \sigma_\alpha^2, \sigma_e^2)'$ of the model (1). Use of a consistent estimator $\hat{\gamma} = (\hat{\mu}, \hat{\sigma}_\alpha^2, \hat{\sigma}_e^2)'$ based on the observed sample data leads to the corresponding quantile estimates $\hat{q}_{\pi_l} = \hat{\mu} + \hat{\sigma} z_{\pi_l}$ from expression (2), giving estimates $\hat{C}_{\alpha ll'} = C_{\alpha ll'}(\hat{\gamma})$, where $\hat{\sigma}^2 = \hat{\sigma}_e^2 + \hat{\sigma}_\alpha^2$. Thus, we may obtain estimators $\widehat{\mathrm{Cov}}(\hat{\tau}_C)$ and $\widehat{\mathrm{Cov}}[\hat{F}_C(\hat{q}_\pi)]$, say.

4. Application to NHANES III Data on ln(LEAD) for Children

4.1 Sampling Scheme and Parameter Estimation

We applied the proposed methods to blood lead data in the natural logarithm scale, ln(LEAD), measured for children of all races aged 1-5 covered by Phase 2 (1991-1994) of the U.S. Third National Health and Nutrition and Examination Survey (NHANES III). For some general background on NHANES III, see National Center for Health Statistics (1996). For analysis purposes, the data may be treated as involving $L = 23$ strata with $n_h = 2$ primary sampling units (usually counties) selected with replacement from each stratum of size $N_h \geq 2$, where $h = 1, \ldots, L$. Selection of PSUs is assumed to be independent across L strata. Additional levels of sampling select secondary units, households and individual persons. Each sampled individual is then interviewed and asked to participate in a medical examination including blood lead measurement. Let y_{hij} be the observed values of Y for person j among sampled n_{hi} individuals from the sampled first-stage cluster (h, i) of size N_{hi} in stratum h, where $i = 1, \ldots, n_h$ and $j = 1, \ldots, n_{hi}$.

Figure 1: The 25×25 Display of the Ratios of the Approximate Covariances under the Two-way Nested Random Effect Model to the Exact Covariances under the IID Model Based on the Phase 2 of NHANES III Data on log(LEAD) for Children of All Races Aged 1-5. The covariance ratios are plotted at the corresponding pairs of the probability values $p = 0.75(0.01)0.99$.



For this group, there are $n = 2,392$ participants selected from $N = \sum_{h=1}^{L}\sum_{i=1}^{N_h} N_{hi}$ individuals, where $n = \sum_{h=1}^{L}\sum_{i=1}^{n_h} n_{hi}$.

Given the sampling weights w_{hij} associated with y_{hij}, a customary design-based sample distribution function at $y \in \mathbb{R}$ is then given by

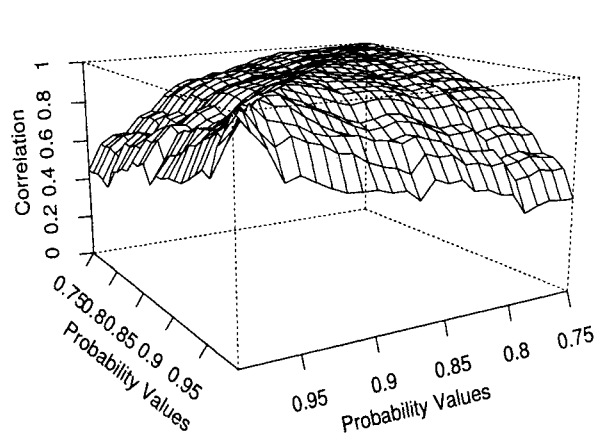
$$\hat{F}(y) = \hat{N}^{-1}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\delta_{y_{hij}}(y),$$

where $\hat{N} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}$ is the estimated total number of ultimate units (persons) in the finite population and $\delta_{y_{hij}}(y) = 1$ if $y_{hij} \leq y$ or 0 otherwise. The corresponding pth design-based sample quantile of Y is defined by

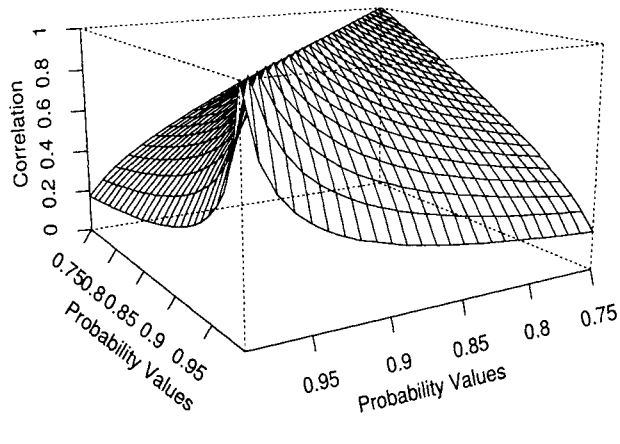
$$\hat{q}_{p:D} = \inf\{y : \hat{F}(y) \geq p\}.$$

To apply the approximate expression in Section 3 to data from stratified multistage sampling, we may need to approximate an unbalanced complex

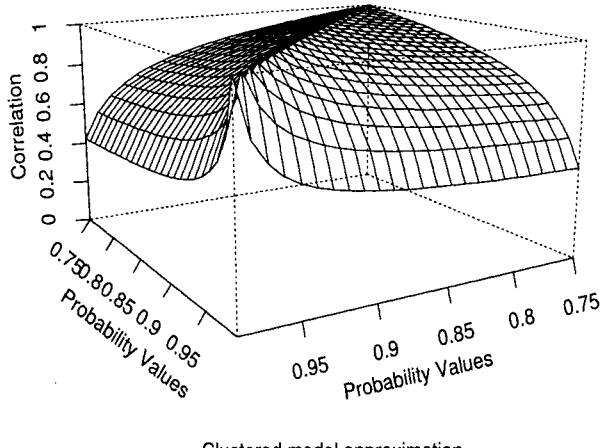
Figure 2: Comparisons of Correlation Matrices of the Estimated Distribution Functions for the Phase 2 of NHANES III Data on log(LEAD) for Children of All Races Aged 1-5.



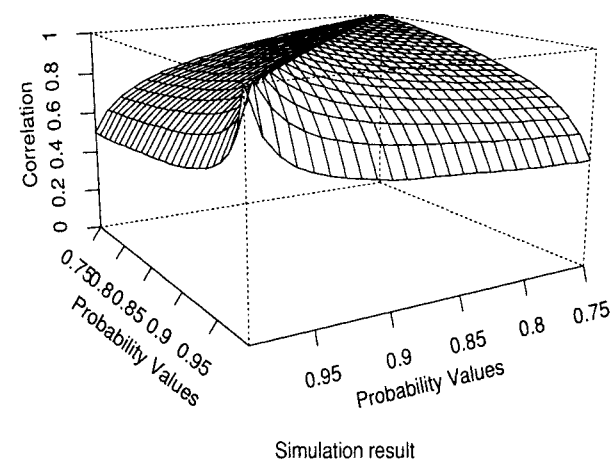
Design-based



IID Model



Clustered model approximation



Simulation result

survey sample design with a simpler balanced cluster sampling framework.

In the light of this reasoning, the parameter vector $\gamma = (\mu, \sigma_e^2, \sigma_\alpha^2)'$ may be estimated by $\hat{\gamma} = (\hat{\mu}, \hat{\sigma}_e^2, \hat{\sigma}_\alpha^2)'$, where

$$\hat{\mu} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}},$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{h=1}^{L}(\bar{y}_{h1\cdot} - \bar{y}_{h2\cdot})^2}{2L},$$

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}(y_{hij} - \bar{y}_{hi\cdot})^2}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}}$$

and $\bar{y}_{hi\cdot} = (\sum_{j=1}^{n_{hi}} w_{hij})^{-1} \sum_{j=1}^{n_{hi}} w_{hij} y_{hij}$. All three parameter estimators may be interpreted as combinations of L estimators, each of which is based on cluster sampling data within each of L strata.

4.2 Application of Proposed Methods

Public health interest in serum lead generally focuses on higher levels of concentration, so we will consider the $k = 25$ probability values $p = 0.75(0.01)0.99$. In addition, some authors have found that ln(LEAD) data are fitted quite well by a normal distribution. See, e.g, Hasselblad et al. (1980), Pirkle et al. (1994) and Park and Eltinge (1999). Also, ad hoc estimation gives $\hat{\mu} = 1.0082$, $\hat{\sigma}_e^2 = 0.4338$ and $\hat{\sigma}_\alpha^2 = 0.0429$. Thus, we have a relatively small between-cluster variance ratio $(\hat{\sigma}_\alpha/\hat{\sigma}_e)^2 = 0.0990$, which indicates that the approximation discussed in Section 3 may be adequate. Moreover, the estimated intracluster correlation $\hat{\rho} = \hat{\sigma}_\alpha^2/(\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2)$ (e.g., Skinner (1989)) is given

as 0.0901 which shows that the within-cluster variation is approximately 9% of the overall variation $\hat{\sigma}^2$ among data. Figure 1 shows that the variance ratio on the diagonal decreases as the probability value becomes larger. In addition, the covariance ratio on the off-diagonal increases as two probability values in a pair are farther apart. Note that all the ratios are greater than 1, which implies that clustering inflates covariances. This observation illustrates the augmented part derived in expression (14).

To examine further the effect of clustering on the covariances independently of the scaling, Figure 2 presents the four associated correlation matrices. Comparison of the top two plots of Figure 2 displays the effect on the correlation matrix of a vector of 25 estimated distribution function values with reference to the Model/IID assumption. On the other hand, the bottom two plots of Figure 2 compare the approximated correlations and simulated correlations with 10,000 replications under the aforementioned two-way nested random effect model. See Park (1999) for a detailed discussion of the simulation design and related analyses.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York : Wiley.

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

Francisco, C. A. and Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Design. *The Annals of Statistics* **19**, 454–469.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Pacific Grove, Calif.: Wadsworth & Brooks/Cole.

Hasselblad, V., Stead, A. G., and Galke, W. (1980). Analysis of Coarsely Grouped Data From the Lognormal Distribution. *Journal of the American Statistical Association* **75**, 771–778.

National Center for Health Statistics (1996). *NHANES III Reference Manuals and Reports, CD-ROM GPO 017-022-1358-4,* Washington, DC United States Government Printing Office.

Park, I. (1999). The Use of Sample Survey Data for Estimation of the Tails of Distribution Functions. Ph.D. dissertation, Department of Statistics, Texas A&M University.

Park, I. and Eltinge, J. L. (1999). Fitting Complex Survey Data to the Tail of A Parametric Distribution. *Proceedings of the Section on Survey Research Methods,* American Statistical Assocation, **76**, 599–604.

Pirkle, J. L., Brody, D. J., Gunter, E. W., Kramer, R. A., Paschal, D. C., Flegal, K. M., and Matte, T. D. (1994). The Decline in Blood Lead Levels in the United States. *Journal of the American Medical Association* **272**, 221–230.

Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990). On Estimating Distribution Functions from survey Data using Auxiliary Information. *Biometrika* **77**, 365–375.

Skinner, C. J. (1989). Introduction to Part A. In C. J. Skinner, D. Holt, and T. M. F. Smith (eds.), *Analysis of Complex Surveys*, pp. 23–58. New York: Wiley.

Tucker, H. G. (1998). *Mathematical Methods In Sample Surveys.* Singapore : World Scientific.