SOME ISSUES IN THE ANALYSIS OF COMPLEX SURVEY DATA

Susana Rubin Bleuer and Ioana Schiopu Kratina, Statistics Canada
Susana Rubin Bleuer, Statistics Canada, 3rd flr, R.H. Coats Bldg., Ottawa, K1A-0T6 (rubisus@statcan.ca)

Key words: Stochastic Model, Complex Survey Data, Superpopulation.

1. INTRODUCTION

We are concerned about inference on a parameter of a stochastic model with an estimator using data from a complex sample. Classical sampling theory concerns inferences for finite population parameters. Hàjek (1960), Krewski and Rao (1981), Binder (1983) and others, studied and obtained results on the asymptotic properties of the sample estimator under simple random sample and some complex designs.

On the other hand, Hartley and Silken (1975), Fuller (1975), Francisco and Fuller (1991) and others, studied the properties of the sample estimator with respect to a model parameter, some times called superpopulation parameter. They obtained asymptotic results for regression sample estimators using data from certain complex sampling designs.

Underlying their set ups, there was the notion of a "superpopulation" defined on a probability space $(\mathbf{\Omega},\boldsymbol{F},\boldsymbol{P})$ and the finite population was a considered a realization of it for an outcome $\omega \in \Omega$. The observed sample would be the second phase in a 2-phase design.

In our study, we represent the 2-phase 'sampling scheme' by means of a "product probability space" which includes both the designed sampling space and the superpopulation. That is, we formalize the space where the two "samples" live together. This allows us to develop methods of inference on a superpopulation parameter, based on data obtained from a wide variety of complex designs. Furthermore, this methodology can be viewed as a means of integrating different approaches to design-based inference and model-based inference with data from a complex sample design.

Indeed , suppose for example that we are interested in the parameter θ_0 defined as the solution of the stochastic model equation

$$E_m\{u(Y,\theta_0)\} = 0.$$

Let θ_N be the maximum likelihood estimator of θ_0 based on the finite population values, and let θ_n be a sample estimator of θ_N. Here N denotes the size of the finite population, if it is unclustered, or the number of clusters in the finite population, while n denotes the number of clusters in the first stage of the sample (which is the second phase sample in our set up). We can express the sample estimator around the model parameter as the sum of of two terms:

$$n^{1/2}(\theta_n - \theta_0) = n^{1/2}(\theta_n - \theta_N) + n^{1/2}(\theta_N - \theta_0) .$$

If we condition on the finite population, that is, if we hold the finite population fixed, then, under certain conditions, the first term is asymptotically normal in the design probability space, while the second term is a constant. If, on the other hand, we let the finite population vary according to the law of the superpopulation, the second term is asymptotically normal, under the usual regularity conditions. The asymptotic properties of these two terms are given in different spaces. The asymptotic limit of the sum makes sense if we think of each term as an entity of the product space defined in this paper.

The convergence of the sum does not follow in a trivial way. We prove asymptotic normality of the sample estimator under a set of minimum conditions and we present applications to inference on a distribution function and other superpopulation parameters. These results extend Fuller (1975) and (parts of) Francisco and Fuller (1991) to more complex estimators and more general designs.

We would like to remark that our result enables us to provide inference for $n/N \to f \geq 0$. It is important to allow the asymptotic sampling rate f to be positive, because even when the first stage sampling rate is small, the variance of the second term may not be negligible and should be accounted for. Korn and Graubard (1998) gave examples where the model variation is non-negligible for different designs and superpopulations.

A specific application is given to the estimation of $\theta_0 = F(x)$ under a two stage stratified sampling design. Examples of consistent variance estimators of the sample estimator and another application to survival data, can be found in Rubin Bleuer (1998).

2. FINITE POPULATIONS AND SAMPLING DESIGNS

Definition 2.1 A finite population U of size N consists of N units labeled $i = 1, 2, ..., N$. To each unit i in the finite population we associate a vector $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ where $\boldsymbol{y}_i$ represents the vector of characteristics of interest and $\boldsymbol{z}_i$ contains the prior information available at the time the survey design is chosen. All components of $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ are real-valued and we assume $\boldsymbol{y}_i \in \mathbb{R}^k$ and $\boldsymbol{z}_i \in \mathbb{R}^q$.

Definition 2.2 We define a sampling design as in Sarndal et al (1991). Let S be the collection of all samples s or "sets " of labels i from $U = \{1, ..., N\}$ that are possible to obtain with a specific sampling procedure. Note that the collection S of samples can include "ordered" samples with repeated units if the sampling scheme allows for replacement of units. These "ordered" samples are not proper subsets of the set of N labels $U = \{1, ..., N\}$, but they will be considered as bonafide elements of the set S. A sampling design on (U, S) is a function

$$p_d : S \times \mathbb{R}^{qN} \to [0,1]$$

such that

(1) $p_d(s, \cdot)$ is Borel-measurable in $\mathbb{R}^{qN}$, $\forall s \varepsilon S$

(2) $p_d(\cdot, z_1, ..., z_N)$ is a probability measure on S, $\forall z_i \varepsilon \mathbb{R}^q$,

Remark 2.1 For the sake of simplicity and without loss of generality, we label $qN = N$, and we shall often deal with scalars only. Similarly we set $k = 1$ for now.

3. SUPERPOPULATION

Definition 3.1 A superpopulation associated with a finite population U of size N associated with vectors $x_i = (y_i, z_i)$, $i = 1, 2, ..., N$, is a sequence of N random vectors $\{X_i\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$,

$$X_i = (Y_i, Z_i): \Omega \to \mathbb{R}^{k \times q},$$

such that for some $\omega_0 \varepsilon \Omega$, $X_i(\omega_0) = x_i$. We say that the $\{X_i\}$ generates the finite population U, or that U is a realization of the superpopulation given by ω_0. The $\{X_i\}$ are assumed stochastically independent, though not necessarily identically distributed.

Example3.1 The superpopulation is composed of L disjoint strata of clusters. Stratum sizes are considered known fixed constants. The h-stratum is composed of N_h clusters and $N = N_1 + N_2 + ... + N_L$. The size of cluster i in stratum h is M_{hi}. A two-stage model can be represented by the finite sequence of random vectors $\{ X_{11} ... X_{1N_1} ... X_{LI} ... X_{LN_L} \}$, where $X_{hi} = (Y_{hi}, M_{hi}, \mu_{hi}, \sigma^2_{hi})$.

The second stage model m_2 is given by setting $Y_{hi} = (Y_{hi1}, ..., Y_{hiM_{hi}})'$, which is composed of M_{hi} random values $(Y_{hij}) \sim$ independent, identically distributed random variables (i.i.d.r.v.) $F_{hi}(\mu_{hi}, \sigma^2_{hi})$. The values M_{hi} depend on the particular outcome ω of the superpopulation, but they are often known at the time of the design, and the vectors Y_{hi} can be observed if we do a census of the finite population. But the mean and variance of the (Y_{hij}) cannot usually be observed. The first stage model m_1 is defined by assuming that $(\mu_{hi}, \sigma^2_{hi})$ are i.i.d.r.v. with distribution function $F_h(\mu_h, \sigma^2_h, \Sigma^{\mu}_h)$. Here $\mu_h = E_{m_1}(\mu_{hi})$, $\sigma^2_h = E_{m_1}(\sigma^2_{hi})$ and $\Sigma^{\mu}_h = V_{m_1}(\mu_{hi})$. Note that while the cluster values (Y_{hij}) are stochastically independent given the second stage model, they are correlated when the overall model is taken into account.

Definition 3.2 Given the superpopulation (Y_i, Z_i), $i = 1, 2, ..., N$ let $Z = (Z_1, ..., Z_N)'$ denote the random vector from the superpopulation containing $N \times q$ elements. Let $\omega \varepsilon \Omega$ determine the finite population $U = U(\omega) = \{Y_1(\omega), ..., Y_N(\omega)\}$, as well as the prior information $Z(\omega) = (Z_1(\omega), ..., Z_N(\omega))' = (z_1, ..., z_N)'$. We write, for a sampling design p_d on the finite population U,

$$p_d(s, \omega) = p_d(s, Z(\omega)).$$

Since $p_d(s, \cdot)$ is Borel -measurable and Z is a random vector on Ω, for each $s \varepsilon S$ the mapping

$$p_d(s, \cdot): \Omega \to [0,1]$$

is a random variable on $(\Omega, \mathcal{F}, P)$.

4. THE PRODUCT SPACE

We wish to define a probability measure on a product space which will contain both the design and the superpopulation that generated the finite population. Let $X^N = (X_1 ... X_N)$ define a superpopulation associated with a finite population U_N and let p_d be a sampling design defined on (U_N, S_N). Recall that N is the number of stochastically independent elements in the superpopulation. Let

$$\Omega_N = S_N \times \Omega = \{(s, \omega) \,/\, s \varepsilon S, \omega \varepsilon \Omega\}$$

Definition 4.1 We define the probability measure $P_{N,d}$ by the expression

$$P_{N,d}(s, F) = \int_F p_d(s, \omega)\, dP(\omega) \qquad (4.1)$$

for every $s \varepsilon S_N$ and $F \varepsilon \mathcal{F}$. Since S_N is a finite set, the σ-algebra generated by S_N is the collection of finite unions of elements in S_N. By abuse of notation we shall write S_N to denote the the σ-algebra generated by S_N when there is no ambiguity. The integral is σ-additive and hence $P_{N,d}$ is a probability measure in the product space. Thus the triple

$$(\Omega_N, S_N \times \mathcal{F}, P_{N,d})$$

is a well defined probability space.

The next large sample result on the product space is the key to our development. We show that if a sequence of random variables converge weakly (in law) in the design probability space then it converges weakly in the product space. Let $\{U_N\}_{N=1}^{\infty}$ be a sequence of finite populations of size N, generated by superpopulations X^N defined on $(\Omega, \mathcal{F}, P)$, not necessarily nested. We assume that for every finite population U_N there is a sampling design p_d defined on (U_N, S_N) with expected (or fixed) sample size n such that $f_N = n/N$ converges to a fixed constant $f \geq 0$ as $N \to \infty$. When $f = 0$ we assume that $n \to \infty$ as $N \to \infty$. Let Ω_N be the associated product space.

Theorem 4.1 Let $T_{N,d}$ be a random variable defined on a design probability space (U_N, S_N, p_d) such that for every real number x, we have

$$\lim_{N\to\infty} p_d\{ s \varepsilon S_N / T_{N,d} \geq x\} = g(x). \qquad (4.2)$$

We assume that (4.2) holds $a.s.\ \omega$ in Ω and that the function g does not depend on ω. Then

$$\lim_{N\to\infty} P_{N,d}\ \{(s,\omega)\ \varepsilon \Omega_X / T_{N,d}(s,\omega) \geq x\} = g(x).$$

Corollary 4.1 Let θ_N be a finite population parameter defined on a finite population of size N, generated by a superpopulation X^N and let p_d define a sampling design on (U_N, S_N) as above. We have

1) If θ_n is a design-consistent estimator of θ_N based on a sample of size n, then θ_n is consistent in the product space, and

2) If for almost all $\omega \varepsilon \Omega$, as N, $n \to \infty$, the design-based distribution of

$$n^{1/2}(\theta_n - \theta_N) = n^{1/2}(\theta_n(s,\omega) - \theta_N(\omega))$$

is asymptotically normal with mean zero and fixed variance independent of ω, then the distribution of

$$n^{1/2}(\theta_n - \theta_N)$$

is asymptotically normal, in the product space $(\Omega_N, S_N \times \mathscr{F}, P_{N,d})$.

Remark 4.1 Krewski and Rao (1981) conditions for asymptotic normality of the sample estimator $\bar{y}$ of the finite population mean $\bar{Y}$, in the design probability space require that, for a realization of the finite population $U_N = \{Y_1(\omega), ..., Y_N(\omega)\}$ and prior information $\mathbf{Z}(\omega)$, these values satisfy certain design-moment properties. These properties translate into the convergence of sequences of numbers and have to be satisfied for almost every $\omega \varepsilon \Omega$ in order to yield convergence in the design space for every possible value of the finite population. For example, under SRSWOR from a finite population generated by n independent and identically distributed random variables (i.i.d.r.v.), $Y_1, ..., Y_N$, one of the Krewski- Rao conditions on $U_N(\omega)$, is that the design variance converge to a positive number $\psi(\omega)$ almost surely in ω:

$$\psi_N(\omega) = V_d(n^{1/2}\ \bar{y}(s,\omega)) \to \psi(\omega)\ a.s.\omega$$

where $\psi(\omega)$ is positive definite $a.s.\omega$. Now, simple conditions on the moments of the superpopulation will ensure this. Indeed, if the model expectation and variance exist and are finite ($E_m(Y_i) < \infty$ and $V_m(Y_i) < \infty$), then by the strong law of large numbers for i.i.d.r.v. we have:

$$\psi_N(\omega) = (1-n/N)\ N/(N-1)\ [\sum_i Y_i^2/N - \bar{Y}^2]$$

$$\to (1-f)\ V_m(Y_i)\ a.s.\omega.$$

Remark 4.2 When the prior information is correlated with the characteristic of interest under the superpopulation assumption, moment conditions for asymptotic convergence become more demanding.

5. ESTIMATION OF MODEL PARAMETERS

In analytical uses of sample surveys, the object is to estimate either a superpopulation model parameter or a finite population parameter whose form is motivated by such a model. Let $X_i = (Y_i, Z_i)$, i = 1, 2 ,..., N be a superpopulation defined on a probability space $(\Omega, \mathscr{F}, P)$. Most superpopulation or finite population parameters can be described by the distribution function $F_y(Y, \theta_0, \varphi)$ of the random vector Y defined on $(\Omega, \mathscr{F}, P)$, where both θ_0 and φ describe completely the distribution function F_y, φ is considered a nuisance parameter, and $\theta_0 \varepsilon \Theta$ is the parameter of interest. Here Θ is the parameter space. In the following we will assume that we know the nuisance parameter, and omit writing it. We assume that θ_0 can be estimated by an unbiased estimating function, that is, a function of the finite population vector and the parameter $u(y, \theta)$, such that

$$W(\theta) = E_m(u(y, \theta)) = 0 \qquad (5.1)$$

if $\theta = \theta_0$. Here E_m denotes expectation under the model m.

Definition 5.1 For a realization $\omega \varepsilon \Omega$ of the superpopulation, the finite population estimating equation W_N is a finite population total defined by

$$W_N = W_N(\omega,\theta) = \sum_{i \varepsilon U_N} u(Y_i(\omega), \theta)$$

for every $\theta \varepsilon \Theta$. The finite population parameter $\theta_N = \theta_N(\omega)$ is defined as the solution of the finite population estimating equation:

$$W_N(\omega, \theta_N(\omega)) = 0.$$

Definition 5.2 For every $\omega \varepsilon \Omega$ and sample $s \varepsilon S_N$, let W_n be a design-consistent estimator of the finite population estimating equation W_N. Thus:

$$W_n = W_n(s,\omega,\theta) = \sum_{i \varepsilon s} w_i(\mathbf{Z}(\omega))\ u(Y_i(\omega), \theta) \qquad (5.2)$$

Here the $w_i = w_i(\mathbf{Z}(\omega))$ are design weights which may depend on the prior information $\mathbf{Z}(\omega)$. The sample estimator of θ_N is $\theta_n = \theta_n(s,\omega)$, defined as the solution of the sample estimating equation:

$$W_n(s,\omega,\theta) = 0.$$

To make inferences about θ_0 by means of the sample parameter, we propose to find the asymptotic distribution of θ_n in the product space. We use Binder's (1983) result on the asymptotic distribution of θ_n in the design probability space and extend it to the product space.

We assume from now on that the function u is differentiable and we set the following notation: for $\omega\varepsilon\Omega$ and $s\varepsilon S_N$ we define the finite population "information matrix" J_N and the sample "information matrix" J_n by

$$J_N(\omega,\theta) = (1/N)(\partial W_N/\partial\theta)(\theta)$$

and respectively

$$J_n(s,\omega,\theta) = (1/N)(\partial W_n/\partial\theta)(\theta),$$

where W_N and W_n are the finite population total and respective sample estimator as defined above.

The next theorem require the structure of the product space so we spell out the set of conditions required for the development. We assume that there is a superpopulation defined on a probability space $(\Omega, \mathcal{F}, P)$ such that the finite population U_N is a realization of the superpopulation. Let (U_N, S_N, p_d) be a sampling space defined on U_N, and let $(\Omega_N, S_N \times \mathcal{F}_N, P_{N,d})$ be the product space generated by the above defined model and design. Here $\Omega_N = S_N \times \Omega$. Let $u(Y,\theta)$ be a real valued function of the data y with $\theta\varepsilon\Theta$ and let W_N, W_n, θ_N and θ_n be defined as above (Definitions (5.1) and (5.2)). We set $W(\theta) = E_m(u(Y,\theta))$ and we assume the following conditions on the superpopulation and the design (p_d)

(i) Model There exists $\theta_0\varepsilon\Theta$ such that

$$W(\theta_0) = E_m(u(Y,\theta_0)) = 0,$$

and $W(\theta)$ is a real valued differentiable function with continuous partial derivatives and $(\partial W/\partial\theta)(\theta_0)$ of full rank.

(ii) $E_m(u(Y,\theta)) < +\infty$ for every $\theta\varepsilon\Theta$. Under (iidrv), this implies the weak law of large numbers: $W_N(\theta)/N \rightarrow W(\theta)$ in P.

(iii) There exists a compact neighbourhood $K(\theta_0)$ of θ_0 on which the sample "information matrix" $J_n(s,\omega,\theta)$ is bounded in design probability, as $N \rightarrow \infty$, uniformly in θ. Similarly we require that the finite population "information matrix" $J_N(\omega,\theta)$ be bounded ($O_p(1)$ as $N \rightarrow \infty$, uniformly in θ, $O_p(1)$ refers to the probability measure P).

Conditions (i), (ii) and (iii) and (vii) below, are required for the consistency of θ_n and θ_N.

(iv) Condition J $J_N = J_N(\omega,\theta) \rightarrow J(\theta_0)$ in probability P as $N\rightarrow\infty$ and $\theta\rightarrow\theta_0$, where $J(\theta_0)$ is a positive definite matrix which is non-stochastic in $(\Omega, \mathcal{F}, P)$.

(v) Asymptotic Variance: Let $\Sigma_N(\theta_0) = (f/N)\sum_{i=1}^{N} V_m(u_i(Y), \theta_0))$. The finite population variance $\Sigma_N(\theta_0)$ converges to $\Sigma(\theta_0)$ in probability P as $N\rightarrow\infty$, where $\Sigma(\theta_0)$ is a positive definite matrix.

(vi) Liapunov Condition: There exists some $\gamma > 2$ such that as $N\rightarrow\infty$ we have

$$\sum_{i=1}^{N} E_m \mid u_i(\theta_0) \mid^{\gamma} = o[N\,\Sigma_N]^{\gamma/2}.$$

Conditions (i) to (vi), ensure the asymptotic normality of θ_N, the solution of the finite population estimating equation $W_N(\theta) = 0$.

(vii) We assume the necessary conditions for the Central Limit Theorem to hold for $X_N(\theta_0) = n^{1/2}\{W_n(\theta_0) - W_N(\theta_0)\}/N$ in the design probability space. Recall that moment conditions in the superpopulation will imply some of the necessary conditions for specific designs (see, for example, Remark 4.1). For some designs, equicontinuity or equiboundedness of the function u is required (see Rubin-Bleuer (2000)).

Theorem 5.1 Asymptotic normality of θ_n If $f = \lim\, inf\, n/N \geq 0$ and conditions (i) to (vii) hold then

$$n^{1/2}(\theta_n - \theta_0) \qquad (5.4)$$

converges to a normal random variable with mean zero and variance Γ in the product space $(\Omega_N, S_N \times \mathcal{F}_N, P_{N,d})$. The variance Γ is the sum of two terms, the variance due to the design and the variance due to the model:

$$J(\theta_0)^{-1}\{\varphi(\theta_0) + f\,\Sigma(\theta_0)\}\,J(\theta_0)^{-1} \qquad (5.5)$$

Here $\varphi(\theta_0)$ represents the limiting variance of $W_n(\theta_N)$, $\Sigma_N(\theta_0)$ is the asymptotic variance defined in (v) and $J = J(\theta_0)$ is the asymptotic limit of the "information matrix", defined in (iv). We use the following lemma for the proof of Theorem 5.1.

Lemma 5.1 Asymptotic Independence. Let T_N be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ such that they converge in law to a random variable T with distribution function $F_T(t)$, for $-\infty < t < +\infty$. Let (U_N, S_N, p_d) be a sampling space defined on a realization U_N of a superpopulation, and let $(\Omega_N, S_N \times \mathcal{F}_N, P_{N,d})$ be the product space generated by the superpopulation and sampling design. Let R_N be a sequence of random variables (vectors) defined in the product space such that for almost every ω, the conditional distribution of R_N given ω converges to a distribution function $G_R(r)$, for $-\infty < r < +\infty$, with parameters independent of ω. That is, we assume that for $-\infty < r < +\infty$, as N increases to infinity, we have

$$G_{R_N}(r,\omega) = p_d\{s \varepsilon S_N / R_N(s,\omega) \le r\} \to G_R(r)$$

$a.s.\omega$ in Ω. Then, the joint distribution function of (R_N, T_N) converges to the product of the two distribution functions $G_R(r) \cdot F_T(t)$, and the random variables R_N and T_N are said to be "asymptotically independent".

For the proof of theorem 5.1, we set $R_N(s,\omega) = n^{1/2}(\theta_n - \theta_N)$ and

$T_N(\omega) = N^{1/2}(\theta_N - \theta_0)$. We note that simple moment conditions in the superpopulation needed for the convergence of T_N will ensure that the parameters of $G_R(r)$ do not depend on ω. Conditions (i) to (vi) yield the asymptotic normality of T_N in $(\Omega, \mathscr{F}, P)$. R_N is asymptotically normal in the product space by condition (vii) and Corollary 4.1. Theorem 5.1 follows from Lemma 5.1.

6. EMPIRICAL DISTRIBUTION FUNCTION

Let the superpopulation be composed of an infinite number of disjoint strata $h = 1, 2, \ldots$ and let the finite population consist of L of these strata with N_h independent clusters in each stratum. The N_h are non-stochastic. There are $N = \sum_{h=1}^{L} N_h$ clusters (primary sampling units) in the finite population. Let M be the number of ultimate units in the finite population U_N and let M_{hi} be the number of ultimate units in cluster i of stratum h. We will consider the M_{hi} and M known at the time of the design. Thus, for our purpose M_{hi} and M are fixed quantities. The characteristic of interest is given by the random vector Y_{hij}, $i = 1, \ldots, M_{hi}$, $i = 1, \ldots, N_h$ and $h = 1, \ldots, L$. Hence from now on we consider the conditional distributions of Y_{hij} given the cluster sizes M_{hi}, and E_m and V_m will denote the expectation and variance with respect to $(\Omega, \mathscr{F}, P)$ given the cluster sizes M_{hi}. We assume that a common overall superpopulation distribution function exists for the characteristic of interest Y_{hij}:

$$F(x) = P(Y_{hij} \le x) \qquad (6.1)$$

for $h = 1, \ldots, L$, $i = 1, \ldots, N_h$ and $j = 1, \ldots, M_{hi}$. Let us assume, for the sake of simplicity, that x is a scalar. The superpopulation distribution function $F(x)$ will refer to the distribution conditional to the cluster sizes M_{hi}. Since the Y_{hij} are independently and identically distributed given M_{hi}, and the M_{hi} are considered fixed, there is no need to conceive a clustered superpopulation. But often operational constraints dictate a stratified 2-stage design, and hence it is convenient for us to group the superpopulation in the sampling design clusters. We define the finite population distribution function by

$$F_N(x) = (1/M) \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} I\{Y_{hij} \le x\}$$

where the indicator function I is defined as $I\{Y_{hij} \le x\} = 1$ if $Y_{hij} \le x$, and 0 otherwise. Under the model, the finite population distribution function is unbiased for F.

Now let us consider a stratified two-stage design in which the clusters ($p.s.u$) are selected with replacement and in which independent subsamples are taken within those *psu's* selected more than once. Suppose $n_h \ge 2$ psu's are selected from the N_h psu's in the h-th stratum with probabilities $p_{hi} > 0$, $i = 1, 2, \ldots, N_h$ and $h = 1, \ldots, L$, where $\sum_i p_{hi} = 1$. Let $n = \sum_{h=1}^{L} n_h$.

Let $\bar{G}_N(x) = (1/M) \sum_{hi\varepsilon s} G_{hi}(x)$, where $G_{hi} = \sum_{j=1}^{M_{hi}} I(Y_{hij} \le x)$. Let $\hat{G}_{hi}(x)$ and $\hat{M}_{hi}$ be a design-unbiased estimators of the respective cluster totals $G_{hi}(x)$ and M_{hi}, based on sampling at the second stage. The second stage sampling rate is $f_{hi} = m_{hi} / M_{hi}$. We consider the sample estimator of the finite population distribution given by the ratio of two sample means:

$$F_n(x) = \hat{\bar{G}}_N(x) / \hat{\bar{M}}_N$$

where

$$\hat{\bar{G}}_N(x) = (1/M) \sum_{hi\varepsilon s} \hat{G}_{hi} / (n_h p_{hi})$$

And similarly,

$$\hat{\bar{M}}_N = \hat{M}/M = (1/M) \sum_{hi\varepsilon s} \hat{M}_{hi} / (n_h p_{hi}).$$

Let V_d denote the design variance. The asymptotic properties of $F_n(x)$ will be examined under the following conditions. We require fewer conditions for the asymptotic normality of the finite population distribution than for the finite population parameter of theorem 5.1, because the former is a sample mean (from the first phase).

C1.The population mean per cluster is M/N. We assume that as $N \to 0$, $M/N \to m > 0$.
C2.$f = n/N$ remains constant as $N \to \infty$.

C3. $V_d(n^{1/2} \hat{\bar{A}}_N(x)) = (1/M^2) \sum_{h=1}^{L} (n/n_h)\, b_n$,

$$b_n = \sum_{i=1}^{N_h} (V_{hi} + A_{hi}^2(x))/p_{hi} - (\sum_{i=1}^{N_h} A_{hi}(x))^2$$

converges in probability P to a positive definite matrix Γ (non stochastic in Ω) as $N \to \infty$, where $A_{hi}(x) = G_{hi}(x) - F(x)\, M_{hi}$ and

$$\hat{\bar{A}}_N(x) = \hat{\bar{G}}_N(x) - F(x)\, \hat{\bar{M}}_N(x).$$

C4. There exists $\delta > 0$ such that as $N \to \infty$
$$(1/M) \sum_{hi} E_m(|G_{hi}(x) - F(x)\, M_{hi}|^{2+\delta}) = O(1).$$
C5.We assume the necessary conditions for the Central Limit Theorem to hold for $\hat{\bar{M}}_N$ and $\hat{\bar{G}}_N(x)$ in the design probability space. (See conditions 1-4 from Krewski and Rao (1981).

Theorem 6.1 Under the model (6.1) and Conditions C1-C5 , for $-\infty < x < +\infty$, we have that
$$n^{1/2}(F_n(x) - F(x)) \qquad (6.2)$$
converges in law, in the product space, to a normal random variable with mean zero and variance $\varphi + \Sigma$ where the structure of φ depends on the design and $\Sigma = (f/m)\, F(x)(1 - F(x))$.

Corollary 6.1 : A consistent estimator of the variance of (6.2) (in the product space) is given by
$$Variance = (g_n + f/m) * F_n(x)(1 - F_n(x)),$$
$$g_n = \sum_h W_h^2 (n/n_h)(1/M_h^2) \{ \sum_1^{N_h} M_{hi}/f_{hi}p_{hi}) - 1 \}.$$

REFERENCES

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Reviews*, **51**, 279-292.
Chow, Y.S. and Teicher, H. (1988). *Probability Theory*, second edition, Springer-Verlag, New York
Francisco, C.A. and Fuller, W.A. (1991). Quantile estimation with a complex survey design, *Annals of Statistics*, **19**, 454-469.
Fuller, W.A. (1975). Regression analysis for sample surveys. Sankhyã **37**, 117-132.
Godambe, V.P and Thompson, M.E. (1986). Parameters of superpopulations and survey population: their relationship and estimation, *International Statistical Review*, **94**, 127-137.
Hájek, J. (1963). Limiting distributions in simple random sampling from finite populations, *Publ. Math.Inst. Hungarian Acad. Sci.*, **5**, 361-374.
Hartley, H.O. and Sielken, R.L. (1975). A "super-population viewpoint" for finite population sampling, *Biometrics*,31,411-422.
Korn, E. L. and Graubard, B.I. (1998). Variance estimation for superpopulation parameters, *Statistica seneca*, **8**, 1131-1151.
Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: Properties of linearization, jackknife and balanced repeated replication methods, *Annals of Statistics* **9**, 1010-1019.
Rubin Bleuer, S. (1998). Inference for parameters of the superpopulation, Part 1, *Research Sabbatical Report, Internal report*, Statistics Canada.
Rubin Bleuer, S. (2000). Some issues in the analysis of complex survey data. *Statistics Canada Series, Methodology Branch, Business Survey Methods Division*, BSMD- 20-001 E.
Särndal, C-E, Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
Şchiopu-Kratina, Ioana (1999). Variance Estimation under Marginal Models for Longitudinal Data Analysis using Complex Survey Data, *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, November 15-17, 1999, Washington, D.C.
Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction With Applications*, Chapman and Hall, New York.
Skinner, C.J., Hold, D. And Smith, T.M.F. (1989). Analysis of Complex Surveys, John Wiley & Sons, Chichester, Wiley, New York, Brisbane, Toronto, Singapore.