東海管理評論
民國八十九年十一月，第二卷，第一期，47-62

Small Area Estimation Employing Poststratification

*Pao-Sheng Shen**

Abstract

A compromise estimator employing **poststratification** is proposed for **small area estimation**. The estimator strikes a balance to deal with the assumption of similarity within a poststratum and small number of observations in subareas constructed by poststratification. Small sample properties are derived for the estimator from simple random samples. By assuming that the probability that poststratum size equal to zero is negiligibly small, the estimator can be justified in **Bayesian** terms based on an inherent superpopulation model. The generalization of poststratification from the simple random sampling to complex design is discussed.

Keywords : Small Area Estimation, Poststratification, Bayesian Analysis.

1. Introduction

Small area estimation has received considerable attention in recent years because of a growing demand for reliable small area statistics. Sample surveys rarely produce enough data to permit accurate estimation of these small areas by using the standard methods based on the selection probabilities. Therefore, alternative estimators that borrow strength from other related small areas have been proposed in literature to improve efficiency. An example concerning small area estimation was given by Ghosh and Meeden (1986) under a normal superpopulation model. Ghosh and Lahiri (1987) proposed robust empirical Bayes estimation of a vector of stratum means under the assumption that the posterior expectation of any stratum mean is a linear function of sample observations. Cressie (1989) used an empirical Bayes approach to correct undercount in U. S. decennial censuses by modeling the subareas within a poststratum to have a common mean and variances inversely proportional to their census counts.

* Department of Statistics, Tunghai University.

Although the superpopulation model approach to small area estimation provides a new avenue for exploration of this problem, the model-dependent estimators may be seriously design biased. Purcell and Kish (1980) warned against the mistake of considering small area estimation as one homogeneous problem and suggested that area size is a key factor in the choice between design-consistent and model-dependent estimators. Hence, this article is aimed at finding an estimator which provides a good compromise for dealing with the assumption of homogeneity within a poststratum and the unacceptably large standard error due to small number of observations in subareas. A class of estimator, which are generalizations of the synthetic estimator (SYN)(Gonzalez and Hoza, 1978), will be proposed. The design-based inferences are developed for the estimators from simple random samples. These include the derivation of their biases and variances. It will be shown that estimators will have both within poststratum and between poststratum components of variance. Based on a random effects superpopulation model, an approximate Bayes rule can be obtained from the class of estimators with respect to a general quadratic loss function within the class of linear combinations of a given set of functions on the sample space. The approximate Bayes rule strikes a balance between SYN which pools information across areas and the classical unbiased estimator (UNB) which depends only on data. It has a much smaller bias than the SYN estimators when the latter is badly biased. In addition, it has the advantage of a considerably reduced variance compared to UNB. These results are supported by a Monte Carlo study. In Section 2 the proposed class of estimators are introduced and small sample properties are derived for the estimators from simple random samples. In Section 3 the estimators are studied through a Bayesian approach. In Section 4 the properties of the estimators as well as SYN and UNB are studied through a Monte Carlo simulation. Finally, Section 5 provides some concluding remarks and discussion of generalization from the simple random sampling to complex design.

2. Estimators

We suppose the finite population of size N is divided into I mutually exclusive areas, labelled $i=1,\ldots,I$. An estimate is required for the total (or mean) of the variable of interest for each area, In practice, the number of areas, I, is often quite large. We further assume that, within each area, units are poststratified into L subareas ; labelled $h=1,\ldots,L$. The subarea sizes N_{hi} resulting from this

cross-classification are assumed to be known. The number of poststrata, L, is usually modest ; in any case, L is assumed small compared to I. Let Y_{hij} ($j = 1,\ldots,N_{hi}$) be the measurement on the j^{th} individual in the hi^{th} subarea and let $\bar{Y}_{\cdot i} = \sum_{h=1}^{L} W_{hi}\bar{Y}_{hi}$ denote the mean for the i^{th} area, where $W_{hi} = N_{hi}/N_{\cdot i}$ and $\bar{Y}_{hi} = \sum_{j=1}^{N_{hi}} Y_{hij}/N_{hi}$, where $N_{\cdot i} = \sum_{h=1}^{L} N_{hi}$. The primary focus is to estimate $\bar{Y}_{\cdot i}$'s. We will consider the case of a simple random sample of size n drawn from the population. Define n_{hi} as the number of units measured in the hi^{th} subarea and $n_{h\cdot} = \sum_{i=1}^{I} n_{hi}$ as the number of units which happen to fall into the h^{th} poststratum.

Define the "indicator variables"

$$\alpha_{hi}(x_{hi}) = \begin{cases} 1, & \text{if } n_{hi} \geq x_{hi}, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$x_{hi} \in S, \quad S = \{\, 0,1,2,\ldots,\min(n, N_{hi}) \,\}$$

and

$$\gamma_h = \begin{cases} 1, & \text{if } n_{hi} \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

We consider a class of estimators of $\bar{Y}_{\cdot i}$ of the form

$$\tilde{\bar{y}}_{\cdot i}(\underline{x}_i) = \frac{\sum_{h=1}^{L} W_{hi}\left[\alpha_{hi}(x_{hi})\bar{y}_{hi} + \beta_{hi}(x_{hi})\bar{y}_{h\cdot}\right]}{\sum_{h=1}^{L} W_{hi}\gamma_h} \qquad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

where

$\underline{x}_i = (x_1, x_2, \ldots, x_{Li})^T$, $\beta_{hi}(x_{hi}) = \gamma_h - \alpha_{hi}(x_{hi})$, $\bar{y}_{hi}$ and $\bar{y}_{h\cdot}$ are the ordinary sample means of the units falling into the hi^{th} subarea and the h^{th} poststratum respectively provided subarea size $n_{hi} \geq 1$ or poststratum size $n_{h\cdot} \geq 1$. The definition of $\bar{y}_{hi}$ or $\bar{y}_{h\cdot}$ for the case $n_{hi} = 0$ or $n_{h\cdot} = 0$ is arbitrary.

The variances of the estimators (1) are unaffected by translation of y values. Now, to derive order of magnitude of their biases and variances, let $E_{2d}(\cdot)$ denote the conditional expectation given n_{hi}'s and $E_{1d}(\cdot)$ denote the expectation over n_{hi}'s. Similarly, we define variances $V_{1d}(\cdot)$, $V_{2d}(\cdot)$ and covariances $Cov_{1d}(\cdot)$, $Cov_{2d}(\cdot)$. First, we write $\tilde{\bar{y}}_{\cdot i}(\underline{x}_i)$ in a simple form as

$$\tilde{y}_{\cdot i}(\underline{x}_i) = \sum_{h=1}^{L} W_{hi}\tilde{y}_{hi}(x_{hi})$$

where

$$\tilde{y}_{hi}(x_{hi}) = \frac{\left(\frac{1}{2}+\varepsilon_{hi}(x_{hi})\right)\bar{y}_{hi} + \left(\frac{1}{2}+\delta_{hi}(x_{hi})\right)\bar{y}_{h\cdot}}{1+\bar{\theta}_i},$$

where

$$\varepsilon_{hi}(x_{hi}) = \frac{\alpha_{hi}(x_{hi}) - \frac{1}{2}E_d(\bar{\gamma}_i)}{E_d(\bar{\gamma}_i)},\quad \delta_{hi}(x_{hi}) = \frac{\beta_{hi}(x_{hi}) - \frac{1}{2}E_d(\bar{\gamma}_i)}{E_d(\bar{\gamma}_i)}$$

$$\bar{\gamma}_i = \sum_{h=1}^{L} W_{hi}\gamma_h,\quad \bar{\theta}_i = \sum_{h=1}^{L} W_{hi}\theta_{hi},\quad \theta_{hi} = \frac{\gamma_h - E_d(\bar{\gamma}_i)}{E_d(\bar{\gamma}_i)},$$

$$E_d(\bar{\gamma}_i) = \sum_{h=1}^{L} W_{hi}E_d(\gamma_h),\ \text{and}\ \ E_d(\bar{\gamma}_h) = 1 - \frac{\binom{N-N_{h\cdot}}{N}}{\binom{N}{n}}$$

Suppose that $|\bar{\theta}_i| < 1$, we are able to expand $\tilde{y}_{hi}(x_{hi})$ in the form

$$\tilde{y}_{hi}(x_{hi}) = \left[\left(\frac{1}{2}+\varepsilon_{hi}(x_{hi})\right)\bar{y}_{hi} + \left(\frac{1}{2}+\delta_{hi}(x_{hi})\right)\bar{y}_{h\cdot}\right]\left[1 - \bar{\theta}_i + (\bar{\theta}_i)^2 - (\bar{\theta}_i)^3 + \ldots\right]$$

from which we obtain the bias of $\tilde{y}_{\cdot h}(\underline{x}_i)$, denoted as $B_d[\tilde{y}_{\cdot h}(\underline{x}_i)]$,

$$\begin{aligned} B_d(\tilde{y}_{hi}(x_{hi})) = {} & \sum_{h=1}^{L} W_{hi}\bar{Y}_{hi}\left\{E_{1d}\left[(\bar{\theta}_i)^2 - \bar{\theta}_i\theta_{hi}\right] + \ldots + (-1)^{m-1}E_{1d}\left[(\bar{\theta}_i)^{m-1}\theta_{hi} - (\bar{\theta}_i)^m\right] + \ldots\right\} \\ & + \sum_{h=1}^{L} W_{hi}(\bar{Y}_{h\cdot} - \bar{Y}_{hi})\left\{\frac{1}{2} + E_{1d}[\delta_{hi}(x_{hi})] + E_{1d}\left[\frac{(\bar{\theta}_2)^2}{2} - \bar{\theta}\delta_{hi}(x_{hi})\right] + \ldots\right. \\ & \left. + (-1)^{m-1}E_{1d}\left[(\bar{\theta}_i)^{m-1}\delta\theta_{hi}(x_{hi}) - \frac{(\bar{\theta}_i)^m}{2}\right] + \ldots\right\} \end{aligned}$$

where $\bar{y}_{h\cdot}$ denotes the population mean for the h^{th} poststratum. Under the

assumption that the poststratum sizes are sufficiently large for approximating the hypergeometric distribution of n_h by a binomial, we can observe that, for any positive integer m,

$$\theta_{hi}^{m} = (-1)^m (1-\gamma_h) + (\frac{\overline{Q}_i}{1-\overline{Q}_i})^m \gamma_h \ ,$$

where

$$\overline{Q}_i = \sum_{h=1}^{L} W_{hi} Q_h^n \ , \ Q_h = 1 - \frac{N_h}{N} \ .$$

Similarly,

$$\frac{1}{2} + E_{1d}[\delta_{hi}(x_{hi})] = \frac{B_{hi}(x_{hi})}{1-\overline{Q}_i},$$

where

$$B_{hi}(x_{hi}) = 1 - Q_h^n - A_{hi}(x_{hi}),$$

$$A_{hi}(x_{hi}) = \sum_{k=x_{hi}}^{\min(N_{hi},n)} \binom{n}{k} P_{hi}^k (1-P_{hi})^{n-k},$$

where $P_{hi} = N_{hi}/N$. Hence, it is not difficult to see that in case of boundedness of $\overline{Y}_{hi}$ for all h, i.

$$B_d(\tilde{y}_{\cdot i}(\underline{x}_i)) \sim O\left(\max_h \{Q_h^n\}\right) + O\left(\max_h \left\{ |\overline{Y}_{h\cdot} - \overline{Y}_{hi}| \ (1 - A_{hi}(x_{hi})) \right\}\right),$$

Next, we derive the variances of the class of the estimators **(1)**. There are two components of variances from the well-known relation

$$V_d(\tilde{y}_{\cdot i}(\underline{x}_i)) = V_{1d}(E_{2d}[\tilde{y}_{\cdot i}(\underline{x}_i)]) + E_{1d}(V_{2d}[\tilde{y}_{\cdot i}(\underline{x}_i)]),$$

where the term $V_{1d}(E_{2d}[\tilde{y}_{\cdot i}(\underline{x}_i)])$ and $E_{1d}(V_{2d}[\tilde{y}_{\cdot i}(\underline{x}_i)])$ are called the between poststrata and the within poststrata components of variance of $\tilde{y}_{\cdot i}(\underline{x}_i)$ and denoted as $V_{dB}[\tilde{y}_{\cdot i}(\underline{x}_i)]$ and $V_{dW}[\tilde{y}_{\cdot i}(\underline{x}_i)]$, respectively. To derive $V_{dB}[\tilde{y}_{\cdot i}(\underline{x}_i)]$, we require the following results. We can observe that, for any positive integer m,

$$\delta_{hi}^{m}(x_{hi})=(\frac{-1}{2})^{m}(1-\gamma_h)+\left[\frac{1+\bar{Q}_i}{2(1-\bar{Q}_i)}\right]^{m}\gamma_h(1-\alpha_{hi2d}(x_{hi})),$$

Hence, it is not difficult to see that

$$Cov_{1d}(\delta_{hi}^{m}(x_{hi}),\delta_{h'i}^{m}(x_{h'i}))\sim O\left(\max_h\{A_{hi}(x_{hi})(1-A_{hi}(x_{hi}))\}\right),$$

$$Cov_{1d}(\theta_{hi},\theta_{h'i})\sim O\left(\max_h\{Q_h^n\}\right).$$

We find a first approximation to $V_{dB}(\bar{y}_{\cdot i}(\underline{x}_i))$ by omitting the terms in $\bar{\theta}_i,\theta_{hi}$ and $\delta_{hi}(x_{hi})$ with degree higher than 2 in the expansion of

$$\begin{aligned}V_{dB}(\bar{y}_{\cdot i}(\underline{x}_i))\cong&\sum_{h=1}^{L}W_{hi}^2\left\{\bar{Y}_{hi}^2V_{1d}(\theta_h-\bar{\theta}_i)+(\bar{Y}_{h\cdot}-\bar{Y}_{hi})^2V_{1d}(\delta_{hi}(x_{hi})-\bar{\theta}_i/2\right\}\\&+\sum_{h\neq h'=1}^{L}W_{hi}W_{h'i}\{\bar{Y}_{hi}(\bar{Y}_{h'\cdot}-\bar{Y}_{h'i})Cov_{1d}((\theta_{hi}-\bar{\theta}_i),(\delta_{h'i}(x_{h'i})-\bar{\theta}_i/2)\\&+\bar{Y}_{hi}\bar{Y}_{h'i}Cov_{1d}((\theta_{hi}-\bar{\theta}_i),(\theta_{h'i}-\bar{\theta}_i))\\&+(\bar{Y}_{h\cdot}-\bar{Y}_{hi})(\bar{Y}_{h'\cdot}-\bar{Y}_{h'i})Cov_{1d}((\delta_{hi}(x_{hi})-\bar{\theta}_i/2),(\delta_{h'i}(x_{h'i})-\bar{\theta}_i/2))\}.\end{aligned}$$

It follows that

$$V_{db}[\bar{y}_{\cdot i}(\underline{x}_i)]\sim O\left(\max_h\{Q_h^n\}\right)+O\left(\max_h\{(\bar{Y}_{hi}-\bar{Y}_{h\cdot})^2A_{hi}(x_{hi})(1-A_{hi}(x_{hi}))\}\right).$$

The within poststrata component $V_{dW}[\bar{y}_{\cdot i}(\underline{x}_i)]$ is given by

$$\begin{aligned}\sum_{h=1}^{L}W_{hi}^2\{&A_{hi}(x_{hi})S_{hi}^2\left[E_{1d}(n_{hi}^{-1}|n_{hi}\geq x_{hi})-N_{h\cdot}^{-1}\right]\\&+B_{hi}(x_{hi})S_h^2\left[E_{1d}(n_{h\cdot}^{-1}|n_{h\cdot}\geq 1)-N_{h\cdot}^{-1}\right]\}\end{aligned}$$

where $E_{1d}(.|n_{hi}\geq x_{hi})$ stands for the conditional expectation given $n_{hi}\geq x_{hi}$, $E_{1d}(.|n_{h\cdot}\geq 1)$ stands for the conditional expectation given $n_{h\cdot}\geq 1$, S_{hi}^2 and $S_{h\cdot}^2$ denote the variance of Y in the hi^{th} subarea and h^{th} poststratum, respectively. When n is sufficiently large, $E_{1d}(n_{hi}^{-1}|n_{hi}\geq x_{hi})$ and $E_{1d}(n_{h\cdot}^{-1}|n_{h\cdot}\geq 1)$ can be approximated to terms of order n^{-2} using

$$\left((n-x_{hi})P_{hi}+x_{hi}\right)^{-1}+(n-x_{hi})P_{hi}(1-P_{hi})/\left[(n-x_{hi})P_{hi}+x_{hi}\right]^3$$

and

$$\left((n-1)P_{h\cdot}+1\right)^{-1}+(n-1)P_{h\cdot}(1-P_{h\cdot})/\left[(n-1)P_{h\cdot}+1\right]^3,$$

respectively, where $P_{h\cdot}=N_{h\cdot}/N$ (Stephan, 1945).

3. A Bayesian Analysis of the Estimators

In this Section, we investigate the property of the class of estimators (1) using a Bayesian Approach. We assume that there is a superpopulation distribution from which all the population values are derived. Consider the variance component model as follows :

$$\begin{aligned} &Y_{hij}=\mu_h+\phi_{hi}+\varphi_{hij} \qquad \cdots\cdots (2)\\ &E_m(\phi_{hi})=0,\ V_m(\phi_{hi})=\sigma_h^2, \qquad h=1,\ldots,L,\\ &E_m(\varphi_{hij})=0,\ V_m(\varphi_{hij})=\tau_h^2, \qquad h=1,\ldots,L, \end{aligned}$$

where $E_m(\cdot)$ and $V_m(\cdot)$ denote model-based expectation and variance respectively. All distributions for the model are assumed independent.

The parameter ϕ_{hi} is the random effect associated with the i^{th} area within the h^{th} poststratum, and φ_{hij} is the random effect associated with the j^{th} individual within the hi^{th} subarea. Since σ_h^2 is in general much larger than τ_h^2/N_{hi}, conditional on $\underline{\mu}=(\mu_1,\mu_2,\ldots,\mu_L)$ and $\underline{\sigma}=(\sigma_1,\sigma_2,\ldots,\sigma_L)$, $\bar{Y}_{h1}$, $\bar{Y}_{h2}$, $\cdots$, $\bar{Y}_{hI}$ are independently distributed for all $h=1,\ldots,L$ with $E_m(\bar{Y}_{hi}|\mu_h)=\mu_h$ and $V_m(\bar{Y}_{hi}|\sigma_h)\cong\sigma_h^2$. Similarly, $E_m(S_{hi}^2|\tau_h)\cong\tau_h^2$. We now state a result that justifies the use of the class of estimators (1). First, the loss function considered is taken to be of the form

$$\sum_{i=1}^{I} f(N_{\cdot i})(\bar{y}_{\cdot i}^{est}-\bar{Y}_{\cdot i})^2 \qquad \cdots\cdots (3)$$

where $f(N_{\cdot i})$ is any positive function $N_{\cdot i}$, the i^{th} area size and $\bar{y}_{\cdot i}^{est}$ is some

arbitrary estimate of $\bar{Y}_{\cdot i}$.

LEMMA

Under the model (2) and assuming that there is vanishingly small probability that poststratum size is zero, there exists $\underline{x}_i^*$'s such that $\bar{y}_{\cdot i}(\underline{x}_i^*)$, $i=1,\ldots,I$, is an approximate Bayes rule with respect to the loss function (3) within the class of decision rules $\hat{\bar{y}}_{\cdot i}(\underline{x}_i)$, $i=1,\ldots,I$, with the form

$$\hat{\bar{y}}_{\cdot i}(\underline{x}_i)=\sum_{h=1}^{L}W_{hi}\sum_{j=1}^{3}\lambda_{hij}(x_{hi})g_{hij}(x_{hi}),$$

where

$g_{hi1}(x_{hi})=\alpha_{hi}(x_{hi})\bar{y}_{hi}$,

$g_{hi2}(x_{hi})=\alpha_{hi}(x_{hi})\bar{y}_{h\cdot}$, and $g_{hi3}(x_{hi})=(1-\alpha_{hi}(x_{hi})\bar{y}_{h\cdot}$;

$\lambda_{hi1}(x_{hi})$, $\lambda_{hi2}(x_{hi})$ and $\lambda_{hi3}(x_{hi})$ are real-valued functions of x_{hi} and satisfy

$$E_mE_d\left[\sum_{j=1}^{3}\lambda_{hij}(x_{hi})g_{hij}(x_{hi})\right]=\mu_h,\quad h=1,\ldots,L.$$

Proof. We wish to find $\underline{\lambda}_{hi}(x_{hi})=(\lambda_{hi1}(x_{hi}),\lambda_{hi2}(x_{hi}),\lambda_{hi3}(x_{hi}))^T$ $(h=1,\ldots,L$; $i=1,\ldots,I)$ to minimize the risk

$$E_mE_d\left[\sum_{i=1}^{I}f(N_{\cdot i})\left(\hat{\bar{y}}_{\cdot i}(\underline{x}_i)-\bar{Y}_{\cdot i}\right)^2\right]$$

$$=\sum_{i=1}^{I}f(N_{\cdot i})\left\{E_m\left[V_d\left(\hat{\bar{y}}_{\cdot i}(\underline{x}_i)\right)\right]+E_m\left[\left(B_d\left(\hat{\bar{y}}_{\cdot i}(\underline{x}_i)\right)\right)^2\right]\right\} \quad \cdots\cdots (4)$$

Under the superpopulation model (2), this is equivalent to finding x_{hi} to minimize the following risk separately for each h and i :

$$E_mE_d\sum_{j=1}^{3}\left(\lambda_{hij}(x_{hi})g_{hij}(x_{hi})-\bar{Y}_{hi}\right)^2.$$

When I is sufficiently large and x_{hi} is given, we obtain by Theorem 2.1 of Goldstein (1975) the approximate Bayes $\hat{\bar{y}}_{\cdot i}(x_{hi})$ as follows :

$$\lambda_{hi1}(x_{hi})=\frac{\sigma_h^2}{\sigma_h^2+q(x_{hi})\tau_h^2},\quad \lambda_{hi2}(x_{hi})=1-\lambda_{hi1}(x_{hi}),\quad \lambda_{hi3}(x_{hi})=1,$$

where

$$q(x_{hi})=\left[(n-x_{hi})P_{hi}+x_{hi}\right]^{-1}+\frac{(n-x_{hi})P_{hi}(1-P_{hi})}{\left[(n-x_{hi})P_{hi}+x_{hi}\right]^3}\ .$$

The associated Bayes risk can be approximated by $R_{hi}(x_{hi})$,

$$\begin{aligned}R_{hi}(x_{hi})=\sigma_h^2\left[1-2\lambda_{hi1}(x_{hi})A_{hi}(x_{hi})+2\lambda_{hi1}^2(x_{hi})A_{hi}(x_{hi})\right]\\ +\tau_h^2\lambda_{hi1}^2(x_{hi})A_{hi}(x_{hi})\end{aligned}$$

Hence, we can obtain the approximate Bayes rule of $\hat{\bar{y}}_{\cdot i}(\underline{x}_i)$, $i=1,\ldots,I$, by finding $\underline{x}_i^{\circ}=(\underline{x}_1^{\circ},\underline{x}_2^{\circ},\ldots,\underline{x}_{Li}^{\circ})^T$, $h=1,\ldots,L$ such that

$$R_{hi}(x_{hi}^{\circ})=\min_{x_{hi}\in S}\{R_{hi}(x_{hi})\},\quad (h=1,\ldots,L\ ;\ i=1,\ldots,I)$$

The proof is completed by choosing x_{hi}^{*} such that $A_{hi}(x_{hi}^{*})\cong A_{hi}(x_{hi}^{\circ})\lambda_{hi1}(x_{hi}^{\circ})$ for $h=1,\ldots,L$; $i=1,\ldots,I$.

From (4), the overall Bayes risk is as follows :

$$\sum_{i=1}^{I}f(N_{\cdot i})\sum_{h=1}^{L}W_{hi}^2R_{hi}(x_{hi}^{\circ}) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5)$$

In situations where σ_h^2 and τ_h^2 are both unknown, we can derive the empirical Bayes estimators by replacing the two variance components in (4) with the nonnegative invariant quadratic estimators (IQE) proposed by Mathew and Sinha (1992). Assuming that $n_{hi}\geq 1$ \$ for all h and i, we can choose the following estimators $\hat{\tau}_h^2$ and $\hat{\sigma}_h^2$ for τ_h^2 and σ_h^2, respectively.

$$\hat{\tau}_h^2=\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_{hi}}(y_{hij}-\bar{y}_{hi})^2}{n_{h\cdot}-I},\quad h=1,\ldots,L.$$

$$\hat{\sigma}_h^2 = a_h\left(\sum_{i=1}^{I} n_{hi}\bar{y}_{hi}^2 - n_{h\cdot}\bar{y}_{h\cdot}^2\right),\quad h=1,\ldots,L.$$

By Theorem 2.3 of Mathew and Sinha (1992) $a_h\,(h=1,\ldots,L)$ is chosen such that $\hat{\sigma}_h^2$ has a uniformly smaller mean squared error than every unbiased IQE of σ_h^2. Mathew and Sinha (1992) show that $\hat{\tau}_h^2\ (h=1,\ldots,L)$ is the unique nonnegative unbiased IQE of $\tau_h^2\,(h=1,\ldots,L)$. When n_{hi} is small or equal to zero, we can aggregate temporarily some of the subareas within the same poststratum to have stable estimators of τ_h^2 and σ_h^2.

4. The Simulation Study

A simulation study was carried out under the superpopulation model (2) to illustrate the characteristics of three different methodologies for producing small area estimation. The three methods evaluated were, SYN, UNB and the approximate Bayes rule, described in Section 3, (denoted by SYN/C). If $n_{hi}=0$ we define UNB estimators to be zero (somewhat arbitrarily, since, strictly speaking, the estimator is then undefined). We consider the case in which $L=2$, $I=10$, $P_{hi}=0.075$ for all h and $i=1,2,3$; $P_{hi}=0.05$ for all h and $i=4,5,6,7$; $P_{hi}=0.025$ for all h and $i=8,9,10$. For the Monte Carlo simulation, 5000 repeated simple random samples, each of size $n=120$, were selected from the superpopulation of $N=1000$ units. The superpopulations were generated using the RANNOR generating function of the SAS library. The parameters were $\mu_1=5$, $\mu_2=10$ and three sets values for (σ_h^2,τ_h^2), namely, (0.2,1), (0.5,1) and (1,1). The results of the corresponding h_{hi}^*'s for the approximate Bayes rules of the three superpopulations are 5, 4 and 4 for $i=1,2,3$; 4, 3 and 2 for $i=4,5,6,7$; 3, 2 and 1 for $i=8,9,10$, respectively.

To look at E_mV_d and $E_mB_d^2$, the estimates were averaged over all 5000 repetitions from the superpopulation model. Table 1 presents the results of E_mV_d, $E_mB_d^2$ (denoted as Var and B^2 respectively) and $MSE=E_mV_d+E_mB_d^2$ for each area of the three superpopulations. The mean sample size taken for each area (denoted by 'ms') and the overall simulated risks with $f(N_i)=1/I$ in (3.3) (denoted by 'av') are also presented in 《Table 1》.

《Tab. 1》 Var, B^2 and MSE of Each of Three Estimators over 5000 Repeated Simple Random Sample averaged over 5000 repetitions for the Superpopulation model with three sets of values for σ_h^2 and τ_h^2.

area (i)	n_i^*	SYN Var^+	SYN $B^{2\ddagger}$	SYN $MSE^\S$	UNB Var	UNB B^2	UNB MSE	SYN/C Var	SYN/C B^2	SYN/C MSE
					$(\sigma_h^2,\tau_h^2)=(0.2,1)$					
1	9.03	0.010	0.074	0.084	0.066	0.000	0.066	0.059	0.000	0.059
2	9.01	0.010	0.061	0.071	0.065	0.000	0.065	0.060	0.000	0.060
3	9.03	0.010	0.059	0.069	0.065	0.000	0.065	0.059	0.000	0.059
4	5.97	0.010	0.069	0.079	0.150	0.000	0.150	0.073	0.001	0.074
5	5.99	0.010	0.081	0.091	0.138	0.000	0.138	0.074	0.002	0.076
6	5.99	0.010	0.090	0.100	0.154	0.000	0.154	0.075	0.002	0.077
7	6.01	0.010	0.072	0.082	0.151	0.000	0.151	0.074	0.002	0.076
8	3.02	0.010	0.106	0.116	1.398	0.097	1.495	0.061	0.035	0.096
9	3.00	0.010	0.118	0.128	1.448	0.102	1.550	0.060	0.032	0.092
10	2.97	0.010	0.085	0.095	1.428	0.099	1.527	0.060	0.029	0.089
av	5.00	0.010	0.082	0.092	0.506	0.030	0.536	0.066	0.010	0.076
					$(\sigma_h^2,\tau_h^2)=(0.5,1)$					
1	9.00	0.011	0.184	0.195	0.066	0.000	0.066	0.060	0.000	0.060
2	9.02	0.011	0.151	0.162	0.065	0.000	0.065	0.059	0.000	0.059
3	9.00	0.011	0.147	0.158	0.065	0.000	0.065	0.060	0.000	0.060
4	5.98	0.011	0.173	0.184	0.150	0.000	0.150	0.097	0.001	0.098
5	5.99	0.011	0.203	0.214	0.138	0.000	0.138	0.096	0.001	0.097
6	5.99	0.011	0.226	0.237	0.154	0.000	0.154	0.100	0.001	0.101
7	6.00	0.011	0.179	0.190	0.151	0.000	0.151	0.098	0.001	0.099
8	3.00	0.011	0.265	0.276	1.402	0.098	1.500	0.177	0.010	0.187
9	2.98	0.011	0.295	0.306	1.445	0.102	1.547	0.178	0.012	0.190
10	3.00	0.011	0.213	0.224	1.440	0.099	1.539	0.180	0.008	0.188
av	5.00	0.011	0.204	0.215	0.507	0.030	0.537	0.111	0.003	0.114
					$(\sigma_h^2,\tau_h^2)=(1,1)$					
1	9.00	0.014	0.368	0.382	0.066	0.000	0.066	0.063	0.000	0.063
2	9.00	0.014	0.303	0.317	0.065	0.000	0.065	0.063	0.000	0.063
3	9.00	0.014	0.293	0.307	0.065	0.000	0.065	0.063	0.000	0.063
4	6.02	0.014	0.346	0.360	0.150	0.000	0.150	0.102	0.001	0.103
5	6.00	0.014	0.406	0.420	0.138	0.000	0.138	0.099	0.001	0.101
6	6.00	0.014	0.452	0.466	0.153	0.000	0.153	0.103	0.001	0.104
7	5.99	0.014	0.359	0.373	0.151	0.000	0.151	0.104	0.001	0.105
8	3.01	0.014	0.529	0.543	1.409	0.098	1.507	0.227	0.002	0.229
9	2.99	0.014	0.590	0.604	1.446	0.101	1.547	0.228	0.002	0.230
10	2.99	0.014	0.427	0.441	1.458	0.100	1.558	0.229	0.002	0.231
av	5.00	0.014	0.408	0.422	0.510	0.030	0.540	0.128	0.001	0.129

$*$: mean sample size taken for each area. $+$: $E_m V_d$, $\ddagger$: $E_m B_d^2$, § : $E_m V_d + E_m B_d^2$.

From《Table 1》, the following conclusions emerge：

A. The SYN estimators are badly biased unless σ_h^2/τ_h^2 is very small. This causes large MSE of the SYN estimators although they consistently have an attractively low variance compared to the UNB and SYN/C estimators. The UNB estimators are essentially unbiased except in the small areas 8, 9 and 10, where the events $n_{hi}=0$, have a large probability. Their variance is consistently higher than that of SYN and SYN/C estimators. The MSE of the UNB estimators is unacceptably large in small areas because $\bar{y}_{hi}$ is given the value 0 when $n_{hi}=0$ (which creates large bias in the UNB estimate).

B. The SYN/C estimators are biased in some areas, but the bias is much less pronounced than that of SYN estimators. The SYN/C estimators have much smaller variance than that of UNB in small areas. In all areas of the three superpopulations, the SYN/C estimators have a smaller MSE than either of SYN and UNB estimators.

5. Discussion

The use of the proposed class of ratio estimators based on poststratification pinpoints one of the fundamental issues of controversy in statistical theory, namely, that of conditional inference (see Fuller (1966) for a conditional approach). The main argument in favor of the unconditional approach is simplicity; the quality of a whole procedure is described by a single number, namely, the Bayes risk (5), when a Bayesian approach is used. It is thus reasonable to use unconditional variance, or mean-squared error at the planning stage to choose among methods. Our estimators can be used to improve the collapsing procedure currently employed by the U.S. Bureau of the Census, where x_{hi}'s were set equal to some selected constants independent of sample outcomes for all h and i. The extension of these results to stratified random sampling is straightforward. However, the extension to more complex designs, such as two-stage sampling involving clustering, is not obvious. The main difficulty is the computation of $E_d(\gamma_h)$ or $A_{hi}(x_{hi})$ which would require the knowledge of subarea sizes in each next to last stage unit. The other difficulty is the complexities due to the correlation of units within the same cluster. However for

many survey designs, it is possible to compute approximations to the $E_d(\gamma_h)$ and $A_{hi}(x_{hi})$ as follows.

Assume that the last stage units are obtained by simple random sampling from next to last stage units. Let the subscript c index denote next to last stage unit. Denote by $p(s)$ the probability that a sample s of next to last units has been drawn by the specified design. Denote by $m_c(s)$ the specified number of last stage units to be drawn from the c^{th} next to last stage units in s. Denote by $P_{hic}(s)$ and $P_{hc}(s)$ the proportion of last stage units in the c^{th} next to last stage unit which is in the hi^{th} subarea and h^{th} poststratum, respectively. Then $E_d(\gamma_h)$ is approximated by

$$E_d(\gamma_h) = \sum_S p(s)\left(1-\left(1-\overline{P}_h(s)^{\overline{m}(s)}\right)\right)$$

where $\overline{P}_h(s)$ is an average value of the $P_{hc}(s)$ and $\overline{m}(s)$ is an average value of the $m_c(s)$. Similarly, $A_{hi}(x_{hi})$ can be approximated by

$$A_{hi}(x_{hi}) = \sum_S p(s) \sum_{k=x_{hi}}^{v(s)} \binom{\overline{m}(s)}{k} \left(\overline{P}_{hi}(s)\right)^k \left(1-\overline{P}_{hi}(s)^{\overline{m}(s)-k}\right)$$

where $\overline{P}_{hi}(s)$ is an average value of the $P_{hic}(s)$, $v(s) = \min(\overline{m}(s), N_{hi})$. Improvements in computation of the $E_d(\gamma_h)$ and $A_{hi}(x_{hi})$ will be left for subsequent investigation.

References

Cressie, Noel (1989), "Empirical Bayes Estimation of Undercount in the Decennial Census", *J. Amer. Statist. Assoc.* 84 : 1033~1044.

Fuller, W. (1966), "Estimation Employing Post Strata", *J. Amer. Statist. Assoc.* 61 : 1172~1183.

Ghosh, M. and G. Meeden (1986), "Empirical Bayes Estimation in Finite Population Sampling". *J. Amer. Statist. Assoc.* 81 : 1058~1062.

_____and P. LAHIRI (1987), "Robust Empirical Bayes Estimation of Means From Stratified Samples", *J. Amer. Statist. Assoc.* 82 : 1153~1162.

Goldstein, M. (1975), "Approximate Bayes Solutions to Some Nonparametric

Problems" , *Ann. Statist.* 3 : 512～517.

Gonzalez, M. E. and C. Hoza (1978), "Small-Area Estimation with Application to Unemployment and Housing Estimates" , *J. Amer. Statist. Assoc.* 73 : 7～15.

Mathew, T. and B. K. Sinha (1992), "Nonnegative Estimation of Variance Components in Unbalanced Mixed Models with Two Variance Components" , *J. Multivariate Anal.* 42 : 77～101.

Purcell, N. J. and L. Kish(1980), "Postcensal Estimates for Local Areas(or Domains)" , *Inter. Statist. Rev.* 48 : 3～18.

Stephan, F. (1945), "The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate" , *Ann. Math. Statist.* 16 : 50～61.

以事後分層估計小區域

沈葆聖*

摘要

本文提出以事後分層法為依據的小區域估計值。該估計值同時考量層內同質性假設之合理性以及小區域樣本數不足的困境。我們推導簡單隨機抽樣下該估計值的小樣本性質。在超族群模式下，經由貝氏分析可以證實估計值的合理性。此外，我們探討複雜抽樣下，該估計值的推廣。

關鍵詞：小區域估計、事後分層、貝氏分析。

* 東海大學統計系。