Empirical Bayes Variance Estimation for Stratified Cluster Sampling with One PSU per Stratum

Sepideh Mosaferi

University of Maryland, College Park, Maryland 20742, USA

E-mail: smosafer@umd.edu

Date: May 2015

SUMMARY

A single primary sampling unit (PSU) per stratum design is a popular design for estimating the parameter of interest. Although, the point estimator of the design is unbiased and efficient, an unbiased variance estimator does not exist. A common method for variance estimation for a 1 PSU per stratum design is based on collapsing or combining two adjacent strata, but the attained estimator of variance is not design-unbiased, and the bias increases as the population means of collapsed strata become more different. Since in some situations, including small area estimation, an unbiased estimator of variance is needed, the one-per-stratum design with collapsed stratum variance estimator cannot be a good choice, and some statisticians prefer a design in which 2 PSUs per stratum are selected. Here, we firstly compare a 1 PSU per stratum design to a 2 PSUs per stratum design. Then, an empirical Bayes estimator for the variance of a 1 PSU per stratum design will be proposed. In a simulation study, we will show that the empirical Bayes method performs better than the classical collapsed stratum variance method.

Some key words: Collapsing strata; Empirical Bayes estimator; One PSU per stratum design; Two PSUs per stratum design; Unbiased estimator; Variance estimation.

1. Introduction

One PSU per stratum design is theoretically efficient for providing an unbiased estimator of a population parameter. However, estimating the variability of the attained estimator is infeasible without considering any implicit assumptions such as collapsing strata; such assumptions produce biased estimator of the variance. Some examples that use the multi-stage one per stratum designs include the Current Population Survey (CPS) conducted by the U.S. Census Bureau, the U.S. Bureau of Labor Statistics, and the National Crime Victimization Survey.

Due to the lack of an unbiased variance estimator for the 1 PSU per stratum design, some survey statisticians would prefer to select two PSUs per stratum since variance estimators for simple estimators like the Horvitz-Thompson are unbiased. Surveys such as the Survey of Income and Program Participation of the U.S. Census Bureau and the U.S. Department of Agriculture's National Resources Inventory use a multi-stage two-per-stratum design and two-per-stratum design, respectively. For these designs, an unbiased variance estimator exists without any implicit assumptions, but the one PSU per stratum design still has its own popularity because it allows deeper stratification.

Collapsed stratum for variance estimation was first introduced by Hansen et al. (1953). This method usually causes an overestimation in the variance of estimator; therefore, two PSUs per stratum design might be prefered since its variance is design-unbiased. Hansen et al. (1953) and Isaki (1983) used some auxiliary variables, which are well-correlated with the expected values of the mean in each stratum, to reduce the bias of variance estimator.

In 1969, Hartley et al. proposed a method of grouping strata where each group contains 7 to 15 strata and then applied a linear regression of the group means on one or more auxiliary variables for estimating the variance of one PSU per stratum design; regression residuals used to estimate the stratum variance components. This method needs additional experience on the performance before being used, and the bias of the variance estimator depends on how well the regression model fits. The idea of stratum boundaries being chosen by a random process prior to the sample selection was proposed by Fuller (1970). The Fuller's method is biased when the stratum boundaries are not randomized beforehand.

Rust and Kalton (1987) examined the effects of collapsing strata in pairs, triples and larger groups on the quality of the variance estimator and found that a greater level of collapsing is desirable when a small sample of PSUs is selected. They mentioned some factors that might be considered in deciding on the extent of collapsing.

Under the assumption that Durbin's (1967) sampling scheme was used within collapsed strata, Shapiro and Bateman (1978) applied the Yates-Grundy (1953) variance estimator. This variance can be biased upwards, but the bias is relatively smaller than the collapsed method, and the variance estimator is more stable; however, for their empirical example, the authors did not consider the same number of units in collapsed strata; this resulted in the collapsed stratum variance performing poorly compared to the situation of having the same number of units in each collapsed strata.

Mantel and Giroux (2009) proposed a new approach based on the components of variance from different stages of sampling. They studied the Canadian Health Measures Survey (CHMS), a 3-stage sample design, in which the number of PSUs per stratum is very small. In the study, they assumed that a randomized PPS systematic (RPPSS) sampling design was used for the CHMS; this assumption can introduce an unknown bias into the variance estimation. They also could not calculate an uncollapsed variance estimate for the Atlantic, a stratum with 1 PSU.

Recently, Breidt et al. (2014) proposed a nonparametric alternative method that replaces a collapsed stratum estimator by kernel-weighted stratum neighborhoods and used deviations from a fitted mean function to estimate the variance. They applied their method to the U.S. Consumer Expenditure Survey to show the superior practical perfomance of their method over the collapsed stratum variance estimator. The estimator that they used is a natural nonparametric extension of linear models, proposed by the pioneers such as Hartley et al. and Isaki.

In fact, most of the alternative recommended methods for the collapsed stratum variance are based on the existence of some concomitant or auxiliary information; nevertheless, this kind of desirable auxiliary information might not be readily available for all of strata. So, finding an acceptable comparative variance estimator for 1 PSU per stratum design without using auxiliary information is needed to decrease the bias and mse.

In section 2 of this paper, we systematically compare the 1 PSU per stratum design *(Design 1)* to 2 PSUs per stratum desgin *(Design 2)* based on the actual variance of the point estimator of population mean via a simulation study. Section 3 gives a comparison of the variance estimation of 2 designs in a theoretical way and in a simulation study. In section 4, a method based on an empirical Bayes estimator will be proposed as a substitution of the collapsed stratum variance for the 1 PSU per stratum design. The findings of comparisons between the empirical Bayes estimator of variance and classical collapsed stratum variance based on a simulation study are summarized in section 5.

2. Comparison of Design 1 and Design 2

For simplicity of exposition, we consider a stratified design with H strata and n_h units drawn from the hth stratum consisting of N_h units using simple random sampling without replacement (SRSWOR), $h = 1,...,H$. In many applications, these units could be a primary stage units (PSU). We are interested in estimating the finite population mean $\bar{Y} = N_T^{-1}\sum_{h=1}^{H} W_h\bar{Y}_h$, where $N_T = \sum_{h=1}^{H} N_h$, $W_h = N_h/N_T$ and $\bar{Y}_h$ is the finite population mean of the hth stratum. The Horvitz-Thompson unbiased estimator of the finite population mean ($\bar{Y}$) is given by:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$$

where $\bar{y}_h$ is the sample mean for the hth stratum. The associated randomization-based variance is given by:

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{1}{n_h}(1 - \frac{n_h}{N_h})S_h^2 \qquad (1)$$

where $S_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(y_{hj} - \bar{Y}_h)^2$.

In this paper, we compare two popular design options: $n_h = 1$ (one PSU per stratum design, say *Design 1*) and $n_h = 2$ (two PSUs per stratum design, say *Design 2*). These two options are widely used in the context of stratified cluster sampling and stratified multi-stage sampling designs. To make a fair comparison of Design 1 and Design 2, we consider $2H$ strata for Design 1 and H groups, each with two strata, for Design 2; therefore, we have equal number of units (PSUs) for both designs.

The relative efficiency of Design 1 relative to Design 2 can be measured by $Eff = \frac{V_2(\bar{y}_{st})}{V_1(\bar{y}_{st})}$, where V_1 and V_2 are the randomization-based variances of the same point estimator $\bar{y}_{st}$ mentioned in (1). If $Eff > 1$, Design 1 is more efficient than Design 2. On the other hand, if $Eff < 1$, Design 1 is less efficient than Design 2. If $Eff = 1$, the two designs are equivalent.

To compare Design 1 and Design 2, we conduct a Monte Carlo simulation study. Following Hansen et al. (1983), we generate a finite population of size $N_T = 20,000$ units by drawing a random sample of size 20,000 from a bivariate superpopulation characterized by two dimensional random vector (X,Y), where X has a gamma distribution with shape 2 and scale 5, $f(x) = .04xexp(-x/5)$, and Y (conditional on X) has a gamma distribution with shape c and scale b^2 with $b = 1.25x^{3/2}(8+5x)^{-1}$ and $c = .04x^{-3/2}(8+5x)^2$.

To compare the effects of the number of strata H in Design 1 on the efficiency, we consider $H = 10, 50$, and 100 strata, which are formed based on the quantiles of X. The results are reported in Table 1. Overally, Design 1 performs better than Design 2 with considerable efficiency for small number of strata for Design 1. However, efficiency diminishes as the number of strata increases.

Table 1: Comparison of Design 1 and Design 2 Based on the Number of Strata

*Design*1			*Design*2			*Comparison*
H	N_h	$V_1(\bar{y}_{st})$	H	N_h	$V_2(\bar{y}_{st})$	$Eff = \frac{V_2(\bar{y}_{st})}{V_1(\bar{y}_{st})}$
10	2000	0.2515	5	4000	0.2759	1.0969
50	400	0.0464	25	800	0.0469	1.0104
100	200	0.0229	50	400	0.0232	1.0109

How does the efficiency depend on the difference in the finite population means and/or variances of the two strata within a group? To be able to assess the effects of differences in population means and population variances within the collapsed strata on the randomization-based variance (1), we considered $H = 10$ strata and 5 strata for Designs 1 and 2, respectively, and employed some changes to the generated population's means and variances within and between groups (collapsed strata), which are shown with the "g" symbol. We separately generated data for each stratum from the Normal distribution, $N(\mu = mean(y), \sigma^2 = var(y))$ and considered different coefficients of k, 2k, 3k, 4k, 5k (k = 1, 2) to implant some changes in μ and

σ^2 for groups 1, 2, 3, 4, and 5, respectively. The populations used for generating the data based on the Normal distribution are given in Table 2, and the results of comparisons associated to the groups of Table 2 are in Table 3.

Table 2: Generated Populations from the Normal Distribution for the Comparison

$\bar{Y}_{g1} \approx \bar{Y}_{g2}$	stratum1 $\sim N(\mu+1,\sigma^2)$ stratum2 $\sim N(\mu+1,\sigma^2)$; stratum3 $\sim N(\mu+2,2\sigma^2)$ stratum4 $\sim N(\mu+2,2\sigma^2)$
$S^2_{g1} \approx S^2_{g2}$	stratum5 $\sim N(\mu+3,3\sigma^2)$ stratum6 $\sim N(\mu+3,3\sigma^2)$; stratum7 $\sim N(\mu+4,4\sigma^2)$ stratum8 $\sim N(\mu+4,4\sigma^2)$
	stratum9 $\sim N(\mu+5,5\sigma^2)$ stratum10 $\sim N(\mu+5,5\sigma^2)$
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$	stratum1 $\sim N(\mu+1,\sigma^2)$ stratum2 $\sim N(\mu+2,\sigma^2)$; stratum3 $\sim N(\mu+2,2\sigma^2)$ stratum4 $\sim N(\mu+4,2\sigma^2)$
$S^2_{g1} \approx S^2_{g2}$	stratum5 $\sim N(\mu+3,3\sigma^2)$ stratum6 $\sim N(\mu+6,3\sigma^2)$; stratum7 $\sim N(\mu+4,4\sigma^2)$ stratum8 $\sim N(\mu+8,4\sigma^2)$
	stratum9 $\sim N(\mu+5,5\sigma^2)$ stratum10 $\sim N(\mu+10,5\sigma^2)$
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$	stratum1 $\sim N(\mu+1,\sigma^2)$ stratum2 $\sim N(\mu+2,2\sigma^2)$; stratum3 $\sim N(\mu+2,2\sigma^2)$ stratum4 $\sim N(\mu+4,4\sigma^2)$
$S^2_{g1} \neq S^2_{g2}$	stratum5 $\sim N(\mu+3,3\sigma^2)$ stratum6 $\sim N(\mu+6,6\sigma^2)$; stratum7 $\sim N(\mu+4,4\sigma^2)$ stratum8 $\sim N(\mu+8,8\sigma^2)$
	stratum9 $\sim N(\mu+5,5\sigma^2)$ stratum10 $\sim N(\mu+10,10\sigma^2)$
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$	stratum1 $\sim N(\mu+1,\sigma^2)$ stratum2 $\sim N(\mu+1,2\sigma^2)$; stratum3 $\sim N(\mu+2,2\sigma^2)$ stratum4 $\sim N(\mu+2,4\sigma^2)$
$S^2_{g1} \neq S^2_{g2}$	stratum5 $\sim N(\mu+3,3\sigma^2)$ stratum6 $\sim N(\mu+3,6\sigma^2)$; stratum7 $\sim N(\mu+4,4\sigma^2)$ stratum8 $\sim N(\mu+4,8\sigma^2)$
	stratum9 $\sim N(\mu+5,5\sigma^2)$ stratum10 $\sim N(\mu+5,10\sigma^2)$

According to Table 3, when population means within groups are different, Design 1 is more efficient than Design 2; on the other hand, when the population means within groups are similar, there is no preference between two designs. In addition, changes in the population variances within the groups do not show any conspicuous effects on the efficiency. In the next section, we will assess the effect of collapsing stratum variance in Design 1 to find out whether Design 1 performs better than Design 2 or not with respect to the coverage probability of mean.

Table 3: Comparison of Design 1 and Design 2 Based on the Differences of Means and Variances

	*Design*1			*Design*2			*Comparison*
Case Study	H	N_h	$V_1(\bar{y}_{st})$	H	N_h	$V_2(\bar{y}_{st})$	$Eff = \frac{V_2(\bar{y}_{st})}{V_1(\bar{y}_{st})}$
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S^2_{g1} \approx S^2_{g2}$	10	2000	1.6198	5	4000	1.6198	1.0000
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S^2_{g1} \approx S^2_{g2}$	10	2000	1.6197	5	4000	1.9021	1.1743
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S^2_{g1} \neq S^2_{g2}$	10	2000	2.4310	5	4000	2.7047	1.1126
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S^2_{g1} \neq S^2_{g2}$	10	2000	2.4476	5	4000	2.4476	1.0000

3. Variance Estimation in Design 1 and Design 2

3.1. Theoretical Expressions

As in section 2, we assume that we have two strata in each of the H groups and let N_{gi} denote population size for the ith stratum within the g group ($g = 1, \cdots, H,\ i = 1, 2$). Let y_{gij} denote the value of the characteristic of interest for the j unit in the i stratum within the gth group ($g = 1, \cdots, H,\ i = 1, 2,\ j = 1, \cdots, N_{gi}$). For simplicity in exposition, we assume $N_{gi} = N$ $\forall g = 1, \cdots, H,\ i = 1, 2$, so $N_T = 2HN,\ W_{gi} = N_{gi}/N_T = \frac{1}{2H}$. For $g = 1, \cdots, H,\ i = 1, 2$, define

$\bar{Y}_{gi} = N^{-1}\sum_{j=1}^{N} y_{gij}$, finite population mean for the ith stratum within the gth group

$S_{gi}^2 = (N-1)^{-1}\sum_{j=1}^{N}(y_{gij} - \bar{Y}_{gi})^2$, finite population variance for the i stratum in the gth group

$\mu_r^{gi} = (N-1)^{-1}\sum_{j=1}^{N}(y_{gij} - \bar{Y}_{gi})^r$, finite population rth central moment ($r \geq 1$). Note that $\mu_1^{gi} = 0$ and $\mu_2^{gi} = S_{gi}^2$.

We also assumed that the Finite Population Correction (FPC) factor is negligible as the sample size is only 1 or 2 per stratum, and N is large.

For Design 1, $N_h = N$, $N_T = 2HN$, and $S_h^2 = \frac{1}{N-1}\sum_{j=1}^{N}(y_{hj} - \bar{Y}_h)^2$. The true variance based upon Design 1 is:

$$V(\bar{y}_{st}) = \frac{1}{4H^2}\sum_{h=1}^{2H} S_h^2 \tag{2}$$

We used (=) symbol instead of ($\approx$) symbol for the formulae of variances. We can rewrite (2) as:

$$V(\bar{y}_{st}) = \frac{1}{4H^2}\sum_{g=1}^{H}(S_{g1}^2 + S_{g2}^2)$$

The collapsed strata variance estimator is given by:

$$v(\bar{y}_{st}) = \frac{1}{2H^2}\sum_{g=1}^{H} s_g^2 \tag{3}$$

where $s_g^2 = \sum_{i=1}^{2}(y_{gi} - \bar{y}_g)^2$, and $\bar{y}_g = \frac{y_{g1}+y_{g2}}{2}$. The method relies on the implicit assumption: $\bar{Y}_{g1} = \bar{Y}_{g2} = \bar{Y}_g$. Estimator (3) is design-biased, and its bias with respect to Design 1 is given by:

$$Bias(v(\bar{y}_{st})) = \frac{1}{4H^2}\sum_{g=1}^{H}(\bar{Y}_{g1} - \bar{Y}_{g2})^2 \tag{4}$$

The proof of (4) is given in the Appendix section.

As it is clear from (4), the bias is not related to the population variances within the groups. Wolter computed the bias of population total given the original sampling design (Wolter, 2007, p. 51). Equation (4) suggests the strategy of how we can group strata to reduce the bias of collapsed stratum variance by putting more similar strata in pairs with respect to the characteristic of interest to minimize the difference $|\bar{Y}_{g1} - \bar{Y}_{g2}|$.

In order to find out the *MSE* of $v(\bar{y}_{st})$, we need its *variance*; with pursuing the mentioned assumptions, the variance is:

$$\begin{aligned} Var(v(\bar{y}_{st})) = \frac{1}{16H^4}\sum_{g=1}^{H}\{&\mu_4^{g1} + \mu_4^{g2} + 2S_{g1}^2 S_{g2}^2 + 4(\bar{Y}_{g1} - \bar{Y}_{g2})^2(S_{g1}^2 + S_{g2}^2) \\ &- (S_{g1}^2 - S_{g2}^2)^2 + 4(\bar{Y}_{g1} - \bar{Y}_{g2})(\mu_3^{g1} - \mu_3^{g2})\} \end{aligned} \tag{5}$$

The proof of (5) is given in the Appendix section. If $\mu_4^{g1} = \mu_4^{g2} = \mu_4^g$, $\mu_3^{g1} = \mu_3^{g2} = \mu_3^g$, and $S_{g1}^2 = S_{g2}^2 = S_g^2$, then

$$Var(v(\bar{y}_{st})) = \frac{1}{8H^4} \sum_{g=1}^{H} \{\mu_4^g + (S_g^2)^2 + 4S_g^2(\bar{Y}_{g1} - \bar{Y}_{g2})^2\}$$

Therefore, the MSE of $v(\bar{y}_{st})$ under all of these equality assumptions is:

$$\begin{aligned} MSE(v(\bar{y}_{st})) &= Var(v(\bar{y}_{st})) + \{Bias(v(\bar{y}_{st}))\}^2 \\ &= \frac{1}{8H^4} \sum_{g=1}^{H} \{\mu_4^g + (S_g^2)^2 + 4S_g^2(\bar{Y}_{g1} - \bar{Y}_{g2})^2\} + \frac{1}{16H^4} \sum_{g=1}^{H} (\bar{Y}_{g1} - \bar{Y}_{g2})^4 \end{aligned}$$

The mse is inversely related to the number of strata H. As a result, according to the asymptotic properties we can expect that as the number of strata H increases, mse decreases.

For Design 2, we consider the above mentioned assumptions and use the standard variance of stratified estimator. In the next section, we will compare the variance estimation of two designs based on a simulation study.

3.2. Simulation Study

We performed a simulation experiment to investigate the differences between two designs in respect of the empirical coverage probability (CP) and average length (AL) of a nominal 95% confidence interval (CI) for $\bar{y}_{st}$ under 2 designs. The length of a CI is related to its coverage probability-wider CIs have higher coverage probability.

The population used for this sub-section is similar to the one used in Table 3. The sample designs are the random selection of 1 PSU and 2 PSUs without replacement in each stratum for Design 1 and Design 2, respectively. The process of sample selection was repeated 10,000 times. For each replication (r), the stratified mean estimator and the general two-sided 95% CI, $\bar{y}_{st} \pm 1.96\sqrt{v(\bar{y}_{st})}$, were computed and then empirical coverage probability and average length in percent for both designs were calculated. The results are given in Table 4.

According to Table 4, the average length of CI under Design 1 is bigger than Design 2, so as a result the coverage probability becomes bigger in Design 1. In addition, the length of CI and coverage probability are greater when the population means of collapsed strata are different; this can again reflect the important effect of $|\bar{Y}_{g1} - \bar{Y}_{g2}|$ in collapsing.

Table 4: Empirical Results of Simulation Study for Comparison of Design 1 and Design 2

	*Design*1		*Design*2	
Case Study	*AL*	*CP%*	*AL*	*CP%*
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S_{g1}^2 \approx S_{g2}^2$	4.7309	88.81	4.7168	88.48
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S_{g1}^2 \approx S_{g2}^2$	5.4929	92.35	5.0995	87.82
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S_{g1}^2 \neq S_{g2}^2$	6.4273	90.90	6.0775	88.88
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S_{g1}^2 \neq S_{g2}^2$	5.7853	89.32	5.7591	88.17

4. Empirical Bayes Variance for Design 1

Let $s_g^2 = \frac{1}{2}(y_{g1} - y_{g2})^2$ denotes the pooled sample variance for the gth group. Here y_{gi} denotes the sampled observation from the ith stratum in the gth group. We assume $\frac{s_g^2}{S_g^2} \sim \chi^2(1)$ and an inverse gamma prior IG(a,a) for S_g^2. The posterior distribution $\pi(S_g^2|s_g^2)$ is

$$\pi(S_g^2|s_g^2) \propto f_{S_g^2}(s_g^2)\pi(S_g^2) = \frac{(s_g^2)^{-1/2} e^{-\frac{s_g^2}{2S_g^2}}}{\Gamma(1/2)(2S_g^2)^{1/2}} \frac{a^a}{\Gamma(a)} (S_g^2)^{-a-1} e^{-a/S_g^2}.$$

This is an inverse gamma distribution with shape $a + \frac{1}{2}$ and scale $a + \frac{s_g^2}{2}$, $IG(a + \frac{1}{2}, a + \frac{s_g^2}{2})$. Under the squared error loss function, $L(S_g^2, \delta(s_g^2)) \equiv (S_g^2 - \delta(s_g^2))^2$, the optimal Bayes estimator of S_g^2 is $E(S_g^2|s_g^2)$, which is,

$$\delta(s_g^2) = \frac{a + \frac{s_g^2}{2}}{a - \frac{1}{2}}$$

Since a is unknown, we estimated it using the method-of-moments. To do this, we need the marginal distribution of s_g^2 which is $F(1, 2a)$ or identically $\sqrt{s_g^2} \sim T(2a)$, the t-distribution with $2a$ d.f. The theoretical second order moment based on the t-distribution is $E((\sqrt{s_g^2})^2) = \frac{a}{a-1}$, which is valid for $a > 1$, is replaced by the empirical moment based on the historic data, $m = \frac{1}{H}\sum_{g=1}^{H} s_g^2$; therefore, the solution is:

$$\hat{a}_{MM} = \frac{\frac{1}{H}\sum_{g=1}^{H} s_g^2}{\frac{1}{H}\sum_{g=1}^{H} s_g^2 - 1}$$

which yields an empirical Bayes predictor:

$$\hat{\delta}_{EB}(s_g^2) = \frac{\hat{a}_{MM} + \frac{s_g^2}{2}}{\hat{a}_{MM} - \frac{1}{2}} = \frac{2\hat{a}_{MM} + s_g^2}{2\hat{a}_{MM} - 1}$$

As we have the constraint $a > 1$, for the situations that $\hat{a}_{MM} \leq 1$, we considered $\hat{a}_{MM} = 1 + \varepsilon$ where $\varepsilon = 1e - 06$. By substituting $\hat{\delta}_{EB}(s_g^2)$ into (3), the optimal estimator for the variance of 1 PSU per stratum design is attained:

$$\tilde{v}(\bar{y}_{st}) = \frac{1}{2H^2} \sum_{g=1}^{H} \hat{\delta}_{EB}(s_g^2) \tag{6}$$

In the next section, we will practically study the performance of this estimator.

5. Simulation Study and Results

To compare the empirical Bayes estimator of variance based on (6) and classical collapsed method based on (3), we conducted a Monte Carlo study with 10,000 replications. The empirical relative mse was found using the following formula:

$$\sum_{r=1}^{10,000} \frac{relative\ mse}{10,000},$$

where

$$Relative\ MSE = \frac{\sqrt{(v(\bar{y}_{st,r}) - V(\bar{Y}_{st}))^2}}{V(\bar{Y}_{st})}.$$

The results of comparisons are given in Table 5. According to Table 5, the relative mse of the variance estimator based on the empirical Bayes estimator (6) is smaller than the mse of the classical collapsed stratum variance estimator (3), which

Table 5: Empirical Results of MSEs for the Variance Comparisons

	Method	
Case Study	*Based on the Collapsed Stratum Variance*(3)	*Based on the Empirical Bayes Estimator*(6)
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S^2_{g1} \approx S^2_{g2}$	0.5275	0.4804
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S^2_{g1} \approx S^2_{g2}$	0.6921	0.5869
$\bar{Y}_{g1} \neq \bar{Y}_{g2}$ $S^2_{g1} \neq S^2_{g2}$	0.6239	0.4454
$\bar{Y}_{g1} \approx \bar{Y}_{g2}$ $S^2_{g1} \neq S^2_{g2}$	0.5247	0.4258

is a good evidence that our empirical Bayes predictor performs well and can smooth the variability caused by the collapsed variance.

6. Concluding Remarks

One PSU per stratum design has the advantage of deep stratification which is efficient for estimating the interested value, but the variance estimator based on this method is infeasible without considering any restrictions such as collapsing strata. The collapsed stratum variance estimator usually suffers from the overestimation. In this paper, by applying an empirical Bayes estimator, we are able to decrease the MSE of 1 PSU variance estimator. This method can easily be applied for the stratified cluster sampling with one PSU per stratum. In the future, we plan to study a constrained empirical Bayes estimator to protect against the over-shrinking of variance estimator of 1 PSU per stratum caused by the empirical Bayes estimator.

References

[1] Breidt, F. J., Opsomer, J. D., and Borrego, I. S. (2014+). Nonparametric Variance Estimation under Fine Stratification: An Alternative to Collapsed Strata. *Journal of the American Statistical Association*.

[2] Durbin, J. (1967). Design of Multi-stage Surveys for the Estimation of Sampling Errors. *Applied Statistics*, **16**, 152–164.

[3] Fuller, W. A. (1970). Sampling with Random Stratum Boundaries. *Journal of the Royal Statistical Society B*, **32**, 209–226.

[4] —– (2009). *Sampling Statistics*. Wiley, New York.

[5] Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*. Vol. II, Wiley, New York.

[6] —— (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, **78**, 776–793.

[7] Hartley, H. O., Rao, J. N. K., and Kiefer, G. (1969). Variance Estimation with One Unit Per Stratum. *Journal of the American Statistical Association*, **64**, 841–851.

[8] Isaki, C. T. (1983). Variance Estimation Using Auxiliary Information. *Journal of the American Statistical Association*, **78**, 117–123.

[9] Mantel, H. and Giroux, S. (2009). *Variance Estimation in Complex Surveys with One PSU per Stratum. Proceedings of the Survey Research Methods Section*, American Statistical Association, 3069–3082.

[10] Rust, K. and Kalton, G. (1987). Strategies for Collapsing Strata for Variance Estimation. *Journal of Official Statistics*, **3**, 69–81.

[11] Shapiro, G. M. and Bateman, D. V. (1978). A Better Alternative to the Collapsed Stratum Variance Estimate. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 451–456.

[12] Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer Verlag.

[13] Yates, F. and Grundy, P. M. (1953). Selection without Replacement from within Strata with Probabilities Proportional to Size. *Journal of the Royal Statistical Society B*, **15**, 253–261.

Appendix

Proofs All of proofs in this section are based on the design-based application without considering any models.

proof of (4):

$$
\begin{aligned}
Bias(v(\bar{y}_{st})) &= E(\sum_{g=1}^{H}(\frac{1}{H})^2\frac{1}{2}s_g^2) - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2 = \sum_{g=1}^{H}(\frac{1}{H})^2\frac{1}{2}E(s_g^2) - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2 \\
&= \sum_{g=1}^{H}(\frac{1}{H})^2\frac{1}{2}E(\sum_{i=1}^{2}(y_{gi}-\bar{y}_g)^2) - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2 = \sum_{g=1}^{H}(\frac{1}{H})^2\frac{1}{2}E\{\frac{1}{2}(y_{g1}-y_{g2})^2\} - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2 \\
&= \sum_{g=1}^{H}(\frac{1}{H})^2(\frac{1}{4})E\{(y_{g1}-\bar{Y}_{g1})-(y_{g2}-\bar{Y}_{g2})+(\bar{Y}_{g1}-\bar{Y}_{g2})\}^2 - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2 \\
&= \sum_{g=1}^{H}(\frac{1}{H})^2\frac{1}{4}E\{(y_{g1}-\bar{Y}_{g1})^2+(y_{g2}-\bar{Y}_{g2})^2+(\bar{Y}_{g1}-\bar{Y}_{g2})^2 \\
&\quad -2(y_{g1}-\bar{Y}_{g1})(y_{g2}-\bar{Y}_{g2})+2(y_{g1}-\bar{Y}_{g1})(\bar{Y}_{g1}-\bar{Y}_{g2})-2(y_{g2}-\bar{Y}_{g2})(\bar{Y}_{g1}-\bar{Y}_{g2})\} - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2
\end{aligned}
$$

Under the stratified SRSWOR sampling design, samples per stratum are selected independently; thus, y_{g1} and y_{g2} are independent. Also, $\bar{Y}_{g1}$ and $\bar{Y}_{g2}$, the population means in each collapsed stratum of group g are fixed. As a result $E(y_{g1} - \bar{Y}_{g1})(y_{g2}-\bar{Y}_{g2})$ equals to 0. Furthermore, $E(y_{g1}) = \bar{Y}_{g1}$ and $E(y_{g2}) = \bar{Y}_{g2}$; we can rewrite $E(s_g^2)$ as follows:

$$
E(s_g^2) = \frac{1}{2}E\{(y_{g1}-\bar{Y}_{g1})^2+(y_{g2}-\bar{Y}_{g2})^2+(\bar{Y}_{g1}-\bar{Y}_{g2})^2\}
$$

Thus;

$$Bias(v(\bar{y}_{st})) = \sum_{g=1}^{H} (\frac{1}{H})^2 \frac{1}{4} E\{(y_{g1}-\bar{Y}_{g1})^2 + (y_{g2}-\bar{Y}_{g2})^2 + (\bar{Y}_{g1}-\bar{Y}_{g2})^2\}$$
$$-\sum_{h=1}^{2H} (\frac{1}{2H})^2 S_h^2 = \sum_{g=1}^{H} \frac{1}{4}\frac{1}{H^2}\{S_{g1}^2 + S_{g2}^2 + (\bar{Y}_{g1}-\bar{Y}_{g2})^2\} - \sum_{h=1}^{2H}(\frac{1}{2H})^2 S_h^2$$

And $E(y_{g1}-\bar{Y}_{g1})^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_{g1j}-\bar{Y}_{g1})^2 = (1-\frac{1}{N})S_{g1}^2 \approx S_{g1}^2$, $E(y_{g2}-\bar{Y}_{g2})^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_{g2j}-\bar{Y}_{g2})^2 = (1-\frac{1}{N})S_{g2}^2 \approx S_{g2}^2$. As $\sum_{g=1}^{H}(S_{g1}^2+S_{g2}^2) = \sum_{h=1}^{2H} S_h^2$, the bias is: $Bias(v(\bar{y}_{st})) = \sum_{g=1}^{H}\frac{1}{4H^2}\{(\bar{Y}_{g1}-\bar{Y}_{g2})^2\}$ □

proof of (5):

$$Var(v(\bar{y}_{st})) = Var(\sum_{g=1}^{H}\frac{1}{2H^2}s_g^2) = \frac{1}{4H^4}\sum_{g=1}^{H} Var(s_g^2) = \frac{1}{4H^4}\sum_{g=1}^{H} Var(\sum_{i=1}^{2}(y_{gi}-\bar{y}_g)^2) = \frac{1}{4H^4}\sum_{g=1}^{H} Var(\frac{1}{2}(y_{g1}-y_{g2})^2)$$
$$= \frac{1}{16H^4}\sum_{g=1}^{H}\{E(y_{g1}-y_{g2})^4 - \{E(y_{g1}-y_{g2})^2\}^2\}$$

where

$$E(y_{g1}-y_{g2})^4 = E\{(y_{g1}-\bar{Y}_{g1})^4 + (y_{g2}-\bar{Y}_{g2})^4 + (\bar{Y}_{g1}-\bar{Y}_{g2})^4 + 6(y_{g1}-\bar{Y}_{g1})^2(y_{g2}-\bar{Y}_{g2})^2 + 6(y_{g1}-\bar{Y}_{g1})^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2$$
$$+6(y_{g2}-\bar{Y}_{g2})^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2 + 4(y_{g1}-\bar{Y}_{g1})^3(\bar{Y}_{g1}-\bar{Y}_{g2}) - 4(y_{g2}-\bar{Y}_{g2})^3(\bar{Y}_{g1}-\bar{Y}_{g2})\}$$
$$= \mu_4^{g1} + \mu_4^{g2} + (\bar{Y}_{g1}-\bar{Y}_{g2})^4 + 6S_{g1}^2S_{g2}^2 + 6S_{g1}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2 + 6S_{g2}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2$$
$$+4\mu_3^{g1}(\bar{Y}_{g1}-\bar{Y}_{g2}) - 4\mu_3^{g2}(\bar{Y}_{g1}-\bar{Y}_{g2})$$

and $\{E(y_{g1}-y_{g2})^2\}^2 = \{S_{g1}^2 + S_{g2}^2 + (\bar{Y}_{g1}-\bar{Y}_{g2})^2\}^2$.

Therefore;

$$Var(v(\bar{y}_{st})) = \frac{1}{16H^4}\sum_{g=1}^{H}\{\mu_4^{g1} + \mu_4^{g2} + (\bar{Y}_{g1}-\bar{Y}_{g2})^4 + 6S_{g1}^2S_{g2}^2$$
$$+6S_{g1}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2 + 6S_{g2}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2 + 4\mu_3^{g1}(\bar{Y}_{g1}-\bar{Y}_{g2}) - 4\mu_3^{g2}(\bar{Y}_{g1}-\bar{Y}_{g2}) - \{S_{g1}^2 + S_{g2}^2 + (\bar{Y}_{g1}-\bar{Y}_{g2})^2\}^2\}$$

As a result,

$$Var(v(\bar{y}_{st})) = \frac{1}{16H^4}\sum_{g=1}^{H}\{\mu_4^{g1} + \mu_4^{g2} + 4S_{g1}^2S_{g2}^2 + 4S_{g1}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2 + 4S_{g2}^2(\bar{Y}_{g1}-\bar{Y}_{g2})^2$$
$$-(S_{g1}^2)^2 - (S_{g2}^2)^2 + 4\mu_3^{g1}(\bar{Y}_{g1}-\bar{Y}_{g2}) - 4\mu_3^{g2}(\bar{Y}_{g1}-\bar{Y}_{g2})\}$$

since

$$\mu_3^{g1} = E(y_{g1}-\bar{Y}_{g1})^3 \qquad \mu_3^{g2} = E(y_{g2}-\bar{Y}_{g2})^3 \qquad \mu_4^{g1} = E(y_{g1}-\bar{Y}_{g1})^4 \qquad \mu_4^{g2} = E(y_{g2}-\bar{Y}_{g2})^4$$

□