Adjusted Variance Estimators Based on Minimizing Mean Squared Error for Stratified Random Samples

Guoyi Zhang and Bruce Sun

Abstract

In survey data, it is common that variance is large compared to bias. In this research, we propose slightly biased variance estimators by multiplying a constant c between 0 and 1, which are determined by minimizing the mean squared error (MSE) of $c \times$ estimator of the variance. This research is an extension of work by Kourouklis (2012) to the field of survey sampling. Simulation studies show that the adjusted variance estimators perform very well in regarding MSE compared to the regular variance estimator for simple random and stratified random samples.

Key Words: Biased variance estimator, mean squared error, simulations, stratified random sampling, survey data

1 Introduction

Consider a random sample $X_1, X_2, \cdots, X_n$ from a population with distribution function $F \in \mathcal{F}$. Assume that X_i has finite fourth moment. In general, the population variance σ^2 is estimated by the sample variance $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$. Many researchers believe that there are estimators of the form cs^2 (where c is a constant between 0 and 1) that have smaller MSE than the sample variance s^2. These work include Stein (1964), Brown (1968), Brewster and Zidek (1974), Strawderman (1974), Maruyama (1998), Yatracos (2005) and Maruyama and Strawderman (2006). Kourouklis (2012) proposed a variance estimator $c_1 s^2$ and showed that this estimator has the smallest MSE among the estimators of the form cs^2.

In this research, we extend Kourouklis (2012)'s work to survey data. A survey design usually involves stratification and clustering, which complicates the variance estimation. In addition, survey data is associated with huge variability since data is collected across the nation or from a large area. An adjusted variance estimator will decrease the variance, therefore, improve the confidence intervals drastically.

This research is organized as follows: section 2 introduces notation in a general survey frame with simple random sample without replacement (SRS) and stratified random sample design; section 3 proposes the adjusted variance estimator which has the smallest MSE for stratified random samples; section 4 performs simulation comparisons among the estimators; and section 5 gives conclusions of the research.

2 Notation

Let $U = \{1, 2, \cdots N\}$ be the index set of the finite population with size N, and $y_1, y_2, \cdots, y_N$ be the values of the character of the sampling units in the population. Let $\bar{y}_U$ be the population mean: $\bar{y}_U = \sum_{i=1}^{N} y_i/N$, and S^2 be the population variance: $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$. Also let $\mu_2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/N$ and $\mu_4 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^4/N$ be the centralized second and fourth moments respectively. At the sample $\mathcal{S}$ level, let n be the sample size. Sample mean $\bar{y}$ and sample variance s^2 are defined as $\bar{y} = \sum_{i\in\mathcal{S}} y_i/n$, and $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$.

Under an SRS, $E(\bar{y}) = \bar{y}_U$, and

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, \tag{1}$$

where $(1 - n/N)$ is called the finite population correction.

In a stratified random sample, population with size N is divided into H non-overlapping strata with size $N_h, h = 1, 2, \cdots, H$, such that $N = \sum_{h=1}^{H} N_H$. Let y_{hj} be the value of the character for the jth sampling unit within stratum h. Let $\bar{y}_{hU}$ be the population mean of stratum h with $\bar{y}_{hU} = \sum_{j=1}^{N_h} y_{hj}/N_h$, and S_h^2 be the population variance with $S_h^2 = \sum_{j=1}^{N_h}(y_{hj} - \bar{y}_{hU})^2/(N_h - 1)$. Population mean $\bar{y}_U$ can also be written as a weighted average of the stratum means such as $\bar{y}_U = \sum_{h=1}^{H} N_h\bar{y}_{hU}/N$.

Within each stratum h, an SRS with size n_h is taken independently. Assume that $n_h \geq 2$ throughout the paper, and $\sum_{h=1}^{H} n_h = n$. Let $\mathcal{S}_h$ be the set of n_h units in the SRS within stratum h. Stratum sample mean $\bar{y}_h$ and sample variance s_h^2 are defined as $\bar{y}_h = \sum_{j\in\mathcal{S}_h} y_{hj}/n_h$ and $s_h^2 = \sum_{j\in\mathcal{S}_h}(y_{hj} - \bar{y}_h)^2/(n_h - 1)$. An unbiased estimator of

the population mean $\bar{y}_U$ is:

$$\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h \bar{y}_h}{N}. \tag{2}$$

By equation (1) and indpedent sampling within each stratum, variance of $\bar{y}_{str}$ is

$$\mathrm{Var}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}, \tag{3}$$

and is estimated by

$$\widehat{\mathrm{Var}}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}. \tag{4}$$

3 Proposed adjusted variance estimator in a stratified random sample

Estimating population mean and total are two main topics in survey sampling. In this section, we first propose an adjusted variance estimator of the mean that minimizes MSE under an SRS setting. Next, we extend the estimator to a stratified random sample. Last, we discuss how to estimate the optimal value c in practice.

3.1 Lemma and Theorem

In an SRS, we adjust the sample variance s^2 by $cs^2, 0 < c < 1$, where c is determined by minimizing the MSE of cs^2. This is equivalent to minimize MSE of $\widehat{\mathrm{Var}}(\bar{y})$. We state the result as the following lemma and give a brief proof.

Lemma 1. *For a size n SRS selected from a population with size N, the optimal value c that minimizes MSE(cs^2) is*

$$c_{srs} = S^4/E(s^4), \tag{5}$$

where

$$E(s^4) = \frac{n^2}{(n-1)^2}(aN\mu_4 + bN^2\mu_2^2), \tag{6}$$

with

$$a = \frac{e_1 - e_2}{n^2} - \frac{2(e_1 - 3e_2 + 2e_3)}{n^3} + \frac{e_1 - 7e_2 + 12e_3 - 6e_4}{n^4},$$

$$b = \frac{e_2}{n^2} - \frac{2(e_2 - e_3)}{n^3} + \frac{3(e_2 - 2e_3 + e_4)}{n^4},$$

$$e_1 = n/N, e_2 = \frac{n(n-1)}{N(N-1)}, e_3 = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}, e_4 = \frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)},$$

and μ_4 and μ_2 are the centralized moments defined in Section 2.

Proof.

$$\begin{aligned} \text{MSE}(cs^2) &= E(cs^2 - S^2)^2 \\ &= E(c^2s^4) - 2E(cs^2S^2) + S^4 \\ &= c^2E(s^4) - 2cS^4 + S^4 \end{aligned}$$

Let $g(c) = c^2E(s^4) - 2cS^4 + S^4$. By setting $g'(c) = 0$, and using the fact that $g''(c) = 2E(s^4) > 0$, the optimal value of c that minimizes MSE(cs^2) is $c_{srs} = S^4/E(s^4)$.

The remaining problem is to find $E(s^4)$ under an SRS. Using the fact that $V(\bar{y}) = (1-n/N)S^2/n$, and following a similar argument as that of section 2a.10 by Sukhatme (1984), we obtain $E(s^4)$ as in equation (6). □

We now extend the adjusted variance estiamtor $c_{srs}s^2$ in an SRS to a stratified random sample, The following theorem gives the result.

Theorem 1. *In a stratified random sample, population is divided into H non-overlapping strata, and an SRS is taken independently from each stratum. Let N_h and n_h be the population and sample size, and S_h^2 and s_h^2 be the population and sample variance within stratum h as defined in section 2. The optimal value of c that minimizes $MSE(c\widehat{Var}(\bar{y}_{str}))$ is:*

$$c_{str} = \frac{(\sum_{h=1}^{H} k_h S_h^2)^2}{\sum_{h=1}^{H} k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j S_i^2 S_j^2}, \tag{7}$$

where $E(s_h^4)$ can be derived by equation (6) by taking an SRS of size n_h from stratum h, and $k_h = (1 - n_h/N_h)(N_h/N)^2/n_h$.

Proof. By equation (2), $\bar{y}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h / N$. Recall equation (4) gives estimator of $Var(\bar{y}_{str})$ as

$$\widehat{\mathrm{Var}}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h},$$

which can be written as $\widehat{\mathrm{Var}}(\bar{y}_{str}) = \sum_{h=1}^{H} k_h s_h^2$ with expected value of $\sum_{h=1}^{H} k_h S_h^2$. Now we want to find a constant c, such that MSE of $c\widehat{\mathrm{Var}}(\bar{y}_{str})$ reaches the minimum.

After some algebra,

$$\begin{aligned} & E\left\{(c\widehat{\mathrm{Var}}(\bar{y}_{str}) - \mathrm{Var}(\bar{y}_{str}))^2\right\} \\ = \; & E\left\{\left(\sum_{h=1}^{H} k_h c s_h^2 - \sum_{h=1}^{H} k_h S_h^2\right)^2\right\} \\ = \; & \sum_{h=1}^{H} k_h^2 (c^2 E(s_h^4) - 2cS_h^4 + S_h^4) + \sum_{i=1}^{H}\sum_{j=1, j\neq i}^{H} k_i k_j (c-1)^2 S_i^2 S_j^2 \\ = \; & h(c) \end{aligned}$$

Set $h'(c) = 0$, the local extreme value is obtained at

$$c_{str} = \frac{(\sum_{h=1}^{H} k_h S_h^2)^2}{\sum_{h=1}^{H} k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1, j\neq i}^{H} k_i k_j S_i^2 S_j^2}, \tag{8}$$

where $E(s_h^4)$ can be derived by equation (6). Notice that $h''(c) = \sum_{h=1}^{H} 2k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1, j\neq i}^{H} k_i k_j S_i^2 S_j^2 > 0$. c_{str} is the optimal value of c that minimizes $\mathrm{MSE}(c\widehat{\mathrm{Var}}(\bar{y}_{str}))$.

□

3.2 Estimating c_{srs} and c_{str}

In practice, c needs to be estimated using a larger survey or using sample information. We can use $(n-1)s^2/n$ to estimate μ_2. But estimating the fourth moment μ_4 is challenging. Some recent estimators of the fourth moment are not unbiased, or are based on h-statistics and U -statistics (Heffernan, 1997), which can be computationally expensive. Espejo, Pineda, and Nadarajah (2013) proposed estimating the fourth population central moment under distribution-free setting, which involves variance and covariance among the lower sample moments.

Most practitioners may not have the mathematical and statistical background to understand or use the general estimators given in literature. Assume that an SRS or a stratified random sample are with large size, and that the selected sample is representative of the finite population and estimation bias is small. We then use the fourth sample moment and plugin method to estimate the optimal values of c_{srs} and c_{str} as follows.

$$\hat{c}_{srs} = s^4/\hat{E}(s^4), \tag{9}$$

where

$$\hat{E}(s^4) = \frac{n^2}{(n-1)^2}(aN\hat{\mu}_4 + bN^2\hat{\mu}_2^2), \tag{10}$$

where $\hat{\mu}_4 = \sum_{i=1}^{n}(y_i - \bar{y})^4/n$, $\hat{\mu}_2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$, and a and b are defined as in section 3. Similarly, extending SRS to a stratified random sample, we have

$$\hat{c}_{str} = \frac{(\sum_{h=1}^{H} k_h s_h^2)^2}{\sum_{h=1}^{H} k_h^2 \hat{E}(s_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j s_i^2 s_j^2}, \tag{11}$$

where $\hat{E}(s_h^4)$ can be derived using equation (10) when an SRS of size n_h is taken from stratum h, and $k_h = (1 - n_h/N_h)(N_h/N)^2/n_h$.

4 Simulation studies

In this section, we perform a simulation study to evaluate performance of the proposed adjusted variance estimator. The constant c can be calculated using population data `agpop.csv` by equations (5) and (8), or estimated using samples by equations (9) and (11). We compare bias, variance and MSE of the adjusted variance estimators:

c·variance estimator (named Estimator 2 thereafter) and $\hat{c}$·variance estimator (named Estimator 3 thereafter), with those of regular variance estimator (named Estimator 1 thereafter) under the settings of SRS and stratified random samples.

4.1 Simulation set up

The population data we consider in the simulation is `agpop.csv`, which is available from the textbook (Lohr, 2010) supplementary material. The U.S. government conducts a census of agriculture every five years, collecting data on all farms in the 50 states. The census of agriculture provides data `agpop.csv` on number of farms, total acreage devoted to farms (*acres92* is the total acreage devoted to farms in 1992 and is the variable of interest in the study), farm size, yield of different crops, and a wide variety of other agriculture measures for $N = 3078$ counties and county-equivalents in the United States. These 3078 counties are divided into four regions (strata) with stratum size N_h: North Central (NC, stratum 1, $N_1 = 1054$,), North East (NE, stratum 2, $N_2 = 220$), South (S, stratum 3, $N_3 = 1382$) and West (W, stratum 4, $N_4 = 422$).

Simulation does $L = 100,000$ times for each setting. Each time, we draw a sample from the population data `agpop.csv` using either SRS with sample size $n = 300$ or stratified proportional allocated random sample with $(n_1, n_2, n_3, n_4) = (103, 21, 135, 41)$. In a general notation, let $\hat{\theta}$ be an estimator of θ. Assume $\hat{\theta}^{(i)}$ represents the estimator of θ from the ith sample, $i = 1, \cdots, L$. The Monte Carlo mean E_{MC}, Monte Carlo bias B_{MC}, Monte Carlo variance V_{MC}, and Monte Carlo

MSE are given by the following formulas

$$\mathrm{E}_{MC}\{\hat{\theta}\} = L^{-1}\sum_{i=1}^{L}\hat{\theta}^{(i)}, \tag{12}$$

$$\mathrm{B}_{MC}\{\hat{\theta}\} = E_{MC}\{\hat{\theta}\} - \theta, \tag{13}$$

$$\mathrm{Var}_{MC}\{\hat{\theta}\} = L^{-1}\sum_{m=1}^{L}[\hat{\theta}^{(i)} - E_{MC}\{\hat{\theta}\}]^2, \tag{14}$$

and the main criterion for determining efficiency: Monte Carlo MSE is defined by

$$\mathrm{MSE}_{MC}\{\hat{\theta}\} = L^{-1}\sum_{i=1}^{L}\{\hat{\theta}^{(i)} - \theta\}^2. \tag{15}$$

True mean $\bar{y}_U$ is the average of y_i's from the population. For SRS, true variance of $\bar{y}$ is calculated by $\mathrm{Var}(\bar{y}) = (1 - n/N)S^2/n$ (equation (1)). For a stratified random sample, variance of $\bar{y}_{str}$ is $\mathrm{Var}(\bar{y}_{str}) = \sum_{h=1}^{H} k_h S_h^2$ (equation (3)). The unajusted variance estimators of $\mathrm{Var}(\bar{y})$ and $\mathrm{Var}(\bar{y}_{str})$ from the ith sample are $\widehat{\mathrm{Var}}^{(i)}(\bar{y}) = (1 - n/N)s^2/n$ and $\widehat{\mathrm{Var}}^{(i)}(\bar{y}_{str}) = \sum_{h=1}^{H} k_h s_h^2$ respectively.

Optimal values of c_{srs} and c_{str} are calculated by equations (5) and (8) using population data `agpop.csv`, and are estimated by averages of the L estimates $\hat{c}_{srs}^{(i)}$ and $\hat{c}_{str}^{(i)}$ from the ith sample using equations (9) and (11). The ajusted variance estimates of $\mathrm{Var}(\bar{y})$ and $\mathrm{Var}(\bar{y}_{str})$ from the ith sample are $c_{srs}\widehat{\mathrm{Var}}^{(i)}(\bar{y})$, $c_{str}\widehat{\mathrm{Var}}^{(i)}(\bar{y}_{str})$ or $\hat{c}_{srs}\widehat{\mathrm{Var}}^{(i)}(\bar{y})$, $\hat{c}_{str}\widehat{\mathrm{Var}}^{(i)}(\bar{y}_{str})$.

4.2 Simulation results

Table 1 gives simulation results under SRS and stratified random sampling settings. Bias, variance and MSE are calculated by equations (12), (13), (14) and (15). Note

that using population data, we have $V(\bar{y}) = 542599828$ and $\text{Var}(\bar{y}_{str}) = 446220740$. Based on this large scale, bias, variance and MSE of $\widehat{\text{Var}}(\bar{y})$ are all huge.

Table 1 shows that under SRS and stratified random samples: (1) bias of Estimators 2 and 3 are both larger than that of Estimator 1, since Estimator 1 is an unbiased; (2) the trade-off of biased estimators are smaller variance of Estimators 2 and 3 compared to that of Estimator 1; (3) the overall measurement MSE of Estimators 2 and 3 are both smaller than that of Estimator 1. For example, under SRS, the percentage of MSE reduction by Estimator 2 (defined as [MSE of Estimator 1 − MSE of Estimator 2]/MSE of Estimator 1) is $(4.731e+16 - 4.076e+16)/(4.731e+16) = 13.8\%$, and percentage of MSE reduction by Estimator 3 (defined as [MSE of Estimator 1− MSE of Estimator 3]/MSE of Estimator 1) is $(4.731e+16 - 3.156e+16)/(4.731e+16) = 33.3\%$. For stratified random sample, the percentage of MSE reduction by Estimator 2 is 15.1%, and by Estiamtor 3 is 32.7%.

As the percentage of MSE reduction by Estimator 3 is much smaller than that of Estimator 2, we want to take a closer look of Estimator 3. Under SRS, $c = 0.8606$, and $\hat{c}_{srs} = 0.9196$ with standard error of 0.0592. Under stratified SRS, $c_{str} = 0.8538$ and $\hat{c}_{str} = 0.9440$ with standard error of 0.0588. All bias, variance and MSE of Estimator 3 are smaller than those of Estimator 2 under the same setting. Figure 1 shows the histogram of $\hat{c}$ from the $L = 100,000$ simulations under SRS setting. This bimodal histogram shows one peak at 0.85 and another one at 0.95, resulting in an estimate of $\hat{c}_{srs} = 0.9196$.

Table 1: Simulation Results under SRS and stratified random sample settings (variable of interest is *acres92*). Estimators 1, 2, and 3 are variance estimators of mean that are unadjusted, adjusted by c, and adjusted by $\hat{c}$ respectively

Sampling method	SRS			Stratified random sampling		
Estimator	1	2	3	1	2	3
c	na	0.8606	na	na	0.8538	na
$\hat{c}$	na	na	0.9196	na	na	0.9440
Bias	6.876e+03	-7.564e+07	-5.315e+07	1.1490e+06	-6.4230e+07	-3.2259e+07
Variance	4.731e+16	3.504e+16	2.874e+16	3.4425e+16	2.5098e+16	2.2128e+16
MSE	4.731e+16	4.076e+16	3.156e+16	3.4426e+16	2.9224e+16	2.3169e+16

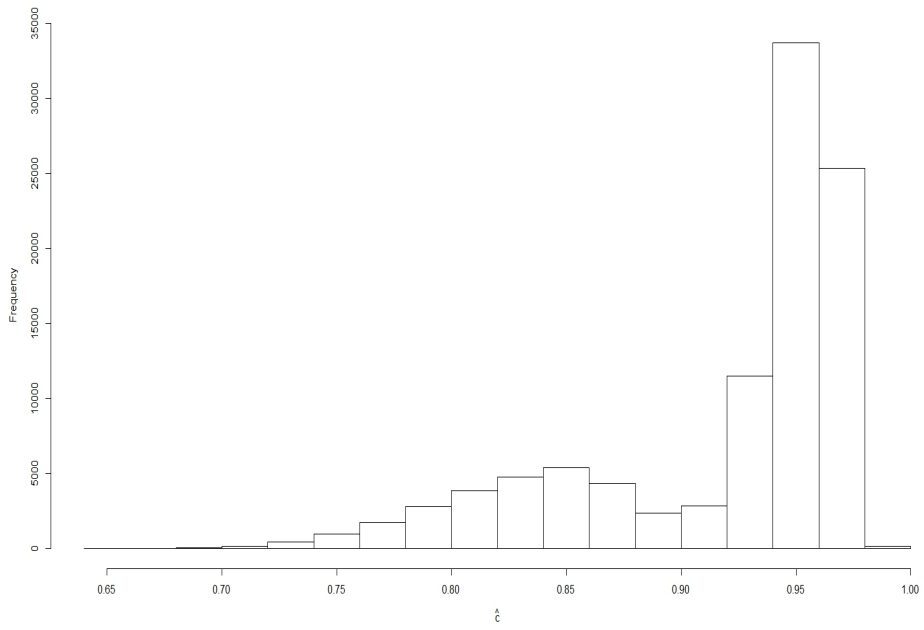

Figure 1: Histogram plot of $\hat{c}_{srs}$ from the 100,000 simulations

Figure 2 shows the sample variance $s^{2(i)}$ versus $\hat{c}_{srs}^{(i)}$ from the ith simulation. Unlike Estimator 2 with a constant adjustment c, $\hat{c}$ seems like a dymamic adjustment with large $\hat{c}$ associated with small s^2 and small $\hat{c}$ associated with large s^2. This makes $\hat{c}s^{2(i)}$ tend to get closer to the true value S^2 and to get closer to each other. Therefore, bias, variance and MSE of Estimator 3 are smaller than those of Estimator 2.

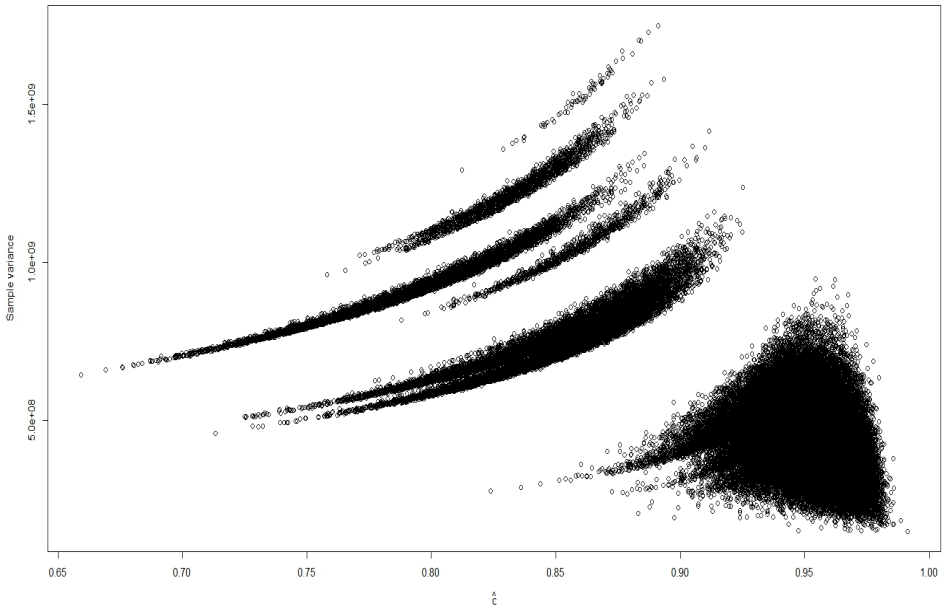

Figure 2: Sample variance versus $\hat{c}_{srs}$ from the 100,000 simulations

5 Conclusions and future study

In this research, we extended Kourouklis (2012)'s work to SRS and stratified random samples. Theoretically, the proposed variance Estimator 2 adjusted by c_{srs} (for SRS) and c_{str} (for stratified samples) has the smallest MSE among the estimators of the

form $c \times$ (variance estimator). In practice, we use sample quantities to estimate the constant c and propose Estimator 3 that is adjusted by $\hat{c}_{srs}$ or $\hat{c}_{str}$. Simulation studies show that the overall measurement MSE of Estimators 2 and 3 are both smaller than that of Estimator 1 (unadjusted estimator). In addtion, $\hat{c}$ acts like dynamic adjustment with large/small $\hat{c}$ associated with small/large variance estimates. As a result, bias, variance and MSE of Estimator 3 are smaller than those of Estimator 2. In practice, we suggest using Estimator 3 to adjust variance estiamtors of mean and total in SRS and stratified random samples as they produce narrower confidence intervals. Future research may extend the adjusted variance estimator to complex surveys such as a two stage stratified cluster survey.

Acknowledgements

The author thanks the referees for their insightful comments and constructive suggestions to improve the manuscript.

References

Brewster, J. F., & Zidek, J. V. (1974). Improving on equivariant estimators. *The Annals of Statistics*, *2*, 21-38.

Brown, L. D. (1968). Inadmissibility of the usual estimators of scale param- eters in

problems with unknown location and scale parameters,. *Annals of Mathematical Statistics*, *39*, 29-48.

Espejo, M., Pineda, M., & Nadarajah, S. (2013). Optimal unbiased estimation of some population central moments. *Metron*, *71*, 39-62.

Heffernan, P. (1997). Unbiased estimation of central moments by using u -statistics. *J. R. Stat. Soc. B*, *59*, 861–863.

Kourouklis, S. (2012). A new estimator of the variance based on minimizing mean squared error. *The American Statistician*, *66*, 234-236.

Lohr, S. (2010). *Sampling: Design and analysis 2nd edition.* Boston, MA: Cengage Learning.

Maruyama, Y.(1998). Minimax estimators of a normal variance. *Metrika*, *48*, 209-214.

Maruyama, Y., & Strawderman, W. E. (2006). A new class of minimax generalized bayes estimators of a normal variance. *Journal of Statistical Planning and Inference*, *136*, 3822-3836.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, *16*, 155-160.

Strawderman, W. E.(1974). Minimax estimation of powers of the variance of a normal population under squared error loss. *The Annals of Statistics*, *2*, 190-198.

Sukhatme, P. V. (1984). *Sampling theory of surveys with applications.* Ames, Iowa: Iowa State College Press.

Yatracos, Y. (2005). Artificially augmented samples, shrinkage, and mean squared error reduction. *Journal of the American Statistical Association*, *100*, 1168-1175.