

# Estimation of Unknown Parameters in Optimum Allocation

Hyeonah Park<sup>a</sup>, Seunghwan Park<sup>a</sup>, Seongryong Na<sup>1,b</sup>

<sup>a</sup>Department of Statistics, Seoul National University

<sup>b</sup>Department of Information and Statistics, Yonsei University

---

## Abstract

The use of pooled standard deviation can reduce the efficiency loss in optimum allocation when strata standard deviations are estimated and several of them are equal. Also shown is that the pooled standard deviation is useful in optimum allocation under a multivariate setting. In addition to theoretical development, we provide the result of simulation study to support the theory.

Keywords: Optimum allocation, pooled standard deviation, proportional allocation.

---

## 1. Introduction

We use optimum allocation to minimize the estimator variance for population parameters under constraints of total survey cost or to minimize cost with a fixed variance. Optimum allocation is directly proportional to the size and the standard deviation for each stratum and inversely proportional to survey costs. Neyman (1934) first proposed the problem of optimum allocation without a cost function. Mahalanobis (1944) imported the concept of cost function and Stuart (1954) verified the optimality using Cauchy inequality.

In practice, there is a limitation on the use of optimum allocation. The efficiency of the estimators from optimum allocation can be decreased when strata standard deviations are absent and subsequently estimated. Alternative allocation methods using other measures (such as the strata range and the strata total of survey variable) instead of standard deviation, have been researched. The power allocation using auxiliary variables and the coefficient of variation of strata was also proposed by Bankier (1988); however, the problem of estimation occurs also in alternative allocations. The variance of an estimator based on the optimum allocation using estimated strata standard deviations has been studied by Evans (1951). The lower bound of the size of preliminary samples to estimate strata standard deviations for optimum allocation was presented by Sukhatme and Sukhatme (1970). Park *et al.* (2007) proposed a method of compromise allocation where optimum allocation using estimated pooled standard deviation is first employed and proportional allocation is used within the combined strata.

In this paper we analyze the optimum allocation using pooled standard deviations among strata and variables. We use the pooled standard deviations in the general case where the number of strata is more than three and the number of the strata with an equal variance is arbitrary. A lower bound of the size of the preliminary sample is also proposed. The use of pooled standard deviation is considered for optimum allocation when the study variable is multivariate. This method proposes a very reasonable

---

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A3003761).

<sup>1</sup> Corresponding author: Professor, Department of Information and Statistics, Yonsei University, 1 Yonsei-daegil, Wonju 220-710, Korea. E-mail: [nasr@yonsei.ac.kr](mailto:nasr@yonsei.ac.kr)

and practical criteria for multivariate optimum allocation that is superior to classical methods (such as optimum and proportional allocations) and that can be verified for specific cases.

This paper is organized as follows. Section 2 uses the pooled standard deviation among strata for optimum allocation and calculate the variance of the estimator using the proposed allocation. Section 3 suggest the use of pooled standard deviation among variables for optimum allocation as well as derives the variance of the estimator of the proposed allocation when the variables have the same variance within each stratum. Section 4 provides the results of the simulation study to investigate the efficiency of the proposed estimators.

## 2. Optimum Allocation Using Pooled Standard Deviation among Strata

Let the parameter of interest be the population mean  $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ , where  $y_i$  is the study variable of unit  $i$  and  $N$  is the size of the finite population. An unbiased estimator under the stratified simple random sampling is  $\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h$ , where  $W_h = N_h/N$  for  $h = 1, \dots, L$ ,  $N_h$  is the population size of stratum  $h$ ,  $L$  is the number of strata,  $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ , and  $y_{hi}$  is the study variable of the  $i^{th}$  unit in the sample of size  $n_h$  from stratum  $h$ . The subject of allocation is of main concern of stratified sampling.

Let  $S_h$  and  $c_h$  denote the population standard deviation of study variable and the survey cost per unit of stratum  $h$  for  $h = 1, \dots, L$ , respectively. The optimum allocation is to decide  $n_h$  so that the variance of  $\bar{y}_{st}$  can be minimized under the condition that the total sample size  $n = \sum n_h$  and the total survey cost  $C = \sum n_h c_h$  are fixed. Note that  $c_h = c$  for all  $h$  is assumed throughout this paper.

In many cases population standard deviations  $S_h$ 's are unknown and subsequently estimated by sample standard deviations that increase the variance of the estimator after optimum allocation using estimated standard deviations. It is true that optimum allocation often reveals poorer efficiency than proportional allocation. Sukhatme and Sukhatme (1970) suggested that the variance of the estimator after optimum allocation with estimated standard deviations and the size of the preliminary samples for the optimum allocation were more efficient than proportional allocation.

We often observe or assume a situation that some of the strata share a common standard deviation that can be taken into account for the estimation of strata standard deviations to reduce the efficiency loss of estimators. Park *et al.* (2007) have considered a simple situation where two of three strata have an equal standard deviation.

In this paper, we study an optimum allocation using pooled standard deviation, where the number of total strata is arbitrary. We assume that there are  $L$  strata and  $S_1 = S_2 = \dots = S_E$  for  $E < L$ . The first  $E$  strata are combined and the pooled estimate for  $S_1$  is provided by

$$\hat{S}_P = \left[ \frac{(n'_1 - 1) \hat{S}_1^2 + (n'_2 - 1) \hat{S}_2^2 + \dots + (n'_E - 1) \hat{S}_E^2}{\sum_{h=1}^E n'_h - E} \right]^{\frac{1}{2}},$$

where  $n'_h$  and  $\hat{S}_h^2$  respectively are the preliminary sample size of stratum  $h$  and the usual sample variance based on the preliminary sample of stratum  $h$  for  $h = 1, \dots, L$ . The allocation equation is given by

$$\begin{aligned} n_h &= n \frac{W_h \hat{S}_P}{\left( \sum_{h'=1}^E W_{h'} \right) \hat{S}_P + \sum_{h'=E+1}^L W_{h'} \hat{S}_{h'}}, & 1 \leq h \leq E, \\ n_h &= n \frac{W_h \hat{S}_h}{\left( \sum_{h'=1}^E W_{h'} \right) \hat{S}_P + \sum_{h'=E+1}^L W_{h'} \hat{S}_{h'}}, & E + 1 \leq h \leq L. \end{aligned} \quad (2.1)$$

The following theorem calculates the variance of the stratified sample mean under the proposed allocation (2.1) in the stratified sampling with replacement.

**Theorem 1.** . Assume that

$$S_1 = S_2 = \dots = S_E = S \quad (2.2)$$

and that the moments of  $\epsilon_h = \hat{S}_h^2 - S_h^2$  higher than 2 can be neglected. Then,

$$\begin{aligned} \text{Var}_{\text{pool}}(\bar{y}_{st}) = n^{-1} & \left\{ \left( \sum_{h=1}^E W_h \right)^2 S^2 + \sum_{h=E+1}^L W_h^2 S_h^2 \right. \\ & + \sum_{h=E+1}^L \sum_{h'=E+1, h \neq h'}^L W_h W_{h'} S_h S_{h'} \left( 1 + \frac{3}{8} D_{h'}^2 - \frac{1}{8} D_h^2 \right) \\ & \left. + \sum_{h'=1}^E W_{h'} \sum_{h=E+1}^L W_h S_h S \left[ 2 + \frac{1}{4} \sum_{j=1}^E \frac{(n'_j - 1)^2}{(\sum_{i=1}^E n'_i - E)^2} D_j^2 + \frac{1}{4} D_h^2 \right] \right\}, \quad (2.3) \end{aligned}$$

where the subscript 'pool' denotes the proposed allocation using  $\hat{S}_P$  and  $D_h = \text{Var}(\hat{S}_h^2)^{1/2}/S_h^2$  for  $1 \leq h \leq L$ .

**Proof:** Note that  $E(\epsilon_h) = 0$ ,  $\text{Var}(\hat{S}_h^2) = E(\epsilon_h^2)$  and  $E(\epsilon_h \epsilon_{h'}) = 0$  for  $h \neq h'$ . Observe that

$$\text{Var}_{\text{pool}}(\bar{y}_{st}) = E \left[ \text{Var}_{\text{pool}}(\bar{y}_{st} | \hat{S}_P, \hat{S}_{E+1}, \dots, \hat{S}_L) \right] + \text{Var} \left[ E_{\text{pool}}(\bar{y}_{st} | \hat{S}_P, \hat{S}_{E+1}, \dots, \hat{S}_L) \right].$$

Replacing  $n_h$  of (2.1), we obtain that

$$\begin{aligned} \text{Var}_{\text{pool}}(\bar{y}_{st} | \hat{S}_P, \hat{S}_{E+1}, \dots, \hat{S}_L) & = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} \\ & = \frac{1}{n} \left\{ \sum_{h=1}^E W_h \hat{S}_P + \sum_{h=E+1}^L W_h \hat{S}_h \right\} \left\{ \sum_{h=1}^E W_h \frac{S_h^2}{\hat{S}_P} + \sum_{h=E+1}^L W_h \frac{S_h^2}{\hat{S}_h} \right\} \\ & = \frac{1}{n} \left\{ \left( \sum_{h=1}^E W_h \right)^2 S^2 + \sum_{h=E+1}^L W_h^2 S_h^2 + \sum_{h=E+1}^L \sum_{h'=E+1, h \neq h'}^L W_h W_{h'} \frac{\hat{S}_h}{\hat{S}_{h'}} S_{h'}^2 \right. \\ & \quad \left. + S^2 \sum_{h'=1}^E W_{h'} \sum_{h=E+1}^L W_h \frac{\hat{S}_h}{\hat{S}_P} + \sum_{h'=1}^E W_{h'} \sum_{h=E+1}^L W_h S_h^2 \frac{\hat{S}_P}{\hat{S}_h} \right\}. \quad (2.4) \end{aligned}$$

Using the Taylor expansion to the extent of the second order, we can see that

$$\begin{aligned} \frac{\hat{S}_h}{\hat{S}_P} & = \frac{S_h}{S} \left\{ 1 + \frac{1}{2S_h^2} \epsilon_h - \frac{1}{8S_h^4} \epsilon_h^2 \right\} \left\{ 1 - \frac{1}{2S^2} \left( \frac{\sum_{j=1}^E (n'_j - 1) \epsilon_j}{\sum_{i=1}^E n'_i - E} \right) + \frac{3}{8S^4} \left( \frac{\sum_{j=1}^E (n'_j - 1) \epsilon_j}{\sum_{i=1}^E n'_i - E} \right)^2 \right\}, \\ \frac{\hat{S}_P}{\hat{S}_h} & = \frac{S}{S_h} \left\{ 1 + \frac{1}{2S^2} \left( \frac{\sum_{j=1}^E (n'_j - 1) \epsilon_j}{\sum_{i=1}^E n'_i - E} \right) - \frac{1}{8S^4} \left( \frac{\sum_{j=1}^E (n'_j - 1) \epsilon_j}{\sum_{i=1}^E n'_i - E} \right)^2 \right\} \left\{ 1 - \frac{1}{2S_h^2} \epsilon_h + \frac{3}{8S_h^4} \epsilon_h^2 \right\} \end{aligned}$$

and for  $E + 1 \leq h, h' \leq L$  and  $h \neq h'$

$$\frac{\hat{S}_h}{\hat{S}_{h'}} = \frac{S_h}{S_{h'}} \left\{ 1 + \frac{1}{2S_h^2} \epsilon_h - \frac{1}{8S_h^4} \epsilon_h^2 \right\} \left\{ 1 - \frac{1}{2S_{h'}^2} \epsilon_{h'} + \frac{3}{8S_{h'}^4} \epsilon_{h'}^2 \right\}.$$

Then, assuming that the moments of  $\epsilon_h$  higher than 2 can be ignored, we have

$$E\left(\frac{\hat{S}_h}{\hat{S}_P}\right) = \frac{S_h}{S} \left( 1 + \frac{3}{8} \sum_{j=1}^E \frac{(n'_j - 1)^2}{(\sum_{i=1}^E n'_i - E)^2} D_j^2 - \frac{1}{8} D_h^2 \right),$$

$$E\left(\frac{\hat{S}_P}{\hat{S}_h}\right) = \frac{S}{S_h} \left( 1 - \frac{1}{8} \sum_{j=1}^E \frac{(n'_j - 1)^2}{(\sum_{i=1}^E n'_i - E)^2} D_j^2 + \frac{3}{8} D_h^2 \right)$$

and for  $E + 1 \leq h, h' \leq L$  and  $h \neq h'$

$$E\left(\frac{\hat{S}_h}{\hat{S}_{h'}}\right) = \frac{S_h}{S_{h'}} \left( 1 + \frac{3}{8} D_{h'}^2 - \frac{1}{8} D_h^2 \right).$$

From (2.4),

$$\begin{aligned} & E\left[\text{Var}_{pool}(\bar{y}_{st} | \hat{S}_P, \hat{S}_{E+1}, \dots, \hat{S}_L)\right] \\ &= n^{-1} \left\{ \left( \sum_{h=1}^E W_h \right)^2 S^2 + \sum_{h=E+1}^L W_h^2 S_h^2 \right. \\ & \quad + \sum_{h=E+1}^L \sum_{h'=E+1, h \neq h'}^L W_h W_{h'} S_{h'} S_h \left( 1 + \frac{3}{8} D_{h'}^2 - \frac{1}{8} D_h^2 \right) \\ & \quad \left. + \sum_{h'=1}^E W_{h'} \sum_{h=E+1}^L W_h S_h S \left[ 2 + \frac{1}{4} \sum_{j=1}^E \frac{(n'_j - 1)^2}{(\sum_{i=1}^E n'_i - E)^2} D_j^2 + \frac{1}{4} D_h^2 \right] \right\}. \end{aligned} \quad (2.5)$$

Finally, we have

$$E_{pool}(\bar{y}_{st} | \hat{S}_P, \hat{S}_{E+1}, \dots, \hat{S}_L) = \bar{Y},$$

which, together with (2.5), implies (2.3).  $\square$

**Remark 1.** Similarly as in Sukhatme and Sukhatme (1970),

$$\text{Var}_{opt}(\bar{y}_{st}) = \frac{1}{n} \left( \sum_{h=1}^L W_h^2 S_h^2 + \sum_h \sum_{h', h < h'}^L W_h W_{h'} S_h S_{h'} \left( 2 + \frac{1}{4} D_h^2 + \frac{1}{4} D_{h'}^2 \right) \right),$$

where  $\text{Var}_{opt}(\bar{y}_{st})$  denotes the variance of the stratified sample mean under the usual optimum allocation using the sample standard deviations. Then, for  $E = 2$  and  $L = 3$ , we obtain that

$$\begin{aligned} \text{Var}_{opt}(\bar{y}_{st}) - \text{Var}_{pool}(\bar{y}_{st}) &= \frac{1}{4n} \left\{ W_1 W_2 S^2 (D_1^2 + D_2^2) + W_1 W_3 S S_3 D_1^2 \left[ 1 - \left( 1 + \frac{W_2}{W_1} \right) \left( 1 + \frac{n'_2 - 1}{n'_1 - 1} \right)^{-2} \right] \right. \\ & \quad \left. + W_2 W_3 S S_3 D_2^2 \left[ 1 - \left( 1 + \frac{W_1}{W_2} \right) \left( 1 + \frac{n'_1 - 1}{n'_2 - 1} \right)^{-2} \right] \right\}. \end{aligned}$$

If  $n'_1 = n'_2 = n'_3$  and  $W_1 = W_2$ , then

$$\text{Var}_{opt}(\bar{y}_{st}) - \text{Var}_{pool}(\bar{y}_{st}) = \frac{1}{8n} \left\{ 2W_1W_2S^2(D_1^2 + D_2^2) + W_3S_3S_3(W_1D_1^2 + W_2D_2^2) \right\} \geq 0.$$

For  $D_h = D$  for  $h = 1, 2, 3$  and  $n'_1 = n'_2 = n'_3$ , we see that

$$\text{Var}_{opt}(\bar{y}_{st}) - \text{Var}_{pool}(\bar{y}_{st}) = \frac{1}{8n} (4W_1W_2S^2 + (W_1 + W_2)W_3S_3S_3)D^2 \geq 0.$$

Thus, if there is reasonable evidence that some strata have a common standard deviation, we can use the pooled standard deviation among strata to increase the efficiency of the stratified sample mean.

**Remark 2.** Assume simply that  $E = 2$  and  $L = 3$ . Then,

$$\begin{aligned} & \text{Var}_{prop}(\bar{y}_{st}) - \text{Var}_{pool}(\bar{y}_{st}) \\ &= \frac{1}{n} \sum_{h=1}^3 W_h (S_h - \bar{S}_w)^2 - \frac{1}{4n} (W_1 + W_2) W_3 S_3 S_3 \left( \frac{(n'_1 - 1)^2}{(n'_1 + n'_2 - 2)^2} D_1^2 + \frac{(n'_2 - 1)^2}{(n'_1 + n'_2 - 2)^2} D_2^2 + D_3^2 \right), \end{aligned}$$

where  $\bar{S}_w = \sum_{h=1}^3 W_h S_h$  and  $\text{Var}_{prop}(\bar{y}_{st})$  is the variance of the stratified sample mean under proportional allocation. It can be seen that if the variations among strata standard deviations are large, our proposed allocation is more efficient than proportional allocation.

**Remark 3.** We assume that  $D_h = D$  and  $n'_h = n'$  for all  $h$ . We can efficiently show that the size of preliminary sample, so that the variance of the estimator under the proposed allocation becomes smaller than that under proportional allocation, is

$$n' > \frac{\beta_2 - 1}{4} \frac{\sum_{h=E+1}^L \sum_{h'=E+1, h \neq h'} W_h W_{h'} S_h S_{h'} + \sum_{h'=1}^E W_{h'} \sum_{h=E+1}^L W_h S_h S_{h'}}{\sum_{h=1}^E W_h (S - \bar{S}_w)^2 + \sum_{h=E+1}^L W_h (S_h - \bar{S}_w)^2},$$

where  $\beta_2 = \mu_4/S^4$ ,  $\mu_4$  is the fourth centered moment and  $S^4$  is the square of variance (cf. chapter 1.10, Sukhatme and Sukhatme, 1970).

### 3. Optimum Allocation Using Pooled Standard Deviations among Variables

In this section we consider the procedure of allocation under multivariate sampling setting. There are  $K$  study variables and we define the population mean of the  $k^{\text{th}}$  variable  $\bar{Y}^k = N^{-1} \sum_{i=1}^N y_i^k$  for  $1 \leq k \leq K$ , where  $y_i^k$  is the  $k^{\text{th}}$  variable of unit  $i$ . Under the stratified sampling technique, an unbiased estimator for  $\bar{Y}^k$  is  $\bar{y}_{st}^k = \sum_{h=1}^L W_h \bar{y}_h^k$ , where  $\bar{y}_h^k$  is the sample mean of stratum  $h$  for the  $k^{\text{th}}$  variable.

We propose the use of pooled standard deviations among variables for the allocation of sample. The pooled standard deviation among variables of stratum  $h$  is defined by

$$\hat{S}_h^{PM} = \left[ \frac{(n_h^{1'} - 1)(\hat{S}_h^1)^2 + (n_h^{2'} - 1)(\hat{S}_h^2)^2 + \dots + (n_h^{K'} - 1)(\hat{S}_h^K)^2}{\sum_{k=1}^K n_h^{k'} - K} \right]^{\frac{1}{2}},$$

where  $n_h^k$  and  $(\hat{S}_h^k)^2$  respectively denote the sample size and the usual sample variance of the preliminary sample of stratum  $h$  for the  $k^{\text{th}}$  variable. The proposed allocation equation is

$$n_h = n \frac{W_h \hat{S}_h^{PM}}{\sum_{h'=1}^L W_{h'} \hat{S}_{h'}^{PM}} \quad (3.1)$$

for  $1 \leq h \leq L$ .

Though this pooling seems artificial, it has the meaning of unifying the  $K$  variances and the problem of allocation becomes simple from the multivariate setting to the univariate one. It can also be understood that the pooled standard deviation is a measure to reflect the total variation about  $K$  variables for each stratum.

The exact calculation of the variance of  $\bar{y}_{st}^k$  is difficult under a general circumstance. We derive its variance for some restricted case in the following theorem. Based on the result, we can ascertain that our proposed allocation can perform better than the usual optimum allocation if the study variables are assumed to have a similar variance within each stratum.

**Theorem 2.** *Assume that*

$$S_h^1 = S_h^2 = \dots = S_h^K = S_h$$

and that the moments of  $\epsilon_h^k = (\hat{S}_h^k)^2 - S_h^2$  higher than 2 can be neglected for  $1 \leq k \leq K$  and  $h = 1, \dots, L$ . Then,

$$\text{Var}_{PM}(\bar{y}_{st}^k) = n^{-1} \left\{ \sum_{h=1}^L W_h^2 S_h^2 + \sum_h \sum_{h \neq h'} W_h W_{h'} S_h S_{h'} \left( 1 + \frac{3}{8} \sum_{l=1}^K (C_h^l)^2 a_l^2 - \frac{1}{8} \sum_{l=1}^K (C_{h'}^l)^2 a_l^2 \right) \right\},$$

where  $C_h^l = (\text{Var}((\hat{S}_h^l)^2))^{1/2} / S_h^2$ ,  $a_l = (n_h^l - 1) / (\sum_{k=1}^K n_h^k - K)$  and the subscript 'PM' means the proposed allocation.

**Proof:** Note that  $E(\epsilon_h^k) = 0$ ,  $\text{Var}[(\hat{S}_h^k)^2] = E[(\epsilon_h^k)^2]$ ,  $E(\epsilon_h^l \epsilon_h^k) = 0$  for  $l \neq k$  and  $E(\epsilon_h^l \epsilon_{h'}^l) = 0$  for  $h \neq h'$ . Observe now that

$$\text{Var}_{PM}(\bar{y}_{st}^k) = E \left[ \text{Var}_{PM}(\bar{y}_{st}^k | \hat{S}_1^{PM}, \dots, \hat{S}_L^{PM}) \right] + \text{Var} \left[ E_{PM}(\bar{y}_{st}^k | \hat{S}_1^{PM}, \dots, \hat{S}_L^{PM}) \right].$$

If fpc is neglected and (3.1) is used for  $n_h$ , then we have

$$\begin{aligned} \text{Var}_{PM}(\bar{y}_{st}^k | \hat{S}_1^{PM}, \dots, \hat{S}_L^{PM}) &= \sum_{h=1}^L W_h \frac{S_h^2}{n_h} \\ &= \frac{1}{n} \sum_{h=1}^L W_h^2 \frac{S_h^2}{\hat{S}_h^{PM}} \sum_{h'=1}^L W_{h'} \hat{S}_{h'}^{PM} \\ &= \frac{1}{n} \left( \sum_{h=1}^L W_h^2 S_h^2 + \sum_h \sum_{h \neq h'} W_h W_{h'} S_h^2 \frac{\hat{S}_{h'}^{PM}}{\hat{S}_h^{PM}} \right). \end{aligned} \quad (3.2)$$

Note that  $\hat{S}_h^{PM} = S_h (1 + \sum_{i=1}^K a_i \epsilon_h^i / S_h^2)^{1/2}$  for  $a_i = (n_h^i - 1) / (\sum_{k=1}^K n_h^k - K)$ . By the Taylor expansion up to the second order, it can be obtained that

$$\frac{\hat{S}_{h'}^{PM}}{\hat{S}_h^{PM}} = \frac{S_{h'}}{S_h} \left( 1 + \frac{\sum_{l=1}^K a_l \epsilon_{h'}^l}{2S_{h'}^2} - \frac{(\sum_{l=1}^K a_l \epsilon_{h'}^l)^2}{8S_{h'}^4} \right) \left( 1 - \frac{\sum_{l=1}^K a_l \epsilon_h^l}{2S_h^2} + \frac{3(\sum_{l=1}^K a_l \epsilon_h^l)^2}{8S_h^4} \right).$$

By the negligibility of the moments of  $\epsilon_h$ ,

$$\begin{aligned} E\left(\frac{\hat{S}_{h'}^{PM}}{\hat{S}_h^{PM}}\right) &= \frac{S_{h'}}{S_h} \left(1 - \frac{\sum_{l=1}^K a_l^2 E(\epsilon_{h'}^l)^2}{8S_{h'}^4} + \frac{3 \sum_{l=1}^K a_l^2 E(\epsilon_h^l)^2}{8S_h^4}\right) \\ &= \frac{S_{h'}}{S_h} \left(1 + \frac{3}{8} \sum_{l=1}^K (C_h^l)^2 a_l^2 - \frac{1}{8} \sum_{l=1}^K (C_{h'}^l)^2 a_l^2\right). \end{aligned}$$

Then, from (3.2),

$$\begin{aligned} &E\left[\text{Var}_{PM}(\bar{y}_{st}^k | \hat{S}_1^{PM}, \dots, \hat{S}_L^{PM})\right] \\ &= \frac{1}{n} \left\{ \sum_{h=1}^L W_h^2 S_h^2 + \sum_h \sum_{h \neq h'} W_h W_{h'} S_h S_{h'} \left(1 + \frac{3}{8} \sum_{l=1}^K (C_h^l)^2 a_l^2 - \frac{1}{8} \sum_{l=1}^K (C_{h'}^l)^2 a_l^2\right) \right\}, \end{aligned}$$

which implies Theorem 2 since  $E_{PM}(\bar{y}_{st}^k | \hat{S}_1^{PM}, \dots, \hat{S}_L^{PM}) = \bar{Y}^k$ . □

**Remark 4.** Set  $n_h'^1 = n_h'^2 = \dots = n_h'^K$  to compare with the equation of variance  $\text{Var}_{opt}(\bar{y}_{st}^k)$  of Sukhatme and Sukhatme (1970). Then,  $C_h^l = C_h$  for all  $l$  and

$$\text{Var}_{opt}(\bar{y}_{st}^k) - \text{Var}_{PM}(\bar{y}_{st}^k) = \frac{1}{4} \left(1 - \frac{1}{K}\right) \sum_h \sum_{h < h'} W_h W_{h'} S_h S_{h'} (C_h^2 + C_{h'}^2).$$

We show that there exists a proper condition where the proposed allocation has better efficiency than the usual optimum allocation using estimated standard deviations.

#### 4. Simulation Results

In this section we perform a simulation study to test our theory. The number of repetition performed for each setting is  $B = 10,000$  and the sample size  $n$  is 100. First, we test whether the proposed allocation (using the pooled standard deviation among strata) is superior to other allocations such as the optimum allocation where standard deviations are estimated separately over all strata and the proportional allocation. The value for the preliminary sample size  $n'_1$  is 5, and the values for  $n'_h$ ,  $2 \leq h \leq 4$  are 5, 10, 30 and 50. The samples are generated by

$$\begin{aligned} y_{1i} &= S_1 z_{1i} + \bar{Y}_1 \\ y_{2i} &= S_2 z_{2i} + \bar{Y}_2 \\ y_{3i} &= S_3 z_{3i} + \bar{Y}_3 \\ y_{4i} &= S_4 z_{4i} + \bar{Y}_4, \end{aligned}$$

where  $z_{hi} \sim N(0, 1)$  for  $1 \leq i \leq n_h$  and  $1 \leq h \leq 4$  and all  $z_{hi}$  are independent. We compute the empirical values of four types of variances  $\text{Var}_{pool}(\bar{y}_{st})$ ,  $\text{Var}_{opt}(\bar{y}_{st})$ ,  $\text{Var}_{prop}(\bar{y}_{st})$  and  $\text{Var}_{top}(\bar{y}_{st})$ , where  $\text{Var}_{top}(\bar{y}_{st})$  is the variance under the optimum allocation when strata standard deviations of population are known.

Each cell in Table 1 contains three values of relative efficiency respectively corresponding to  $\text{Var}_{opt}(\bar{y}_{st})/\text{Var}_{pool}(\bar{y}_{st})$ ,  $\text{Var}_{prop}(\bar{y}_{st})/\text{Var}_{pool}(\bar{y}_{st})$  and  $\text{Var}_{opt}(\bar{y}_{st})/\text{Var}_{top}(\bar{y}_{st})$  in this order. The population weights are  $W_1 = 0.4$ ,  $W_2 = 0.3$ ,  $W_3 = 0.2$  and  $W_4 = 0.1$ . The population means are given by

Table 1: Empirical values for  $\text{Var}_{opt}(\bar{y}_{st})/\text{Var}_{pool}(\bar{y}_{st})$ ,  $\text{Var}_{prop}(\bar{y}_{st})/\text{Var}_{pool}(\bar{y}_{st})$ ,  $\text{Var}_{opt}(\bar{y}_{st})/\text{Var}_{top}(\bar{y}_{st})$ 

$n'_2 = n'_3 = n'_4$	$S_4/S = 2$	$S_4/S = 4$	$S_4/S = 6$	$S_4/S = 8$
5	1.098	1.085	1.075	1.068
	1.014	1.379	1.855	2.335
	1.151	1.140	1.147	1.131
10	1.063	1.055	1.051	1.047
	1.038	1.421	1.918	2.416
	1.090	1.076	1.085	1.072
30	1.050	1.048	1.043	1.039
	1.048	1.441	1.947	2.459
	1.065	1.055	1.061	1.046
50	1.049	1.045	1.042	1.038
	1.050	1.444	1.952	2.465
	1.062	1.049	1.057	1.042

Table 2: Empirical values for  $\text{Var}_{opt}(\bar{y}_{st}^k)/\text{Var}_{PM}(\bar{y}_{st}^k)$ ,  $\text{Var}_{prop}(\bar{y}_{st}^k)/\text{Var}_{PM}(\bar{y}_{st}^k)$ ,  $\text{Var}_{opt}(\bar{y}_{st}^k)/\text{Var}_{top}(\bar{y}_{st}^k)$  for  $k = 1$ 

$n'_2, n'_3$	$S_2 = 10, S_3 = 20$	$S_2 = 20, S_3 = 30$	$S_2 = 30, S_3 = 40$	$S_2 = 40, S_3 = 50$
5, 10	1.066	1.066	1.062	1.062
	1.084	1.163	1.256	1.330
	1.093	1.092	1.086	1.087
10, 20	1.046	1.042	1.038	1.034
	1.092	1.173	1.267	1.343
	1.064	1.058	1.053	1.047
20, 30	1.040	1.035	1.029	1.027
	1.095	1.177	1.271	1.347
	1.055	1.048	1.040	1.037
30, 50	1.038	1.032	1.026	1.024
	1.096	1.178	1.273	1.349
	1.052	1.043	1.036	1.033

$\bar{Y}_1 = 10$ ,  $\bar{Y}_2 = 5$ ,  $\bar{Y}_3 = 40$  and  $\bar{Y}_4 = 20$ . The value for  $S = S_1 = S_2 = S_3$  is fixed as 10 but  $S_4$  takes various values.

Table 1 shows that the estimator using the proposed allocation is more efficient than the estimators using the usual optimum allocation and the proportional allocation. We can see that both the proposed allocation and the optimum allocation are superior to the proportional allocation when the difference between strata standard deviations becomes large. It can also be seen that the efficiency of the proposed allocation nearly becomes the same as that of the optimum allocation with known strata standard deviations as the preliminary sample size and the difference of strata standard deviations become large.

Second, we simulate the use of pooled standard deviations among variables. There are three study variables and three strata. The value for  $n'_1$  is 5 and the values for  $(n'_2, n'_3)$  are (5, 10), (10, 20), (20, 30) and (30, 50). In each stratum, all variables have the same population variance, that is,  $S_1^1 = S_1^2 = S_1^3 = S_1$ ,  $S_2^1 = S_2^2 = S_2^3 = S_2$  and  $S_3^1 = S_3^2 = S_3^3 = S_3$ . The population weights are  $W_1 = 0.45$ ,  $W_2 = 0.35$  and  $W_3 = 0.2$ . The sample for the  $k^{th}$  variable in stratum  $h$  is generated by

$$y_{hi}^k = S_h z_{hi}^k + m_h^k,$$

where  $z_{hi}^k \sim N(0, 1)$  for  $i = 1, \dots, n_h$ ,  $h = 1, 2, 3$  and  $k = 1, 2, 3$  and all  $z_{hi}^k$ 's are independent. We set  $(m_1^1, m_1^2, m_1^3) = (15, 20, 30)$ ,  $(m_2^1, m_2^2, m_2^3) = (5, 5, 20)$  and  $(m_3^1, m_3^2, m_3^3) = (10, 15, 40)$ . The value for  $S_1$  is fixed as 10 but  $(S_2, S_3)$  takes various values. We compute four empirical variances  $\text{Var}_{PM}(\bar{y}_{st}^k)$ ,  $\text{Var}_{opt}(\bar{y}_{st}^k)$ ,  $\text{Var}_{prop}(\bar{y}_{st}^k)$  and  $\text{Var}_{top}(\bar{y}_{st}^k)$  for variable  $k$ .



Each cell in Table 2 contains three values of relative efficiency respectively corresponding to  $\text{Var}_{opt}(\bar{y}_{st}^k)/\text{Var}_{PM}(\bar{y}_{st}^k)$ ,  $\text{Var}_{prop}(\bar{y}_{st}^k)/\text{Var}_{PM}(\bar{y}_{st}^k)$  and  $\text{Var}_{opt}(\bar{y}_{st}^k)/\text{Var}_{topt}(\bar{y}_{st}^k)$  for  $k = 1$  in this order. We remark that the relative efficiency for  $k = 2, 3$  is almost the same. As anticipated, Table 2 shows that our proposed allocation using pooled standard deviations among variables outperforms conventional optimum allocation and proportional allocation.

From a theoretical viewpoint, it seems that the size of the preliminary sample should be appropriately large so that  $\epsilon_h$  may be neglected at the higher moment. Nevertheless, our simulation result shows that the proposed allocation is superior to other methods even for a moderate size of the preliminary sample. We can summarize that the stratified sample mean will increase in efficiency if variances are equal over strata or variables and pooled standard deviations are used for the allocation of samples. We suggest the use of the proposed allocation based on the pooled standard deviations if there is a sufficient evidence that the population standard deviations are equal for several strata or variables.

## References

- Bankier, M. D. (1988). Power allocations: Determining sample sizes for subnational areas, *The American Statistician*, **42**, 174–177.
- Evans, W. D. (1951). On stratification and optimum allocations, *Journal of the American Statistical Association*, **46**, 95–104.
- Mahalanobis, D. C. (1944). On large-scale sample surveys, *Philosophical Transactions of the Royal Society of London. Series B*, **231**, 329–451.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of purposive selection, *Journal of the Royal Statistical Society. Series B*, **97**, 558–606.
- Park, H., Na, S. and Jeon, J. (2007). Compromise Allocation in Univariate Stratified Sampling, *Communications in Statistics: Theory and Methods*, **36**, 265–271.
- Stuart, A. (1954). A sample presentation of optimum sampling results, *Journal of the Royal Statistical Society. Series B*, **16**, 239–241.
- Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*, 2nd edition, Iowa State University Press.