Communications for Statistical Applications and Methods
2013, Vol. 20, No. 2, 97–105

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.2.097

Simple Compromise Strategies in Multivariate Stratification

Inho Park[1,a]

[a]Department of Statistics, Pukyong National University

Abstract

Stratification (among other applications) is a popular technique used in survey practice to improve the accuracy of estimators. Its full potential benefit can be gained by the effective use of auxiliary variables in stratification related to survey variables. This paper focuses on the problem of stratum formation when multiple stratification variables are available. We first review a variance reduction strategy in the case of univariate stratification. We then discuss its use for multivariate situations in convenient and efficient ways using three methods: compromised measures of size, principal components analysis and a K-means clustering algorithm. We also consider three types of compromising factors to data when using these three methods. Finally, we compare their efficiency using data from MU281 Swedish municipality population.

Keywords: Stratum boundaries, sample allocation, principal components analysis, measure of size, K-means clustering algorithm.

1. Introduction

Assume that a stratified simple random sample of n units is selected from a population of N units with forming H strata U_h of size N_h, where $h = 1, 2, \ldots, H$. As often the case in practice, a single variable x_1 say, is assumed to be chosen for stratification as the main indicator of the variables of study interest for the corresponding stratified mean estimator

$$\bar{x}_{1,str} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{x}_{1,h} \tag{1.1}$$

to have its variance

$$V_{str}(\bar{x}_{1,str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{n_h} - \frac{\delta_h}{N_h}\right) \sum_{i \in U_h} \frac{\left(x_{1,hi} - \bar{X}_{1,h}\right)^2}{(N_h - \delta_h)} \tag{1.2}$$

as small as possible, where δ_h is the indicator of without-replacement sampling for stratum h, $x_{1,hi}$ is the value of x_1 for unit (hi), and $\bar{x}_{1,h} = n_h^{-1} \sum_{i=1}^{n_h} x_{1,hi}$ and $\bar{X}_{1,h} = N_h^{-1} \sum_{i=1}^{N_h} x_{1,hi}$ are the stratum sample and population means, respectively. Expression (1.2) indicates that the efficiency of the stratified random sampling design for $\bar{x}_{1,str}$ depends on stratum formation (U_h) and sample allocation (n_h). This is because the former determines the breaks of x_1-homogeneity $(S^2_{1,h}(\delta_h))$ and the sizes (N_h) of strata and the latter the sample information shares $(a_h = n_h/n)$ among strata, where $S^2_{1,h}(\delta_h) = (N_h - \delta_h)^{-1} \sum_{i \in U_h} (x_{1,hi} - \bar{X}_{1,h})^2$ denotes the stratum variance of x_1.

[1] Assistant Professor, Department of Statistics, Pukyong National University, 599-1 Daeyeon 3-Dong, Nam-Gu, Busan 608-737, Korea. E-mail: ipark@pknu.ac.kr

To better understand the efficiency gain of (univariate) stratification over simple random sampling, some simplification is made with the following two design options:

(D1) with-replacement selection ($\delta_h = 0$);

(D2) proportional allocation ($a_h = n_h/n = N_h/N$).

Let $\sigma^2_{1,h} = S^2_{1,h}(0) = N_h^{-1} \sum_{i \in U_h} (x_{1,hi} - \bar{X}_{1,h})^2$ denotes the stratum variance of x_1 when the within-stratum sample selection is with-replacement. Under (D1) and (D2), the variance of $\bar{x}_{1,str}$ can be written from (1.2) as

$$V_{str}\left(\bar{x}_{1,str}|D1, D2\right) = \frac{1}{n} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \sigma^2_{1,h}. \tag{1.3}$$

If a with-replacement simple random sample of the same size n is selected without stratification, then the variance of a simple mean estimator $\bar{x}_{1,srs} = n^{-1} \sum_{i=1}^{n} x_{1,i}$ is given as

$$V_{srs}\left(\bar{x}_{1,srs}|D1\right) = \frac{1}{n} \sigma^2_1, \tag{1.4}$$

where $\sigma^2_1 = N^{-1} \sum_{i \in U} (x_{1,i} - \bar{X}_1)^2$ denotes the population variance of x_1. Then, the decomposition of the population variance (*e.g.*, Särndal *et al.*, 1992, p.109)

$$\sigma^2_1 = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \sigma^2_{1,h} + \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \left(\bar{X}_{1,h} - \bar{X}_1\right)^2,$$

or alternatively from (1.3) and (1.4),

$$V_{srs}\left(\bar{x}_{1,srs}|D1\right) = V_{str}\left(\bar{x}_{1,str}|D1, D2\right) + \frac{1}{n} \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \left(\bar{X}_{1,h} - \bar{X}_1\right)^2$$

indicates that the smaller the weighted within-stratum variances (thus the larger the dispersion among the stratum means), the larger the gain in precision due to stratification over simple random sampling that achieves the reduction of $n^{-1} \sum_{h=1}^{H} (N_h/N)(\bar{X}_{1,h} - \bar{X}_1)^2 \geq 0$ under (D1) and (D2).

In multivariate sampling it is often required to stratify the population with respect to more than one stratification variable (*i.e.*, stratifier). Thus, a compromise must be reached to form strata that are efficient for as many stratifiers as possible. In this paper, we first discuss two simple ways to develop a single (compromised) stratifier from multiple stratifiers for use in line with the above-mentioned variance reduction strategy for univariate stratification: one based on compromised size measures in Section 2 and the other on the principal components analysis in Section 3. Second, we discuss the use of the K-means clustering algorithm for multivariate stratification to form strata in a way to minimize the sum of multiple variances $V_{str}(\bar{x}_{q,str})$ for $q = 1, 2, \ldots, Q$ in Section 4. The K-means clustering algorithm can be readily applicable, since it is available through many statistical software. Both stratification methods (based on multivariate techniques) allow the principal components analysis and the K-means clustering algorithm are often applied to multivariate stratification in the sampling literature (see, *e.g.*, Jarque, 1981). Together with the above three stratification methods, we also consider three types of compromising factors to data to reflect their relative importance when stratifying the population that includes equal-compromise, scale-compromise and size-compromise. Section 5 compares several stratification strategies (*i.e.*, pairs of a stratification method and a type of compromising factors for a given sample allocation rule) based on a number of efficiency measures using data from MU281 Swedish municipality population. Section 6 includes a brief discussion about our findings.

2. Compromised Measure of Size

Assume that $\mathbf{x} = (x_1, x_2, \ldots, x_Q)'$ denotes a set of Q stratifiers with their relative importance or compromising factors $\mathbf{c} = (c_1, c_2, \ldots, c_Q)'$ in multivariate stratification. One of the simplest ways of compromising the Q stratifiers to be used with the variance reduction strategy in Section 1 is to construct a measure of size (MOS) in the form

$$y_L = \sum_{q=1}^{Q} c_q x_q.$$

The linear combination y_L (L-CMOS for brevity) can be rewritten as a simple sum of transformed variables $z_q = c_q x_q$ given as

$$y_L = \mathbf{1}'\mathbf{z} = \sum_{q=1}^{Q} z_q, \tag{2.1}$$

where $\mathbf{1}$ is a $Q \times 1$ vector of ones, $\mathbf{z} = (z_1, z_2, \ldots, z_Q)'$ is a vector of the Q transformed variables and $q = 1, 2, \ldots, Q$. Let $z_{q,hi}$ and $y_{L,hi}$ denote the values of z_q and y_L for unit (hi), respectively, where $h = 1, 2, \ldots, H$ and $i = 1, 2, \ldots, N_h$. Further, define $\bar{y}_{L,str} = \sum_{h=1}^{H}(N_h/N)\bar{y}_{L,h}$ and $\bar{y}_{L,h} = n_h^{-1}\sum_{i=1}^{n_h} y_{L,hi}$ as the stratified mean estimator and the stratum sample mean of y_L, respectively. Then, strata U_h can be formed straightforwardly by minimizing the objective function $V_L(U_h) = V_{str}(\bar{y}_{L,str})$, where $V_{str}(\bar{y}_{L,str})$ is the variance of $\bar{y}_{L,str}$ defined similarly as in (1.2) for y_L. See, *e.g.*, Kozak and Verma (2006) for a univariate algorithm that determines optimal stratum boundaries for a given objective function.

The objective function $V_L(U_h)$ for the optimization can also be written from (2.1) as

$$V_L(U_h) = \sum_{q=1}^{Q} V_{str}\left(\bar{z}_{q,str}\right) + \sum_{q \neq q'} \mathrm{Cov}_{str}\left(\bar{z}_{q,str}, \bar{z}_{q',str}\right), \tag{2.2}$$

where $V_{str}(\bar{z}_{q,str})$ is the variance of $\bar{z}_{q,str} = c_q\bar{x}_{q,str}$ and $\mathrm{Cov}_{str}(\bar{z}_{q,str}, \bar{z}_{q',str})$ is the covariance between $\bar{z}_{q,str}$ and $\bar{z}_{q',str}$ for $q \neq q' = 1, 2, \ldots, Q$. We see that y_L involves a nuisance of $(Q-1) \times (Q-1)$ covariance terms $\mathrm{Cov}_{str}(\bar{z}_{str,q}, \bar{z}_{str,q'})$ in the optimization for constructing strata, which would be negligible when z_q are all near independent.

Another form for a compromised measure of size in multivariate stratification is the Euclidean norm of $\mathbf{z}$ given as

$$y_E = \sqrt{\mathbf{z}'\mathbf{z}} = \left(\sum_{q=1}^{Q} z_q^2\right)^{\frac{1}{2}}.$$

Since y_E is not linear in z_q, neither is the variance $V_E = V_{str}(\bar{y}_{E,str})$ of the corresponding stratified mean estimator $\bar{y}_{E,str}$ in $V_{str}(\bar{z}_{q,str})$ and $\mathrm{Cov}_{str}(\bar{z}_{q,str}, \bar{z}_{q',str})$. However, it is interesting to see that y_E can serve as a measure of size to determine the selection probability adoptable for multivariate with-replacement probability proportionate to size (*i.e.*, pps) sampling that minimizes the sum of Q variances $V_{pps}(\bar{z}_{q,pps}|D1)$, where $\bar{z}_{q,pps} = (nN)^{-1}\sum_{i=1}^{n}(z_{q,i}/p_i)$ is the pps mean estimator for z_q with the selection probability $p_i \geq 0$ such that $\sum_{i=1}^{N} p_i = 1$ and its variance

$$V_{pps}\left(\bar{z}_{q,pps}|D1\right) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{z_{q,i}}{Np_i} - \bar{Z}_q\right)^2. \tag{2.3}$$

If the selection probability p_i^* is set proportional to $y_{E,i} = (\sum_{q=1}^{Q} z_{q,i}^2)^{1/2}$, then

$$p_i^* = \arg\min_{p_i} \sum_{q=1}^{Q} V_{pps}\left(\bar{z}_{q,pps}|D1\right). \tag{2.4}$$

See Malec (1995) for the proof.

A gain in efficiency from using y_E, however, may not be greater with the stratified simple random sampling as a stratifier than with the *pps* sampling as a measure of size. Both *pps* sampling and stratified random sampling are concurrent design options based upon the availability of quantitative auxiliary information at the individual unit-level; however, their efficiency gains are achieved via different structural forms in the sampling design. The *pps* sampling takes into account the inherent variation in the their individual values directly to the associated selection probabilities but the stratified random sampling relies on the homogeneous grouping of the population units as discussed in Section 1. For example, in the case of univariate situation ($Q = 1$), p_i^* in (2.4) is proportional to $z_{1,i}$ and thus reduction in $V_{pps}(\bar{z}_{1,pps}|D1)$ in (2.3) is expected since the difference among $z_{1,i}/p_i^*$ vanishes. See Ardilly and Tillé (2006) for further discussion on the concurrency of the two sampling design options for various univariate situations.

One may generalize a construction of a compromised measure of size through a form $y_\phi = \phi(\mathbf{z}) = \phi_{\mathbf{c}}(\mathbf{x})$ for a given pair of a compromise function $\phi : R^Q \to R$ and a type of compromising factors $\mathbf{c} \in R^Q$. By letting $y_{\phi,hi} = \phi(\mathbf{z}_{hi})$ denote the value of the compromised measure of size $y_\phi = \phi(\mathbf{z})$ for unit (hi), one can form strata by minimizing the objective function $V_\phi(U_h) = V_{str}(\bar{y}_{\phi,str})$, where the resulting stratified mean estimator $\bar{y}_{\phi,str}$ of y_ϕ and its variance $V_{str}(\bar{y}_{\phi,str})$ are defined similarly as in (1.1) and (1.2), respectively.

3. First Principal Component

In the principal components analysis, one seeks to maximize the variance of a linear combination of the variables. Let $\Sigma_{\mathbf{z}}$ denote the $Q \times Q$ variance-covariance matrix of $\mathbf{z}$ in the population U. Then, the first principal component (FPC for short) is defined as

$$y_1 = \mathbf{a}_1'\mathbf{z} = a_{11}z_1 + a_{12}z_2 + \cdots + a_{1Q}z_Q, \tag{3.1}$$

that maximizes $V(y_1) = \mathbf{a}_1'\Sigma_{\mathbf{z}}\mathbf{a}_1$ for a vector $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1Q})'$ such that $\mathbf{a}_1'\mathbf{a}_1 = 1$. The qth principal component is defined as $y_q = \mathbf{a}_q'\mathbf{z}$ that maximizes $V(y_q) = \mathbf{a}_q'\Sigma_{\mathbf{z}}\mathbf{a}_q$ for $\mathbf{a}_q = (a_{q1}, a_{q2}, \ldots, a_{qQ})'$ such that $\mathbf{a}_q'\mathbf{a}_q = 1$, $\mathbf{a}_q'\mathbf{a}_{q'} = 0$ and $\mathrm{Cov}(y_q, y_{q'}) = \mathbf{a}_q'\Sigma_{\mathbf{z}}\mathbf{a}_{q'} = 0$, where $q > q' = 1, 2, \ldots, Q$. Using the multivariate theory, one can show that

$$\mathrm{Var}(y_q) = \mathbf{a}_q'\Sigma_{\mathbf{z}}\mathbf{a}_q = \lambda_q,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_Q \geq 0$ are the eigenvalues of $\Sigma_{\mathbf{z}}$ with their corresponding eigenvector $\mathbf{a}_q = (a_{q1}, a_{q2}, \ldots, a_{qQ})'$ for $q = 1, 2, \ldots, Q$. See, for example, Mardia *et al.* (1980) for the details.

The FPC has been adopted in the literature along with the variance reduction strategy in Section 1 for use in multivariate stratification. See, for example, Hagood and Bernart (1945), Golder and Yeomans (1973), Kish and Anderson (1978) and Jarque (1981). Following Golder and Yeomans (1973), strata can be formed by minimizing $V_F(U_h) = V_{str}(\bar{y}_{1,str})$, where $\bar{y}_{1,str}$ is the stratified mean estimator for y_1 with its variance $V_{str}(\bar{y}_{1,str})$ given as

$$V_F(U_h) = \sum_{q=1}^{Q} a_{1q}^2\, V_{str}\left(\bar{z}_{q,str}\right) + \sum_{q \neq q'} a_{1q}a_{1q'}\, \mathrm{Cov}_{str}\left(\bar{z}_{q,str}, \bar{z}_{q',str}\right), \tag{3.2}$$

where $V_{str}(\bar{z}_{q,str})$ and $\mathrm{Cov}_{str}(\bar{z}_{q,str}, \bar{z}_{q',str})$ are defined as in (2.2). The purpose of using y_1 is to provide the maximal parsimonious summarization of the variance-covariance structure of $\mathbf{z}$ by extracting a single linear combination. As compared with $V_L(U_h)$ in (2.2), the objective function $V_F(U_h)$ in (3.2) involves extra coefficients (*i.e.*, loadings) a_{1q}'s for transforming the original variables $\mathbf{z}$ but common in the inclusion of the covariance terms $\mathrm{Cov}_{str}(\bar{z}_{q,str}, \bar{z}_{q',str})$. Both objective functions V_F and V_L become equivalent when all a_{1q}'s are set to a constant; therefore, when the first principal component is the simple average of $\mathbf{z}$. As an effort for better explanation of the variability among the variables, Samita and Kumari (2006) further consider an aggregation of a number of principal components to construct a measure of size. The efficiency gain due to such an aggregation may not be fruitful with respect to individual stratifiers due to the involvement of the covariance terms in the resulting objective function for the optimization.

In developing y_1 in (3.1), a common practice is to use standardized data, that is, $\mathbf{c} = (\sigma_1^{-1}, \ldots, \sigma_Q^{-1})'$ without the loss of generality. See, for example, Kish and Anderson (1978) and Jarque (1981). Jarque (1981) also considers to use the size-adjusted data of compromising factors $\mathbf{c} = (X_1^{-1}, \ldots, X_Q^{-1})'$ to the original data.

4. K-means Clustering Algorithm

The K-means clustering algorithm was adopted to solve the problem of multivariate stratification in the literature. See, *e.g.*, Golder and Yeomans (1973) and Jarque (1981). In order to form H (or K as indicated in its nomenclature) strata, the objective function of the K-means clustering algorithm can be written as

$$V_K(U_h) = \sum_{q=1}^{Q} V_{str}\left(\bar{z}_{q,str} | D1, D2\right). \tag{4.1}$$

When $c_q \equiv 1$ for all q's, the objective function becomes

$$V_K\left(U_h | c_q \equiv 1\right) = \sum_{q=1}^{Q} V_{str}\left(\bar{x}_{q,str} | D1, D2\right),$$

or alternatively as commonly seen in the cluster analysis literature (*e.g.*, Jain *et al.*, 1999) as

$$V_K\left(U_h | c_q \equiv 1\right) \propto \sum_{h=1}^{H} \sum_{q=1}^{Q} N_h \sigma_{q,h}^2 = \sum_{h=1}^{H} \sum_{i \in U_h} \left\| \mathbf{x}_{hi} - \bar{\mathbf{X}}_h \right\|^2,$$

where $\|\mathbf{x}_{hi} - \bar{\mathbf{X}}_h\| = \{\sum_{q=1}^{Q} (x_{q,hi} - \bar{X}_{q,h})^2\}^{1/2}$ denotes the Euclidean distance between two vectors $\mathbf{x}_{hi} = (x_{1,hi}, \ldots, x_{Q,hi})'$ and $\bar{\mathbf{X}}_h = (\bar{X}_{1,h}, \ldots, \bar{X}_{Q,h})'$.

As seen from (4.1), the K-means clustering algorithm does not contain any covariance terms in the objective function. Therefore, as compared to those of the preceding two methods based on compromised measures of sizes y_L and y_1, one may anticipate that the effect of the minimization by the K-means clustering algorithm is more direct on the Q individual variances $V_{str}(\bar{z}_{q,str})$ due to its exclusion of covariance terms from the minimization process. It should be noted that $V_L(U_h) = V_K(U_h)$ when z_q's are mutually uncorrelated, to form the equivalent strata. The biggest advantage of the use of equations (2.1) and (3.1) is that stratification can be simply accomplished using a single stratifier, that of the K-means clustering algorithm that is readily used with the existing statistical software such as R and SAS.

Table 1: Summary statistics of MU281 variables for stratification.

Variable	Name	Statistics			
		$\bar{X}_q$	σ_q	$CV_q(\%)$	γ_q
x_1	P75	24.263	23.244	95.799	2.273
x_2	SS82	22.039	7.135	32.375	0.703
x_3	S82	47.178	10.415	22.076	1.160

Table 2: Relative sizes of compromising factors for three MU281 variables.

	Compromising Factors (c_q)	$\frac{c_1}{\sum_q c_q}$	$\frac{c_2}{\sum_q c_q}$	$\frac{c_3}{\sum_q c_q}$
Equal-compromise	1	0.333	0.333	0.333
Scale-compromise	$1/\sigma_q$	0.154	0.502	0.344
Size-compromise	$1/X_q$	0.382	0.421	0.197

Table 3: Loadings of the first principal component.

	Compromising Factor (c_q)	a_{11}	a_{12}	a_{13}	% of total variance
Equal-compromise	1	−0.905	−0.205	−0.373	72.928
Scale-compromise	$1/\sigma_q$	−0.583	−0.550	−0.599	63.223
Size-compromise	$1/X_q$	−0.952	−0.234	−0.197	75.430

5. Numerical Comparisons

The performance of stratification strategies discussed in the preceding sections were compared using data from the MU281 municipality population. The MU281 population consists of 281 Swedish municipalities with the largest cities discarded from the MU284 population available from Särndal *et al.* (1992, Appendix B). Three variables ($Q = 3$) were considered for stratification that include (1) population in 1985 (P85, x_1), (2) number of Social-Democratic seats (SS82, x_2) and (3) total number of seats (S82, x_3) in municipal council. Table 1 lists their population mean ($\bar{X}_q$), standard deviation (σ_q), coefficient of variation (CV_q) and skewness (γ_q). For the purpose of comparisons, we first considered three individual stratifiers (denoted by x_q), that is, three MU281 variables for stratification. In addition, we considered three multivariate stratification methods described in Sections 2, 3 and 4 based on the L-CMOS (L), the first principal component (FPC) and the K-means clustering algorithm (K). For each of the three methods, three types of compromising factors were also applied to the original data that includes equal-compromise ($c_q = 1$), scale-compromise ($c_q = 1/\sigma_q$) and size-compromise ($c_q = 1/X_q$) for $q = 1, 2, 3$. Table 2 shows the relative value of each set of compromising factors. Applying three sets of compromising factors, three sets of the transformed variables z_1, z_2 and z_3 were obtained. Table 3 displays their loadings a_{1q} in (3.1) and the percentage of the total variance of the corresponding FPC y_1 (*i.e.*, $100\lambda_1/\sum_{q=1}^{Q}\lambda_q$). Since a_{1q}'s are all negative, the FPCs were all computed by multiplying minus 1 to force them positive. To indirectly evaluate the stratification effect of each stratifier on the three MU281 variables, we first computed their linear correlation coefficients $\rho_{yq} = \mathrm{Corr}(y, x_q)$ for $q = 1, 2, 3$ in Table 4. A larger efficiency gain may be anticipated using stratifiers with larger coefficients. Table 4 shows that correlations ρ_{yq} of either L-CMOS or FPC are larger than those of the individual stratifiers x_q's except for the self-correlation of ρ_{qq}. Table 4 also shows the overall linear associations of a stratifier to all three MU281 variables evaluated by the following formulae:

$$\mu_\rho = \frac{1}{Q}\sum_{q=1}^{Q}\rho_{yq} \quad \text{and} \quad \psi_\rho = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(\rho_{yq} - 1)^2}.$$

Table 4: Correlations and associations for MU281 population variables.

Stratification Strategy		Correlations			Associations	
CMOS	Compromising factor	ρ_{y1}	ρ_{y2}	ρ_{y3}	μ_ρ	ψ_ρ
x_1	1	1.000	0.681	0.874	0.852	0.198
x_2	1	0.681	1.000	0.752	0.811	0.233
x_3	1	0.874	0.752	1.000	0.875	0.161
y_L	1	0.974	0.806	0.944	0.908	0.117
	$1/\sigma_q$	0.926	0.882	0.952	0.920	0.085
	$1/X_q$	0.977	0.814	0.927	0.906	0.116
y_1	1	0.995	0.732	0.914	0.881	0.162
	$1/\sigma_q$	0.974	0.802	0.945	0.907	0.119
	$1/X_q$	0.997	0.728	0.901	0.875	0.167

The first quantity reveals the average coefficient of ρ_{yq}'s and the second quantity indicates the standardized average deviation of the coefficients from the perfect correlation. Table 4 shows that any compromised stratifier has larger μ_ρ and smaller ψ_ρ than any individual stratifier x_q. The only exception is for a stratification by x_3 with its $\psi_\rho = 0.161$ being slightly smaller than $\psi_\rho = 0.167$ of the size-compromised FPC. A compromised stratification with the scale-compromise is better than other compromises of the same stratification method. Among others, the scaled-compromised L-CMOS is the best in terms of the linear association.

To simplify our discussion, we assumed to have $H = 4$ strata for the MU281 population and a sample of size $n = 30$. If we denote $V^*_{str}(\bar{x}^*_{str,q})$ as the variance of the stratified sample mean $\bar{x}^*_{str,q}$ of x_q under the Neyman allocation (that is, $n_h \propto N_h S_{q,h}$). Then we have $V^*_{str}(\bar{x}^*_{str,1}) = 0.941$, $V^*_{str}(\bar{x}^*_{str,2}) = 0.138$ and $V^*_{str}(\bar{x}^*_{str,3}) = 0.501$, respectively. To compare each stratification strategy (*i.e.*, a combination of a stratification method and a type of compromising factors), the relative efficiency for each of the three MU281 variables was computed via the following formula

$$e_q = \frac{V_{str}\left(\bar{x}_{str,q}\right)}{V^*_{str}\left(\bar{x}^*_{str,q}\right)}$$

and the overall efficiencies with

$$\mu_e = \frac{1}{Q}\sum_{q=1}^{Q} e_q \quad \text{and} \quad \psi_e = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(e_q - 1\right)^2}\,.$$

Figure 1 displays the values of e_q for each stratification strategy.

Figure 1 shows that efficiency gain was the greatest for each variable when the stratification was accomplished based on the corresponding variable itself. In addition, the relative efficiency is greater for variables of stronger correlations. For example, both e_1 and e_3 are simultaneously smaller for most of the stratification strategies. However, such patterns are significantly alleviated for both methods based on L-CMOS and the K-means clustering algorithm with a scale-compromise. Table 5 confirms these observations, where both overall efficiency measures μ_e and ψ_e are listed. As expected, both values are smaller for any strategy with the scale-compromise among the others using the same method. For example, the K-means clustering algorithm with the scale-compromise leads to the greatest efficiency gain with $\mu_e = 2.396$ and $\psi_e = 1.618$ under proportional allocation and with $\mu_e = 2.251$ and $\psi_e = 1.415$ under the Neyman allocation. However, the Neyman allocation does not always guarantee better efficiency for all variables as compared to proportional allocation. A multivariate study will often result in improvements in one variable; however, they often result in some inefficiency (*i.e.*, larger variance) in other variables.

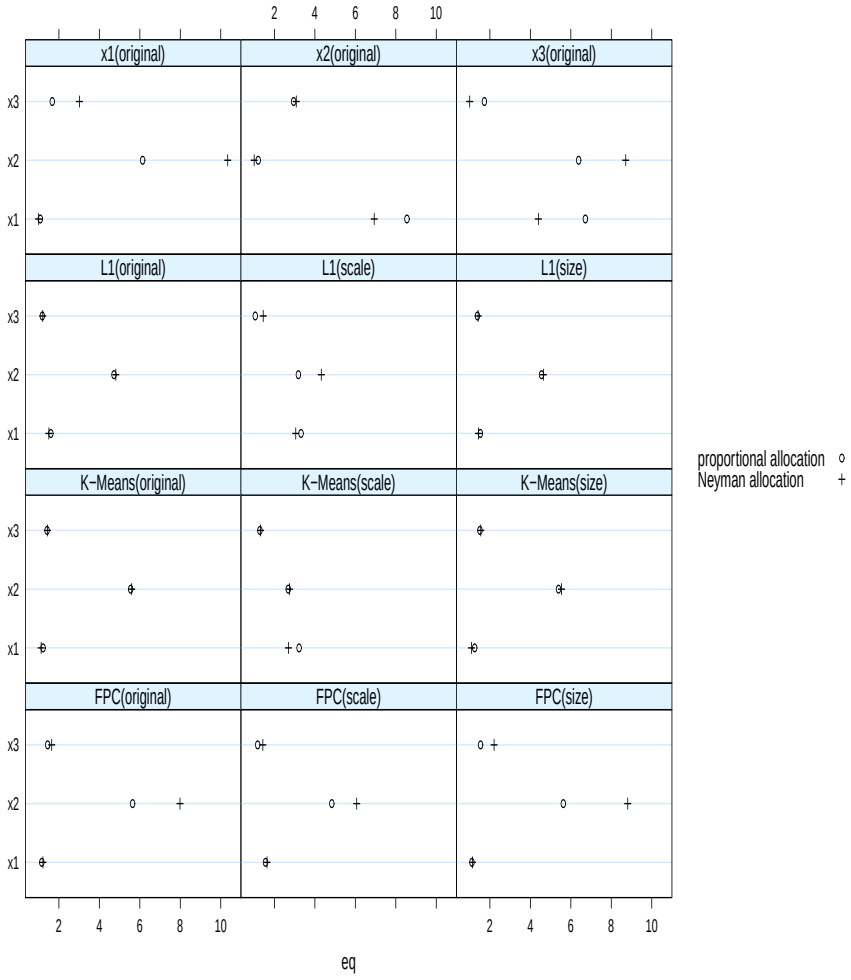

Figure 1: *Efficiency comparisons for various stratification strategies.*

6. Discussion

This paper investigated three multivariate stratification methods that are convenient and efficient to implicate a variance reduction strategy for univariate stratification. Our comparisons show that: First, some multivariate stratification strategies produce the efficiency gain for several variables as compared to any univariate stratification. Second, the overall variance reduction tends to be achieved when the K-means clustering algorithm is applied to data with compromising with respect to their scales; in addition, a similar gain was also shown when the stratification was done using the L-CMOS to the scale-compromised data. Additional efficiency gains may be possible through a multivariate allocation of the sample to the strata such as in Bethel (1989); however, this is beyond the scope of this paper. If the objective is to study some functional form of the variables, $y_M = \psi_M(\mathbf{x})$ then one would need to develop a stratifier that should strongly correlated to the form. In conclusion, the K-means clustering algorithm with the scale-compromise appears to better synthesize the information from multiple stratifiers that are readily applicable using existing software.

Table 5: Efficiency Comparisons for Stratifying MU281 Population.

Stratification		Proportional Allocation		Neyman Allocation	
Method	Compromising factor	μ_e	ψ_e	μ_e	ψ_e
x_1	1	2.965	2.993	4.786	5.518
x_2	1	4.240	4.521	3.672	3.631
x_3	1	4.944	4.556	4.703	4.866
FPC	1	2.747	2.694	3.614	4.051
	$1/\sigma_q$	2.524	2.241	3.116	3.051
	$1/X_q$	2.760	2.692	4.051	4.564
L-CMOS	1	2.501	2.179	2.507	2.224
	$1/\sigma_q$	2.519	1.840	2.925	2.235
	$1/X_q$	2.488	2.086	2.499	2.132
K-means	1	2.727	2.639	2.718	2.668
	$1/\sigma_q$	2.396	1.618	2.251	1.415
	$1/X_q$	2.716	2.557	2.726	2.641

Acknowledgements

We are grateful to the Editor and two anonymous reviewers for their substantive comments and suggestions.

References

Ardilly, P. and Tillé, Y. (2006). *Sampling Methods: Exercise and Solutions*, Springer-Verlag, New York.

Bethel, J. (1989). Sample allocation in multivariate surveys, *Survey Methodology*, **15**, 47–57.

Golder, P. A. and Yeomans, K. A. (1973). The use of cluster analysis for stratification, *Applied Statistics*, **22**, 213–219.

Hagood, M. J. and Bernert, E. H. (1945). Component indexes as a basis for stratification in sampling, *Journal of the American Statistical Association,* **40**, 330–341.

Jain, A. K., Murty, M. N. and Flynn, P. L. (1999). Data clustering: A review, *ACM Computing Surveys,* **31**, 264–323.

Jarque, C. M. (1981). A solution to the problem of optimum stratification in multivariate sampling, *Applied Statistics,* **30**, 163–169.

Kish, L. and Anderson, D. W. (1978). Multivariate and multipurpose stratification, *Journal of the American Statistical Association,* **73**, 24–34.

Kozak, M. and Verma, M. R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency, *Survey Methodology*, **32**, 157–163.

Malec, D. (1995). Selecting multiple-objective fixed-cost sample designs using an admissibility criterion, *Journal of Statistical Planning and Inference*, **48**, 229–240.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980). *Multivariate Analysis*, Academic Press, London.

Samita, S. and Kumari, W. M. R. (2006). Multivariate based stratification as an alternative to multistage stratification in stratified random sampling, *Sri Lankan Journal of Applied Statistics*, **7**, 55–69.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New-York.

Received October 29, 2012; Revised January 24, 2013; Accepted January 30, 2013