مجلّه‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار دوره‌ی ۴، شماره‌ی ۲، پاییز و زمستان ۱۳۸۶، صص ۲۱۷-۲۲۵

J. Statist. Res. Iran **4** (2007): 217-225

"Equivalent Linear Composition" as an Efficient Stratification Factor in Multipurpose Surveys

Arman Bidarbakht-nia* and Vahed Maroufy

Statistical Centre of Iran

Abstract. Horticulture survey is a multi-purpose survey which is conducted ad hoc by Statistical Centre of Iran (SCI). Availability of survey variables in the sampling frame suggests a multivariate stratification in each province based on its desired variables for acquiring a higher efficiency. There are several ways to stratify the sampling frame considering all stratification variables, such as using sum of observations on all variables, clustering, using first principal component, and specially an almost new method which uses multiple-frame techniques. We introduce, *the equivalent linear composition factor*, and illustrate how it works more efficiently than the other methods in this particular survey.

Keywords. stratified sampling; clustering; multiple-frame designs; principal components, equivalent linear composition.

1 Introduction

In sample surveys, auxiliary variables can be used to stratify units in the sampling frame in order to attain precise estimators via homogenous strata. When we have only one auxiliary variable for stratification, it is easy to stratify units through methods explained in sampling handbooks or even using cluster analysis methods. Dalenius-Hodges method (Dalenius and Hodges, 1959) is one of the most common methods on this issue which has been used since long ago.

* Corresponding author

The challenging problem arises when we have more than one auxiliary variable correlated with the study variable(s) in the survey. In such a situation the statistician has to find a subset of auxiliary variables significantly correlated with the survey variable(s) and find a strategy of stratification that optimizes the design regarding all the auxiliary variables.

Horticulture survey (HS) 2008 sampling design is one of the most complicated designs, which is conducted to obtain useful estimates for some economically important products in each province of Iran. Indeed, this is a multistage and multipurpose design in which the survey parameter values are considerably different from a province to another.

Since the sampling frame used for the HS sampling design is based on 2003 census of agriculture, all precious information related to the survey variable(s) (related products) are available (for the primary sampling units (PSUs), cities, and villages). Therefore, in order to obtain cost efficient estimates for the parameters of interest, one may stratify PSUs to gain homogenous strata with respect to any of the stratification variables.

To achieve such an objective, there are various techniques explained in section 2. Section 3 contains an explicit introduction of *equivalent linear composition*. Section 4 is devoted to a comparison among methods based on their efficiency.

2 Stratification Methods

In stratified sampling design with more than one stratification variable, statisticians usually desire to find a suitable linear composition of all stratification variables as a new stratification variable covering relative impact of all basic stratification variables which are correlated significantly with the survey variable(s). By such a method, statisticians not only manage to apply almost all fundamental variables in their study, but also have a new variable for calculating the variance needed in sample allocation process. The explanation of some common methods and their pros and cons are provided in the following subsections.

2.1 Mean and Sum as Stratification Factors

While all stratification variables are of the same unit in the HS, it is wise to select the mean or sum of the entire fundamental stratification variables as a linear combination for stratification. In HS there were two choices to make. Firstly, using the sum (or mean) of the amount of production of the main products and secondly using the sum (or mean) of the area devoting to the main products. In this paper, we apply both approaches and compare them

with some other methods. However, this method has two drawbacks, which cannot be ignored. First, this method optimizes strata for the variable(s) (or product(s)) included in the combination, that has absolutely the largest amount of production (area) in comparison to the others and second, for variables with different units it is completely worthless.

2.2 First Principal Component

Principal Component Analysis (PCA) is one of the statistical methods which considers the variance-covariance structure of a set of variables through a number of independent linear combinations of the variables (Johnson and Wichern, 2007). Although the most important objective of PCA is to reduce the number of variables to make the interpretation simpler, the first principal component can be used as a stratification variable, because it reflects a considerably large percentage of variation. The first principal component of amount of production for main products is one of the criteria applied to HS using the correlation matrix of stratification variables for each province. One major disadvantage of this method comes to pass when fundamental variables are not correlated with each other considerably, or in the especial case that there are some independent subsets of variables. In such situations first principal component can not optimize stratification for all fundamental stratification variables.

2.3 Multiple-frame Method

Skinner et al. (1994) adopting estimating methods in multiple-frame designs, introduced a new method for stratification in multi-purpose surveys. In this method, population units are stratified with respect to all stratification variables, separately. Each time an independent sub-sample is selected and the final sample is a combination of the subsets.

Suppose U is a finite population with N units. If there are A stratification variables, the population is stratified A times with respect to each of the stratification variables and H_i, $(i = 1, 2, \ldots, A)$, strata are formed in the ith stratification. Let n_i be the sample size of the ith sample (the sample chosen from ith stratification). Then the final sample size satisfies $n \leqslant \sum_{i=1}^{A} n_i$, where $n_i = \sum_{h=1}^{H_i} n_{ih}$, and n_{ih} is the sample size for stratum h in the ith stratification. The inequality occurs when some units are appeared in more than one sample. After allocating n_i units to H_i strata, the total of the variable of interest, Y, will be estimated by the following formula:

$$\hat{Y} = \sum_{i=1}^{A} \sum_{j=1}^{n_i} w_j y_j,$$

where $w_j = \left(\sum_{i=1}^{A} \pi_{ij}\right)^{-1}$, $\pi_{ij} = n_{ih_i(j)}/N_{ih_i(j)}$ is the inclusion probability of the jth unit in the ith stratification design, and $h_i(j)$ demonstrates the stratum which includes the jth sample of the ith stratification design. In HS, number of stratification variables – which is the number of main products – varies and reaches to 10 variables in some provinces. This causes n_i's to be too small to obtain precise estimates based on the sub-samples, which is considered as a drawback of this method.

2.4 Clustering

Clustering is one of the multivariate methods which helps statisticians have a better realization of the relation among variables. In this process elements are classified with respect to some similarities (Johnson and Wichern, 2007).

In clustering, elements of each cluster are similar regarding their magnitude. Since this characteristic is the objective followed by the stratification design, clustering could be used effectively as a method for stratification in multipurpose designs. One disadvantage of clustering that can not be ignored is the problem of sample size allocation to each cluster. There are some solutions to this problem, such as optimum allocation and proportionate allocation (Hensen et al., 1953), but these alternatives consider only one ancillary (stratification) variable. There are some other methods for allocation introduced by Dalenius and Hodges (1957), Yates (1960), and Cochran (1977) which are so complicated. Some of these methods are time consuming and need complicated calculations and some do not result in specific and clear answers.

2.5 Equivalent Linear Composition (ELC)

In HS we are going to find a linear combination of production amount of main products that not only demonstrates the degree of importance of all products ignoring their relation and correlation with each other (despite of PCA method), but also is not notably sensitive to the products with much more yield than the others. In fact, in the proposed method we try to remove all deficiencies introduced by the other methods.

As mentioned before, in HS each province has its own sampling design with respect to the main products of the province. However sampling designs for all provinces are multi-stage sampling with cities and villages as PSUs. In this study we are merely focusing on the first stage, i.e., stratification of PSUs, so the calculated variance for each method is only due to the between PSU variation.

If province p has A main products, then its ELC for the jth sampling unit

(PSU) will be as follows:

$$T_{ej} = \sum_{i=1}^{A} \alpha_i x_{ij}, \tag{1}$$

where x_{ij} is the amount of production of the ith product for the jth unit in the sampling frame, and coefficients α_i's are obtained such that same level of importance and significance is allocated to any main product in each province, so the main products with smaller amount of production would not be considered as relatively valueless products. This is due to the objective of the survey design which is to estimate the amount of production for all of the products, thus all of the products have equal degree of importance according to the design objectives, i.e. α_i's are obtained such that

$$\alpha_i X_i = \alpha_{i'} X_{i'} \qquad \text{for all } i, i'$$

where X_i is the amount of production for the ith product in a particular province. Moreover, in order to have the same characteristic as the first principal component, i.e. $\sum_{i=1}^{A} \alpha_i = 1$, α_i's must satisfy the following condition:

$$\alpha_i = \frac{\prod_{k \neq i} x_k}{\sum_{i<i'} x_i x_{i'}}, \qquad i, i' = 1, 2, \ldots, A \tag{2}$$

where x_i is the amount of production for the ith product in a particular province. If the constraint $\sum_{i=1}^{A} \alpha_i = 1$ is substituted by another constraint like $\sum_{j=1}^{N} T_{ej} = \sum_{i=1}^{A} x_i = x$ then we will have

$$\alpha_i = \frac{\sum_{i=1}^{A} x_i}{A x_i}. \tag{3}$$

Since both types of multipliers, α_i's (2) and (3) end up with same results in stratification and also using (3) computation becomes much easier, we selected the latter for the rest of our calculations.

3 Conclusion

The HS sampling design is optimized at province level. Therefore in each province a stratified cluster sampling is conducted in which PSUs are cities and villages for urban and rural areas, respectively. Then final sampling

Table 1. The efficiency of uni-product in compare to other methods.

var (of method)/var (uni-product)				
Method	**Peach**	**Pomegranate**	**Almond**	**Grape**
Multiple-frame	468	190	18	7
ELC	4553	1657	150	13
Area	71624	2228	1144	21
Sum of the production	23377	943	1033	5
Clustering	20189	1274	1377	1
Principal component	6778160	4	7551	251
Uni-product	1	1	1	1

units (households) are selected through one or two stages depending on the size of the PSU. We chose Markazi province with four main products (peach, pomegranate, almond, grape) to compare the stratification methods. The population of PSU's and first stage sample size in Markazi are respectively 1327 and 120 PSUs. It should be noted that we apply different methods only for stratifying the PSUs, and it does not affect the estimator and variance formulas for the sampling design. So since in stratified sampling design the total estimator is unbiased, it is reasonable to compare estimators regarding their variances through efficiency criteria. However, in this research the variance for each method is calculated from a simulation study with 1000 replications. Since the distribution of horticulture products are highly skewed, we applied fuzzy clustering to stratify PSU's in all seven stratification schemes. Bidarbakhtnia and Maroufy (2008) show that this method is considerably more efficient than Dalenius-Hodges' method in skewed populations. The estimation of production for all main products (variables) were generated 1000 times for each method in Markazi province and then the variance of each main variable was compared with the variance obtained from the method by which sampling units were stratified regarding the particular variable. We call the latter method uni-product method which, ignoring the allocation problem, is the most efficient way of stratification.

Tables 1 and 2 contain variance and efficiency of uni-product method for each of the main products in comparison to the other methods. From Table 1 we find out that the four methods, stratification according to area, sum of the production, first principal component, and clustering, do not give stable efficiencies for all stratification variables, although they give high level of efficiencies for a few number of variables.

As expected, stratification with respect to area and sum of the production optimizes the results for variables with higher amount of production and area. For example, the results for grape production in Markazi province is

Table 2. The efficiency of ELC in compare to multiple-frame method.

Product	multiple-frame variance / ELC variance
Apple	3.2
Pear	802
Glass	7.3
Apricot	6.6
Peach	250.5
Plum	63.2
Fig	66.9
Orange	5.1
Citrus	2.2
Pomegranate	40.1
Date	2.4
Almond	4.5
Walnut	19
Grape	43.4
Tabrizy	50.6

significantly efficient based on these two methods, but this is not the case for other products. First principal component method optimizes the stratification for variable(s) which do not have strong and significant correlation with other main variables.

For Markazi province, correlation matrix shows that pomegranate production amount does not have strong correlation with other variables, therefore first principal component optimizes the stratification for it and the efficiency of its estimates is much higher than that for other variables. In case of clustering, results are the same as those derived from sum and area. The only two relatively stable efficiencies are for stratifications based on ELC and multiple-frame methods which stratify the frame considering all stratification variables with (say) minimum risk; as we can see, the efficiency of peach production amount estimate is higher for these methods while for the other variables we can also have estimates with reasonable efficiencies in comparison to the other methods.

In the second step, in order to compare the two methods, multiple-frame and ELC, and to observe how well these methods work for country level estimates, efficiency of the total production estimates for country (sum of all provinces' production) was calculated via both methods. To do this, stratification was performed by both methods and the variances were computed by a simulation

with $n = 1000$ (repeat). Table 2 shows the efficiency of ELC in comparison to multiple-frame method for all estimations, which are greater than 1 for all of them.

Considering the foregoing statements, we find out that not only ELC is the best method with the highest efficiency among others, but also is easier in calculation. In the multiple-frame, other than insufficiency in sample size for each stratification round, the allocation of sample size is still the case while for ELC method we have already changed the scenario to univariate stratification which practically is very convenience to work with. So in case of Iran, ELC has been used as a stratification variable and it is suggested to consider this factor as an alternative for stratifying sampling frames in case of multivariate stratification designs (usually needed for multipurpose surveys) provided that all variable are in the same scale. However sometimes we already have a subjective index that may work better than ELC. As in case of livestock survey that there is a "livestock unit" defined in cooperation with the ministry of agriculture which is used for stratifying sampling frame in that survey.

Acknowledgment

The authors wish to thank the editor, and the referees for helpful comments that greatly improved the presentation.

References

Bidarbakhtnia, A. and Maroufy, V. (2008). A comparison between Dalenius-Hodges stratification and fuzzy clustering in skewed data. *Gozideh-Matāleb-e Āmāri* **18**, 59-71 (in Persian).

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

Dalenius, T. and Hodges, J.L. (1957). The choice of stratification points. *Skandinavisk Aktuarietidskrift* **40**, 198-203.

Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. *J. Amer. Statist. Assoc.* **54**, 88-101.

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

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis.* Prentice-Hall.

Skinner, C.J., Holmes, D.J., and Holt, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review* **62**, 333-347.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*, 3rd ed. Griffin, London.

Arman Bidarbakht-nia
Statistical Centre of Iran,
Dr. Fatemi Ave.,
Tehran 14146 63111,
Iran.
e-mail: *bidar_a@yahoo.com*

Vahed Maroufy
Statistical Centre of Iran,
Dr. Fatemi Ave.,
Tehran 14146 63111,
Iran.
e-mail: *vahed_maroofy@yahoo.com*