International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Weights Proportional to Size (WPS)-Multi Systematic Sampling Design

N. Uthayakumaran

NIE, ICMR, Chennai, Tamilnadu, India

Abstract: ***Researchers are always in pursuit of developing estimators with increased precision by incorporating the information of suitable auxiliary (size) variables either in the selection or in the estimator. For single dimensional population, in probability proportional to size sampling schemes, size information of auxiliary variable is used in the selection as well as in the estimation stage in pursuit of developing estimators of variable of interest with increased precision than even the gold standard simple random sampling. Motivated by this, an attempt has been made towards incorporating size information in the estimator in the selection as well as in the estimation stage in pursuit of developing estimator of population total of the study variable with increased precision for multi-dimensional population exhibiting linear trend.***

Keywords: Multi-systematic: Identifying labels in a linear or cyclic fashion in multiple dimensions, **Multi-dimension:** Multiple dimensions, **Linear trend:** Uniformly increasing trend

1. Introduction

Sample survey is universally recognized as a vital mode of data collection to derive a reliable statistic in a number of fields. Various sampling strategies have been developed in the last five decades or so to meet the growing demand for reliable statistics, starting with the pioneering works of Mahalonobis, Yates, Hansen and Horvitz, Sukhatme, Cochren and others. The problem of estimating finite population total has received the attention of many researchers. For a detailed review of these works, (Hanif and Brewer, 1980), (Sunter, 1986) etc., can be referred. The estimating strategies by them can be broadly classified into two categories namely those designs which use auxiliary information either in the selection stage or in the estimation stage and those not using the auxiliary information at all. Some of the methods which use auxiliary information in single dimension population to estimate population parameter are Midzuno sampling (Midzuno, 1952), Rejective sampling (Sampford, 1967), PPS systematic sampling scheme (Sunter, 1986), Markov sampling with auxiliary information (Suresh Chandra and Sampath, 1992), Markov systematic sampling using auxiliary information (Sampath and Uthayakumaran, 1996), Geographical information based sampling (Uthayakumaran and Selvaraj, 2015) etc. The method which uses auxiliary information for two dimensional population to estimate population parameter is WPS dual circular systematic sampling (Uthayakumaran and Venkatasubramanian, 2013). The method which uses auxiliary information for three dimensional population to estimate population parameter is WPS cubical circular systematic sampling (Uthayakumaran and Venkatasubramanian, 2015). In these WPS dual and cubical circular systematic sampling methods, only equal dimensional populations are dealt for estimation with selection of sampling units in a cyclic fashion. Modifying and generalizing this technique so as to be flexible for both equal and unequal multiple dimensions with the selection of sampling units in a linear or cyclic fashion, a design called WPS-Multi systematic sampling has been developed. Using this sampling design, an attempt has been made towards incorporating size information in the estimator in the selection as well as in the estimation stage in pursuit of developing estimator of population total of the study variable with increased precision for multi-dimensional population exhibiting linear trend.

2. WPS-Multi systematic sampling design

A multiple dimensional population element may be represented by study variable Y_{i_g} , $i_g = 1, 2,\ldots, N_g$; $g = 1, 2,\ldots, h$. Here, Y_{i_g} is the value of the h dimensional elements. Let X_{i_g} , $i_g = 1, 2,\ldots, N_g$; $g = 1, 2,\ldots, h$ be the size variable accordingly. Each dimensional size variable total D_{i_g} , $i_g = 1, 2,\ldots, N_g$; $g = 1, 2,\ldots, h$. are defined on the respective dimensional units for X_{i_g} , $i_g = 1, 2,\ldots, N_g$; $g = 1, 2,\ldots, h$.. T_N denotes the total size of the population. A WPS-Multi systematic sample can be drawn as follows:

In sampling units ($n_1 \times n_2 \times n_3 \times \ldots \times n_h$) with this procedure, the cumulative dimensional size variable totals T_{i_g} are determined on the respective dimensional size variable total D_{i_g} , $i_g = 1, 2,\ldots, N_g$; $g = 1, 2,\ldots, h$. The population contains ($N_1 \times N_2 \times N_3 \times \ldots \times N_h$) units. The sampling interval k_g is T_N/n_g.

A WPS-Multi systematic sample is selected in a linear fashion by drawing multiple independent starting coordinates r_g at random, each between 1 and k_g. A sample of size ($n_1 \times n_2 \times n_3 \times \ldots \times n_h$) contains all units whose coordinates are of the form

$$\{r_g + \gamma k_g\}, \gamma = 0, 1,\ldots, (n_g-1); g = 1, 2,\ldots, h \quad (2.1)$$

If $T_N \neq n_g k_g$, $g = 1, 2,\ldots, h$., a WPS-Multi systematic sample is selected in a cyclic fashion by drawing multiple independent starting coordinates r_g at random, each between 1 and T_N. The sampling interval k_g is the integer part of the

ratio T_N/n_g. The coordinates of the sampling units are of the form

$$\begin{aligned} &\{r_g + \gamma k_g\} \quad \text{if } 1 \le r_g + \gamma k_g \le T_N \\ &\{r_g + \gamma k_g - T_N\} \text{ if } r_g + \gamma k_g > T_N, \ \gamma = 0, 1, .., (n_g - 1) \\ &\qquad\qquad g = 1, 2, .., h \end{aligned} \tag{2.2}$$

For the values obtained from the above form of coordinates, sample units are identified with the labels selected using the corresponding cumulative dimensional size total of T_{i_g}, $i_g = 1, 2, \ldots, N_g$; $g = 1, 2, \ldots, h$. Let D_{i_g}, $i_g = 1, 2, \ldots, N_g$; $g = 1, 2, \ldots, h$ be integers corresponding to dimensional size variable totals of the population units respectively when they are arranged in a linear order of the population.

Estimation of population total of study variable

Survey analysts normally need to prevail over the problem of estimating population total for the study variable with size information. For the sampling design described above, an estimator of the population total of the study variable (Y) is given by

$$\left(\hat{y}_{WPS-Multi} \right)_{r_g} = \frac{1}{\prod_{g=1}^{h} n_g} \left\{ \prod_{g=1}^{h} \sum_{i_g=1}^{n_g} \left(\frac{Y_{i_g}}{W_{i_g}} \right)_{r_g} \right\} \tag{2.3}$$

where $W_{i_g} = \prod_{g=1}^{h} D_{i_g} / (T_N)^h$ (2.4)

The approximate expression for the variance of estimator is given by following (Hartley and Rao, 1962)

$$V\left(\hat{y}_{wWPS-multi}\right)_{r_g} = \frac{1}{\prod_{g=1}^{h} n_g} \prod_{g=1}^{h} \sum_{i_g=1}^{Ng} \left\{ \frac{Y_{i_g}}{W_{i_g}} - Y \right\}^2 W_{i_g} \left(1 - (\prod_{g=1}^{h} n_g - 1) W_{i_g} \right) \tag{2.5}$$

where population total of study variable

$$Y = \prod_{g=1}^{h} \sum_{i_g=1}^{Ng} Y_{i_g} \tag{2.6}$$

Demonstration of WPS-Multi systematic sampling design with two-dimensional population (h=2) using hypothetical data exhibiting linear trends:

$Y_{i_g} = i_1 + i_2$, where $i_1 = 1, 2, \ldots, 4$; $i_2 = 1, 2, \ldots, 6$

$X_{i_g} = \beta (i_1 + i_2)$, where $\beta = 100$ (2.7)

The random starts for two dimensions and fixing the cells with the use of size information establish the sample. Let populations at taluks and villages are considered as known details. If one deals with the population at taluks and appropriate and easily available variable such as population at villages which are linearly arranged, it is possible to arrive at very closer estimate of the population total of study variable. In this demonstration, row dimensions are 4 groups of taluks and column dimensions are 6 groups of villages according to its range of population. Labels of group of taluks (I_1) are given from 1 to 4; labels of group of villages (I_2) are given from 1 to 6 according to its ascending order of row and column dimensional population sizes D_{i_1} and D_{i_2} respectively. The study variable and size variable can be generated under the linear trends (2.7).

Let Y_{i_g} and X_{i_g} can be the respective prevalence cases and population. Let D_{i_1}, and D_{i_2} be the total populations of 4 groups of taluks, 6 groups of total populations of grouped villages respectively. Let T_{i_1} and T_{i_2} be the cumulative totals of population groups of taluks D_{i_1} and population groups of villages D_{i_2} respectively. Let Y be the total prevalence cases. Let T_N be the total population. Towards this, the technique of WPS-Multi systematic sampling can be demonstrated by applying it to the case of sampling **6** units from the two-dimensional population of 24 units for the hypothetical data generated using the models (2.7).

Table-1: Distribution of study and size variables for the hypothetical data according to taluks and villages

I_2 / I_1	1 <200	2 200-300	3 301-400	4 401-500	5 501-600	6 >600	D_{i_1}	T_{i_1}
1	**2** (200)	**3** (300)	**4** (400)	**5** (500)	**6** (600)	**7** (700)	2700	2700
2	**3** (300)	**4** (400)	**5** (500)	**6** (600)	**7** (700)	8 (800)	3300	6000
3	**4** (400)	**5** (500)	**6** (600)	**7** (700)	**8** (800)	**9** (900)	3900	9900
4	**5** (500)	**6** (600)	**7** (700)	**8** (800)	**9** (900)	10 (1000)	4500	14400
D_{i_2}	1400	1800	2200	2600	3000	3400		
T_{i_2}	1400	3200	5400	8000	11000	14400		T_N=14400

- **Numbers in bold are prevalence cases Y_{i_g} - study variable**
- **Numbers in parenthesis are cell population X_{i_g} - size variable**

Step 1:
A random number is drawn from 1 to 7200 of I_1 (say 5000). Sampling interval $k_1 = T_N/n_1 = 14400/2 = 7200$, using (2.1), the following coordinates are identified.

{5000, 12200}

For the values obtained above, corresponding i_1th labels of sampling units are selected using the cumulative total of row dimension.

{2, 4}

Step 2:
A random number is drawn from 1 to 4800 of I_2 (say 2000). Sampling interval $k_2 = T_N/n_2 = 14400/3 = 4800$, using (2.1), the following coordinates are identified.

{2000, 6800, 11600}

For the values obtained above, corresponding i_2^{th} labels of sampling units are selected using the cumulative total of column dimension.

{2, 4, 6}

Step 3:
The following sample units of two-dimensional population are selected using the labels selected from the step1and step 2.

(2, 2) (2, 4) (2, 6) (4, 2) (4, 4) (4, 6)

In the above demonstration, if it is assumed that the sample survey is carried out in the randomly selected 2 group of taluks and 3 group of villages (2 x 3 = 6 cells) to find out the estimate of study variable, it can be noted that value of the study variable deducted from the selected cells are 4, 6, 8, 6, 8 and 10.

Using (2.3), the estimated total for the study variable is 143.9 while population total of the study variable is 144.

Similarly, this exercise may be executed for more than two (h > 2) dimensional populations by considering associate factors like, districts, taluks, villages etc.

Interval estimation
In practice, interval estimation is very much helpful to revise and understand the estimate to its valid conclusions. It is also an effective tool to derive simple and intuitive way to interpret the results. In combination with the statistical power, the interval estimates help to interpret the findings in a more sensible way. So, the 95% confidence interval **(CI)** for the estimate can be obtained by using the formula for the variance of estimation. The 95% CI for estimate in the above situations is shown in Table-2.

Table 2: The 95% CI for the population total of the study variable in hypothetical data

Source	Population	Sample	SE	95% CI
Hypothetical data	144	143.9	1.3	141.4 - 146.4

3. Discussion

In the WPS-Multi systematic sampling design, approach of using size variable in multi-dimensional population is advocated in this paper. This is an attempt to reduce the variance and enhance the quality of estimate of the population total of the study variable by using size information in the selection stage. The weights based on size information (2.4) used in estimator (2.3) are providing strength to estimate the population total of study variable.

The requirement and specific arrangement of the multi-dimensional population in the WPS-Multi systematic sampling design discussed in this paper, in effect ensure - the observance of the study variable on a much enhanced setup. The use of size variable in arranging the total population for the selection of the sample indirectly satisfy the linear trend assumption for the study variable. The random starts for multiple dimensions and fixing the cells with the use of size information uniquely determine the sample. Also, this approach provides a good representative

sample, as care is being taken to spread the population units in the sample.

The hypothetical data generated through models (2.7) towards explaining the WPS-Multi systematic sampling design demonstrate its practical utility for estimation.

4. Conclusion

It is significant to note that complete sampling frame is not required for this sampling design. The linear arrangement of the population in multi-dimensions by size variable ensures the linear trend assumption for the study variable. By this arrangement the proposed sampling design ensures closeness of the estimate of the study variable to the population. For population arranged as (N_1 x N_2 x …..,x N_h) cells, the suggested design is useful in selecting the sample. To estimate the population total of study variable, incorporating size information in the estimator in the selection as well as in the estimation stage, it is possible to arrive at reliable estimator of variable of interest with increased precision.

In real life situation, more number of dimensions will give huge variation and ambiguity. Generally, sampling design with three dimensional arrangements is sufficient for large scale surveys. The researchers can decide the number of dimensions to accomplish a reliable estimator using proper assessment, existing information and arrangement of associate factors.

This sampling design will be an alternate and beneficial to all the large scale, multi-stage surveys. The use of WPS-Multi systematic sampling design in real life situations is anticipated with its practical utility for estimation. More research with the multi-stage surveys using this proposed sampling design will enhance and provide reliable estimates in the estimation of parameters of variable of interest in the field of sampling.

Reference

[1] W.G Cochran, 1977: *Sampling Techniques*, Third Edition, Wiley Eastern, P227-
[2] Hansen, M.H and W.N. Hurwitz (1943): On the theory of sampling from finite populations, Ann. Math. Statist, Vol 14, P3333-362.
[3] Hanif M, and Brewer K.R.W. (1980): Sampling with unequal probabilities without replacement: A review, Int. Statist. Rev., 48, 317-335.
[4] Hartley H.O. and Rao J.N.K (1962): Sampling with unequal probability without replacement, Ann. Math. Stat., 33, 350-374.
[5] Mahalanobis, P.C, 1940: Report on the sample census of jute in Bengal, Ind. Central Jute Committee.
[6] Midzuno, H., 1952: *On the sampling system with probabilities proportionate to sum of sizes*, Annals of Institute of Statistical Mathematics, 2, 99-108.
[7] Sampath S. and Uthayakumaran N. (1996): Markov Systematic sampling using auxiliary information, Statistica, anno LVI, n.4, 439-443.
[8] Samford M.R. (1967): On sampling without replacement with unequal probabilities of selection, Biometrika, 54, 499-513.
[9] Sukhatme, P.V, 1950: Efficiency of sub sampling designs in yield surveys, J. Ind. Soc. Agr. Statist., Vol. 2, P212-228.
[10] Sunter, A., 1986: *Solutions to the problem of unequal probability sampling without replacement*, In. Stat. Rev., 54, 33-50.
[11] Suresh Chandra K. and Sampath S. (1992): Markov sampling with auxiliary information, Statistical Papers, 33, 83-91.
[12] N. Uthayakumaran, and S. Venkatasubramanian, 2013: *Dual circular systematic sampling methods for disease burden estimation*, International journal of statistics and analysis, Accepted for publication.
[13] N. Uthayakumaran, and S. Venkatasubramanian, (2015): An alternate approach to multistage sampling: UV Cubical *circular systematic sampling method*, International journal of statistics and applications, 5(5), 169-180.
[14] N. Uthayakumaran, and V. Selvaraj (2015): *Geographical information based sampling schemes*, International journal of statistics, Vol. 39, Issue 2, 1131-1138.
[15] Yates, F., 1948: *Systematic sampling*, Transactions Royal Society, London, A 241, 345-377.