International Journal of Computational and Theoretical Statistics
ISSN (2384-4795)
Int. J. Comp. Theo. Stat. 2, No. 1 (May-2015)

Improved Variance Function in Cluster Sampling

Shukla A.K. [1], **Yadav S.K.**[2*] **and Tiwari V.** [1]

[1] *Department of Statistics, D.A-V. College, Kanpur, U.P., INDIA*
[2] *Department of Mathematics and Statistics (A Centre of Excellence),Dr. RML Avadh University, Faizabad-224001, U.P., INDIA*

Received Nov. 29, 2014, Revised Feb. 3, 2015, Accepted Feb. 20, 2015 Published May 1, 2015

Abstract: In the present article, we suggest nonlinear regression models for variance function in cluster sampling. The estimates of parameters of suggested and existing models have been obtained. The suggested and existing models have been fitted to data sets given in Sukhatme et al (1984) and Zakkula Govindarajulu (1999). An improvement has been shown over existing models in the sense of best fitting, that is having lesser residual mean Square S^2 for the relationship between variance within cluster S_w^2 and size of the cluster ***M***.

Keywords*:* Non-linear regression, variance function, Efficiency.

1. INTRODUCTION

In cluster sampling, clusters are the sampling units and are the group of basic units. The clusters may be of equal sizes or may be of unequal sizes. In our study we have considered the case of equal size. So let the population consists of N clusters each of size M, thus having NM basic units.

The variance S^2 among all basic units in the population is given by,

$$S^2 = [M(N-1)S_b^2 + N(M-1)S_w^2]/(NM-1) \quad (1)$$

If N is large, then the expression (1) can be rewritten as

$$S^2 = S_b^2 + (M-1)S_w^2/M \quad (2)$$

So that

$$S_b^2 = S^2 - (M-1)S_w^2/M \quad (3)$$

Where

$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - y_{..})^2/(NM-1)$, (Variance of the whole population)

$S_w^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - y_{i.})^2/N(M-1)$, (Variance within cluster in the population)

$S_b^2 = \sum_{i=1}^{N}(\bar{y}_{i.} - \bar{y}_{..})^2/(N-1)$, (Variance between cluster means in the population)

d $\bar{y}_{..} = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}/(NM)$, $\bar{y}_{i.} = \sum_{j=1}^{M} y_{ij}/M$, y_{ij} = value of the j^{th} unit in the i^{th} cluster.

The problem is how to find the optimum value of M, for predicting the variance S_b^2 between units in the population, as a function of M. We can find S_b^2 if S^2 and S_w^2 are known. Since S^2 is the variance among all elements, it is not affected by the size of the cluster. However, S_w^2 will be affected by the size of the cluster if the size changes.

E-mail: alokshukladav@gmail.com, drskystats@gmail.com, vishwas.stats@gmail.com

In cluster sampling, it has been the area of research to establish the appropriate relationship between the size of the cluster and the variance within the cluster. Many authors, including Smith (1938), Jessen (1942), Hansen and Hurwitz (1942), Mahalanobis (1940, 1942) etc., have studied the problem of the determination of the optimum cluster size from both point of view of variance as well as cost function. They have given almost the same functional form describing the relationship between size of the cluster and the variation within cluster. For a given sample size, the sampling variance increases with the cluster size and it decreases with the number of clusters. On the other hand, the cost decreases with cluster size and it increases with the number of clusters. Therefore, it is necessary to determine a balancing point by finding the optimum cluster size and the number of clusters in the sample so that variance is the minimum for a fixed cost or vice-versa.

A function describing the relationship between variance among clusters and size of the clusters is very useful in cluster sampling studies. This variance function may help to determine estimates of variance between cluster means for any cluster size, rather than only those selected for sample survey. It will also help in the determination of the optimum cluster size. Several related functions, about the cluster variance and size of the cluster, have been studied in literature. These functions are different in their nature. Smith (1938) suggested the cluster size is related with the variance between cluster means as,

$$S_b^2 = S^2 M^{-g}, \qquad (4)$$

where g is a constant.

Jessen (1942) gave the relationship between within cluster variance and the size of the cluster by a non-linear form as follows:

$$S_w^2 = \alpha M^{\beta}, \qquad (5)$$

where α and β are parameters to be determined by a suitable method of estimation. Note that Mahalanobis (1940) also suggested a similar relation and Sukhatme et al. (1984) described the detailed procedure of obtaining the optimum cluster size under different conditions using (5).

Hansen *et al*. (1953) established a relation, in which intra-class correlation was related to cluster size as

$$\lambda = \alpha M^{\beta}, \qquad (6)$$

where λ is the intra-class correlation of the clusters.

Misra et al. (2010) suggested the use of asymptotic regression model having following deterministic component

$$S_w^2 = \alpha + \beta\rho^M \quad 0 < \rho < 1 \ , \ M > 1 \qquad (7)$$

to describe the relationship between within cluster variance and size of the cluster and showed that their model described the phenomenon in a better manner as compared to that of Jessen (1942) model.

Tiwari and Misra (2011) suggested the following model for describing relationship between within cluster variance and size of the cluster as

$$S_w^2 = \alpha + \beta M + \rho/M \ , \ M > 1 \qquad (8)$$

where *α*, *β*, and *ρ* are the parameters of the model. Tiwari and Misra (2011) show that the model in (8) describes the phenomenon in a better manner as compared to that of Jessen (1942) and Misra *et al*. (2010) models. Cochran (1977) criticized on the functions (4), (5), and (6) that S_w^2 increases without bound as *M* increases and also suggested that a finding some more appropriate function form of S_w^2, which approaches an upper bound with large *M*, and good fit (estimated values of S_w^2) over the range of *M*.

2. SUGGESTED MODELS

After examining the related models in literature mentioned in Section 1, we have suggested the following three models for explaining the relationship between within cluster variance and the size of the cluster as follows:

$$S_w^2 = \alpha + \delta M + \beta\rho^M, 0 < \rho < 1 \text{and } M > 1 \quad (9)$$
$$S_w^2 = \alpha + \delta / M + \beta\rho^M, 0 < \rho < 1 \text{and } M > 1 \quad (10)$$
$$S_w^2 = \alpha + \delta / M^2 + \beta\rho^M, 0 < \rho < 1 \text{and } M > 1 \quad (11)$$

where α, δ, β, and ρ are the parameters of the suggested models to be determined from survey data.

The models (9)-(11) are known as asymptotic regression models and these models predict the behaviour of S_w^2 with change in the value of *M*. These models are extensively used in agricultural, fisheries, psychological researches etc. The parameter α defines the asymptotic value of the models (10) and (11). The computation of its parameters can be made by non-linear least-squares estimation for which several statistical software programs are easily available. The suggested models (10) and (11) do not increase without bound but approach their asymptotic value of α, therefore, they do not suffer from drawbacks, as mentioned by Cochran (1977). Contrary to functions (4)-(8), the suggested relation in (10) fits (estimated values of S_w^2) extremely well to data set having large cluster sizes.

3. FITTING OF MODELS

The above models are classified by Draper and Smith (1998) in two groups as intrinsically linear and intrinsically non-linear models, model (5) is intrinsically linear can be transformed into a form in which parameters appear linearly. The direct application of least square method is possible to estimate parameters of models (5) and (7). The Model (8) and models (9) to (11) are intrinsically non-linear models, then parameters of model (8) and models (9) to (11) are estimate by iterative procedure as Levenberg-Marquardt's method.

Residual mean Square- s^2

Residual mean Square- s^2 for our models will be $s^2 = \dfrac{\sum_{i=1}^{n}\left(S_w^2 - \hat{S}_w^2\right)^2}{n-p} = \dfrac{\sum_{i=1}^{n} e_i^2}{n-p}$

Where n is number of observation and p is number of parameters in the model.

A smaller value of s^2 indicate that regression due to error is small which imply that sum of square due to regression is high enough to equal to total sum of squares. A small value of s^2 reflects the appropriateness of the fitted model.

4. DETERMINATION OF VARIANCE FUNCTION

If the whole population is considered as a single cluster, it will contain NM elements (i.e. treating the population as a single cluster with NM elements), then S_w^2 will be the total variance (S^2) for (10), thus

$$S^2 = \alpha + \delta / NM + \beta\rho^{NM} \quad (12)$$

Between clusters variance for the suggested model is obtained by putting (10) and (12) in (3) as

$$\hat{S}_b^2 = [\hat{\alpha} + \frac{\hat{\delta}}{NM} + \hat{\beta}\hat{\rho}^{NM} - \frac{(M-1)}{M}\{\hat{\alpha} + \frac{\hat{\delta}}{M} + \hat{\beta}\hat{\rho}^M\}] \quad (13)$$

where $\hat{\alpha}$, $\hat{\delta}$, $\hat{\beta}$, and $\hat{\rho}$ are the estimated parameters, M is the cluster size and N is the number of clusters.

The variance of sample mean in cluster sampling is,

$$V(\bar{y}_c) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 \quad (14)$$

5. EMPIRICAL STUDY

As an illustration, we used two data sets given by Sukhatme *et al*. (1984) and Zakkula Govindarajulu (1999). Here the values of S_w^2 have been calculated for clusters of different size in (acre)^2, where the study variable is the area under wheat crop. The estimated values of S_w^2 for different values of M, using suggested models (9) – (11) and the models of Jessen (1942), Misra *et al*. (2010), and Tiwari and Misra (2011) expressed by (5), (7), and (8), respectively, are given in Table-1 and Table-2. Note that the estimated values of S_w^2 for different values of *M* using the suggested models (9) – (11) have been obtained by the SPSS 17.0 software.

Table 1(a). Estimated values of S_w^2 by different models

M	Observed value of S_w^2	Fitted value of S_w^2 for model (1.5)	Fitted value of S_w^2 for model (1.7)	Fitted value of S_w^2 for model (1.8)	Fitted value of S_w^2 for model (2.1)	Fitted value of S_w^2 for model (2.2)	Fitted value of S_w^2 for model (2.3)
2	78.10	81.53	79.84	77.39	78.364	78.089	78.060
4	84.28	84.25	82.71	85.64	83.643	84.320	84.507
8	88.92	87.05	87.60	89.80	89.500	88.875	88.613
16	93.50	89.95	94.75	91.96	93.291	93.516	93.624
$NM =$ 1176	108.33	110.22	108.17	108.34	108.331	108.330	108.325

Table 1(b). Estimated values of parameters in different models

Model	**Parameters**				
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\beta}$	$\hat{\rho}$	s^2
Model (1.5)	78.886	-	0.0473	-	10.479
Model (1.7)	108.171	-	31.530	0.948	4.410
Model (1.8)	93.813	-	0.012	-32.888	2.756
Model (2.1)	93.999	0.012	-23.548	0.815	0.856
Model (2.2)	108.347	-20.635	-21.074	0.973	0.004
Model (2.3)	108.325	-25.812	-25.535	0.961	0.163

Table 1(c). Estimated values of S_b^2 from equation (3) by using different models

M	Observed value of S_b^2 from equation (1.3)	Estimated value of $\hat{S}_b^2$ for model (1.5)	Estimated value of $\hat{S}_b^2$ for model (1.7)	Estimated value of $\hat{S}_b^2$ for model (1.8)	Estimated value of $\hat{S}_b^2$ for model (2.1)	Estimated value of $\hat{S}_b^2$ for model (2.2)	Estimated value of $\hat{S}_b^2$ for model (2.3)
2	69.28	69.45	68.25	69.64	69.14	69.28	69.29
4	45.12	47.03	46.13	44.11	45.59	45.09	44.94
8	30.52	34.05	31.53	29.76	30.18	30.56	30.78
16	20.69	25.89	19.34	22.12	20.87	20.65	20.55

Table 2(a). Estimated values of S_w^2 by different models

M	Observed value of S_w^2	Fitted value of S_w^2 for model (1.5)	Fitted value of S_w^2 for model (1.7)	Fitted value of S_w^2 for model (1.8)	Fitted value of S_w^2 for model (2.1)	Fitted value of S_w^2 for model (2.2)	Fitted value of S_w^2 for model (2.3)
15	0.05	0.1176	0.0361	0.0289	0.0392	0.0514	0.0422
20	0.08	0.1239	0.0898	0.1008	0.0845	0.0731	0.0834
25	0.11	0.1290	0.1349	0.1440	0.1288	0.1212	0.1303
30	0.18	0.1334	0.1728	0.1727	0.1722	0.1731	0.1729
35	0.22	0.1372	0.2046	0.1933	0.2148	0.2211	0.2094
$NM =$ 8820	0.37	0.3747	0.3717	0.3727	0.3711	0.3701	0.3717

Table 2(b). Estimated values of parameters in different models

Model	**Parameters**				s^2
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\beta}$	$\hat{\rho}$	
Model (1.5)	0.072	-	0.182	-	0.0039
Model (1.7)	0.372	-	-0.566	0.966	0.0004
Model (1.8)	0.3163	-	0.0000061	-4.311	0.0009
Model (2.1)	2.354	0.00001	-2.456	0.996	0.0003
Model (2.2)	0.369	8.488	-1.630	0.960	0.0001
Model (2.3)	0.372	20.110	-0.793	0.958	0.0002

Table 2(c). Estimated values of S_b^2 from equation (3) by using different models

M	Observed value of S_b^2 from equation (3)	Estimated value of $\hat{S}_b^2$ for model (5)	Estimated value of $\hat{S}_b^2$ for model (7)	Estimated value of $\hat{S}_b^2$ for model (8)	Estimated value of $\hat{S}_b^2$ for model (9)	Estimated value of $\hat{S}_b^2$ for model (10)	Estimated value of $\hat{S}_b^2$ for model (11)
15	0.320	0.265	0.338	0.346	0.335	0.322	0.332
20	0.294	0.257	0.286	0.277	0.291	0.301	0.292
25	0.264	0.251	0.242	0.234	0.247	0.254	0.247
30	0.196	0.246	0.205	0.206	0.205	0.203	0.205
35	0.156	0.241	0.173	0.185	0.162	0.155	0.168

6. CONCLUSION

The models in literature generally do not fit well since some of these models increase without bound and are not suitable for large cluster sizes, whereas, suggested models fit very well to data and from Table1(b) and Table2(b), it is easily seen that the suggested model (10) is the best one in the sense of having the least value of residual mean square(s^2).Table1(c) and Table2(c) show that the suggested model (10) is the best one in the sense of having the estimated value of S_b^2 very close to observed value. The asymptotic value of the suggested model is the same as that of single cluster variance (i.e. population treating as single cluster). Single cluster variance never goes up to variance for the simple random sampling and this is also proven by the asymptotic value of the suggested model. Thus the suggested model (10) removes all drawbacks as mentioned by Cochran (1977). Thus the suggested model should be preferred for the optimum cluster size in cluster sampling.

ACKNOWLEDGMENT

The authors are very much thankful to the editor in chief and the unknown referees for critically examining the manuscript and suggestions to improve the earlier draft.

REFERENCES

[1] Cochran,W.G., Sampling Techniques. 3rd Ed., John Wiley & Sons, 1977.

[2] Draper, N.R. and Smith, H., Applied regression analysis. 3rd Ed., John Wiley & Sons, 1998.

[3] Hansen, M.H. and Hurwitz, W.N. "Relative efficiencies of various sampling units in population enquiries", JASA, 37, pp. 89-94, 1942.

[4] Hendricks, W.A. "The relative efficiencies of groups of farms as sampling units", JASA, 39, pp. 366-376, 1944.

[5] Jessen, R.J. "Statistical investigation of sample survey for obtaining farm facts. Iowa Agricultural Experiment Station, Research Bulletin", 304, 1942.

[6] Mahalanobis, P.C., "A sample survey of acreage under jute in Bengal. Sankhya", 4, pp. 511-530, 1940.

[7] Mahalanobis, P.C., "General report on the sample census of area under jute in Bengal. Indian Central Jute Committee", 1942.

[8] McVay, F.E., "Sampling methods applied to estimating numbers of commercial orchards in commercial peach area", JASA, 42, pp. 533-540, 1947.

[9] Misra, G.C., Yadav, S.K., Shukla, A.K., and Raj B., "Use of a non-linear model for improved estimation in cluster sampling", Journal of Reliability and Statistical Studies, 3 (2), pp. 73-78, 2010.

[10] Smith, H.F., "An empirical law describing heterogeneity in the yields of agricultural crops", Journal of Agricultural Science, 28, pp. 1-23, 1938.

[11] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C., "Sampling theory of surveys with applications", Indian Society of Agricultural Statistics, 1984.

[12] Tiwari R.B. and Misra G.C., " Estimation of optimum cluster size", IFRSA's International Journal of Computing, l (4), pp. 717-722, 2011.

[13] Zakkula Govindarajulu, " Elements of Sampling Theory and Methods", Printice Hall Publication, 1999.

International Journal of Computational and Theoretical Statistics
ISSN (2384-4795)
Int. J. Comp. Theo. Stat. 2, No. 1 (May-2015)

E-mail: alokshukladav@gmail.com, drskystats@gmail.com, vishwas.stats@gmail.com