Journal of Reliability and Statistical Studies [ISSN: 0974-8024 (Print), 2229-5666 (Online)]
Vol. 3, Issue 2 (2010): 73-78

USE OF A NON-LINEAR MODEL FOR IMPROVED ESTIMATION IN CLUSTER SAMPLING

Govind Charan Misra[*], Subhash Kumar Yadav[**], Alok Kumar Shukla[*] and Raj Bahadur[1]

[*]Dept of Statistics, D.A-V College, Kanpur, India
E Mail: drmisragovind@gmail.com
[**]Dept of Mathematics & Statistics, Dr R M L Avadh University, Faizabad
E Mail: drskystats@gmail.com

Abstract

Several researchers have attempted to develop a general law to predict a general relationship between variance within cluster S_w^2 and size of the cluster M for purposes like determination of optimum cluster size etc. In the present study a non-linear model has been suggested for describing the relationship between S_w^2 and M which has shown improvement over existing models and results have also been verified with the help of an example.

Key Words: optimum cluster size, Non-linear regression, variance function, Efficiency.

1. Introduction

It has been the area of interest in cluster sampling to find the appropriate functional relationship between the size of the cluster and the variance within the cluster. Many authors including Smith (1938), Jessen (1942), Hansen & Hurwitz (1942), Mahalanobis (1940, 1942) etc. have studied the problem of determination of optimum cluster size from both point of view of variance as well as cost function. They have given almost same functional form describing relationship between size of the cluster and 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 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 minimum for a fixed cost or vice-versa.

Let $\bar{y}$ be the sample mean of the variable under study for a sample of size n. Then we know that the variance of $\bar{y}$ in cluster sampling is

$$V(\bar{y}) = \frac{(1-f)}{n} S_b^2 \tag{1}$$

Where $f = \frac{n}{N}$ is finite population correction, n is size of the sample and S_b^2 is variance between cluster means defined as:

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\bar{Y}_i - \bar{\bar{Y}} \right)^2$$

Where $\bar{Y}_i = \frac{Y_i}{M}$ is the mean per element for the ith cluster and $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \frac{Y}{NM}$ is population mean per element along with $Y = \sum_{i=1}^{N} Y_i$ as population total, N total number of clusters and M size of the cluster.

It is of interest to know how variance $V(\bar{y})$ behaves with the cluster size M. This involves knowing relationship between S_b^2 and M. By analysis of variance S_b^2 can be found if we know

(i) The variance S^2 between all elements in the population.

(ii) The variance S_w^2 within all elements of the same cluster.

Where S^2 and S_w^2 are respectively defined as:

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{\bar{Y}}\right)^2 \text{ and } S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2$$

Thus we have to estimate S^2 and S_w^2 to find S_b^2 by analysis of variance. The data obtained from the sample gives the estimates of S^2 and S_w^2 for actual size of the units or clusters used in sample. Since S^2 is 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. It might be expected to increase as the size of the large cluster increases. It is, therefore, of interest to know how S_w^2 and M are functionally related to each other. Through an investigation McVay (1947) suggested that if the large clusters under investigation differ slightly in size from the clusters actually used, a first approximation is to regard S_w^2 as constant and has observed that this approximation may often be satisfactory. For a better approximation Jessen (1942), Mahalanobis (1940,42) and Hendricks (1944) have attempted to develop a general law to predict how S_w^2 changes with size of the cluster M. On the basis of several agricultural surveys, they observed that S_w^2 appears to bear a relation with M having empirical form as:

$$S_w^2 = aM^b, \quad b > 0 \tag{2}$$

where a and b are constants to be determined from the survey data.

The empirical relation (2) suffers from several drawbacks. Hendricks (1944) pointed out that the relation (2) does not fit well to data having large cluster sizes. Cochran (1999) has remarked that in the relation (2), S_w^2 increases without bound as M increases, and he has strongly recommended to search a relation between S_w^2 and M which approaches an upper bound.

2. Suggested Non-Linear Model

Keeping in view the drawbacks of the model (2) and suggestions of Cochran (1999), we have suggested the following non-linear model to describe functional relationship between S_w^2 and M as

$$S_w^2 = a + b\rho^M \tag{3}$$

where a, b and ρ are constants to be determined from survey data and $0 < \rho < 1$. The relation (3) is known as asymptotic regression model or monomolecular model and this model predicts the behaviour of S_w^2 with change in the value of M. It has been used extremely in agricultural, fisheries, psychological researches etc. The parameter a defines the asymptotic value of the function (3). Draper and Smith (1998) have classified it as an intrinsically non-linear model. The computation of its parameters can be made by non-linear least-squares estimation for which several statistical software are easily available. The suggested model does not increase without bound but approaches its asymptotic value a, therefore, it does not suffer from drawbacks as mentioned by Cochran (1999). Contrary to function (2), the suggested relation fits extremely well to data sets having large cluster sizes.

3. Determination of optimum cluster size

It is well known in cluster sampling that S^2, S_b^2 and S_w^2 are related as:

$$(NM-1)S^2 = N(M-1)S_w^2 + M(N-1)S_b^2$$

so that

$$S_b^2 = \frac{1}{M(N-1)}\left[(NM-1)S^2 - N(M-1)S_w^2\right] \tag{4}$$

If the whole population is considered as a single cluster, it will contain NM elements and the total variance in this case is

$$S^2 = a + b\rho^{NM} \tag{5}$$

Putting this value of S^2 from (5) in (4) and using (3), we have variance between clusters as:

$$S_b^2 = \frac{1}{M(N-1)}\left[(NM-1)(a+b\rho^{NM}) - N(M-1)(a+b\rho^M)\right] \tag{6}$$

Mahalanobis (1940, 1942) has determined the optimum cluster size using model (2) by minimizing the variance for fixed cost or minimizing the cost for fixed variance.

We are determining the optimum cluster size using suggested model. Let the cost function be defined as

$$C = C_0 + nMC_1 + C_2\sqrt{n} \tag{7}$$

Where C_0 is overhead cost, C_1 is the cost of enumerating an element including the cost of travel between elements within the same cluster and C_2 is the cost per unit distance traveled between clusters. Due to Mahalanobis (1940) and Jessen (1942) in practice

mostly C_1 is less than C_2 and minimum distance traveled between n randomly located points is proportional to $\sqrt{n}$.

Now we shall find the optimum value of M by minimizing the variance for the fixed cost of the survey. The variance function obtained from (1) using (4) after ignoring finite population correction (f.p.c.) is

$$V = \frac{S_b^2}{n} = \frac{1}{n}\left[S^2 - \frac{(M-1)}{M}(a + b\rho^M)\right] \tag{8}$$

Using Lagrange's multiplier the expression to be minimized is

$$\phi = C + \lambda V = C_0 + nMC_1 + C_2\sqrt{n} + \lambda V \tag{9}$$

Differentiating partially ϕ with respect to n & M and equating to zero the first order derivatives, we get

$$\frac{\partial \phi}{\partial n} = 0 \quad \Rightarrow \quad C_1 M + \frac{C_2}{2\sqrt{n}} = -\lambda \frac{\partial V}{\partial n}$$

$$\Rightarrow \quad C_1 M + \frac{C_2}{2\sqrt{n}} = -\lambda \frac{V}{n} \tag{10}$$

and

$$\frac{\partial \phi}{\partial M} = 0 \quad \Rightarrow \quad C_1 n = -\lambda \frac{\partial V}{\partial M} \tag{11}$$

Dividing (11) by (10), we get

$$\frac{C_1 n}{C_1 M + \frac{C_2}{2\sqrt{n}}} = -\frac{n}{V}\frac{\partial V}{\partial M}$$

or

$$\frac{M}{V}\frac{\partial V}{\partial M} = -\frac{1}{1 + \frac{C_2}{2\sqrt{n} C_1 M}} \tag{12}$$

As the total cost is fixed, so solving the quadratic equation (4) in $\sqrt{n}$, we get

$$\sqrt{n} = \frac{-C_2 + \sqrt{C_2^2 + 4C_1 M (C - C_0)}}{2C_1 M} \tag{13}$$

Putting this value of $\sqrt{n}$ in equation (12) and on simplifying, we get

$$\frac{M}{V}\frac{\partial V}{\partial M} = \left[1 + \frac{4C_1 M (C - C_0)}{C_2^2}\right]^{-\frac{1}{2}} - 1$$

Using (8), the above equation gives

$$\frac{M(M-1)b\rho^M \log \rho + (a + b\rho^M)}{MS^2 - (M-1)(a + b\rho^M)} = 1 - \left[1 + \frac{4C_1 M (C - C_0)}{C_2^2}\right]^{-\frac{1}{2}} \tag{14}$$

Now M can be obtained from above equation by iterative procedure and it can be done very easily though Statistical software and on substituting this value of M in equation (13), we can obtain the optimum value of n.

4. Empirical study

As an illustration, we have considered the following example. The table 1 shows estimated values of S_w^2 for different values of M using suggested model (3) and also for models of Jessen (1942), Mahalanobis (1942) and Hendricks (1944) expressed by equation (2). The estimated values of S_w^2 for different values of M using model (3) have been obtained using SPSS 17.0 software.

Data Source: [Sukhatme et.al. (1984), table 7.8, page 283]

Table 1

M	Observed values of S_w^2	Fitted values of S_w^2 For Model (2)	Fitted values of S_w^2 for Model (3)
2	78.10	81.53	79.84
4	84.28	84.25	82.71
8	88.92	87.05	87.60
16	93.50	89.95	94.75
NM=1176	108.33	110.22	108.17

It is observed from table1 that the suggested model (3) fits extremely well even to clusters of large sizes like 1176.

The table 2 gives the estimates of parameters a, b, ρ and residual mean squares (s^2) for model (3) and (2) for the above data set. It is observed from the table2 that the suggested model has considerably smaller value of s^2 as compare to model (2). Thus the suggested model gives more efficient result as compare to model (2).

Table 2

a	b	ρ	s^2 Model (2)	s^2 Model (3)
108.171	-31.53	0.948	10.4791	4.410

Efficiency Comparison

When N is large, the equation (4) can be written as:

$$S_b^2 = S^2 - \frac{(M-1)}{M} S_w^2 \qquad (15)$$

The estimates of variance of cluster means have been calculated using relation (15) and relative efficiency of the variance of cluster means based on suggested model (3) with model (2) have been calculated, which are shown in following table 3.

Table 3

M	2	4	8	16
Efficiency	101.766	101.946	108.031	133.864

It is evident from table that use of suggested model gives more efficient estimates of variance of cluster means as compared to existing model.

5. Conclusions

The drawbacks of existing model that it increases without bound and does not fit well to large cluster sizes have been removed by the suggested model. It not only describes S_w^2 in a better way but also attains an upper bound. It fits well to very large values of cluster sizes. As it provides more efficient estimates of variance of cluster means as compared to existing model, the optimum value of cluster size based on suggested model would also be more precise.

Acknowledgement

The authors are highly thankful to the referees for their valuable suggestions for improving this manuscript. The financial assistance of University Grant Commission, New Delhi, India under Major Research Project F.No.33-56/2007 (SR) is also acknowledged.

References

1. Cochran,W.G.(1999). Sampling Technques. 3rd Ed., John Wiley & Sons.
2. Draper, N.R. and Smith, H.(1998). Applied regression analysis. 3rd Ed., John Wiley & Sons.
3. Hansen, M.H. and Hurwitz, W.N. (1942). Relative efficiencies of various sampling units in population enquiries. J A S A., 37, p. 89-94.
4. Hendricks, W.A. (1944). The relative efficiencies of groups of farms as sampling units. J A S A., 39, p. 366-376.
5. Jessen, R.J. (1942). Statistical investigation of sample survey for obtaining farm facts. Iowa Agricultural Experiment Station, Research Bulletin, 304.
6. Mahalanobis, P.C. (1940). A sample survey of acreage under jute in Bengal. Sankhya, 4, p. 511-530.
7. Mahalanobis,P.C. (1942).General report on the sample census of area under jute in Bengal. Indian Central Jute Committee.
8. McVay, F.E. (1947) Sampling methods applied to estimating numbers of commercial orchards in a commercial peach area. J A S A., 42, p. 533-540.
9. Smith, H.F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. J. Agr, Sci., 28, p.1-23.
10. Sukhatme,P.V., Sukhatme, B.V., Sukhatme S. and Asok, C.(1984). Sampling theory of surveys with applications. Indian Society of Agricultural Statistics.