Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8160

Research Paper | Agriculture

Comparison of Cluster Sampling with Simple Random Sampling in context to Agriculture

M. Iqbal Jeelani	Division of Agricultural Statistics, S.K. University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar 191121
S. Maqbool	Division of Agricultural Statistics, S.K. University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar 191121
Nageena Nazir	Division of Agricultural Statistics, S.K. University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar 191121
S. A. Mir	Division of Agricultural Statistics, S.K. University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar 191121
I. Khan	Division of Agricultural Statistics, S.K. University of Agricultural Sciences and Technology of Kashmir, Shalimar, Srinagar 191121

ABSTRACT

The present work is an attempt to show that cluster sampling is more efficient than simple random sampling provided the mean square within the clusters is maximum and there is a negative intra-class correlation coefficient between elements within clusters as relative efficiency of cluster sampling increases with increase in mean square within clusters. Different estimators of cluster sampling are applied and their results are compared with simple random sampling using the same sample size. Different computer programmes are developed using R-software. All these functions are run on real data set generated on Apple crop from district Ganderbal of Kashmir valley.

KEYWORDS: Cluster sampling, simple random sampling, intra-class correlation coefficient, R-software.

1. INTRODUCTION

Cluster sampling is a sampling technique used when natural groupings are evident in a population. Cluster sampling is operationally more convenient, less time consuming and importantly cost wise efficient as compared to simple random sampling**.** In cluster sampling problems like imperfect sampling frames, improper stratification, hidden periodicity etc do not arise. Cluster sampling is a technique where the entire population is divided into groups or clusters and a random sample of the selected clusters are included in the sample. As a simple rule, the number of units in a cluster should be small and the number of clusters should be large, but there should be homogeneity between cluster means. Each cluster should be a small scale representation of total population.

The efficiency of cluster sampling has been studied by Smith (1938), where it has been discussed that the relative efficiency of cluster sampling increases with the increase in mean square within clusters. On the basis of many agricultural surveys Jessen (1942) and Mahalonobis (1944) developed a general law to predict how mean square within clusters changes with the size of cluster. Hansen and Hurtwiz, (1944) discussed that in many practical situations, cluster size is positively correlated with the variable under study and in these cases, it is advisable to select the clusters with probability proportional to the number of elements in the cluster. A good discussion of numerical values of intra-class correlation coefficient for different elements within cluster in cluster sampling have been given by Hurtwiz and Madow (1953), they have shown the intra-class correlation coefficient as a "measure of homogeneity" of the clusters in cluster sampling.

In this paper a description of cluster sampling as compared to simple random sampling in reference to apple data has been given. The computations have been done by newly developed functions in R-software the details of the function are given in the appendix-I and only names of the functions are depicted in the paper. A complete description of R-software is given in Pinheiro and Bates (2007).

2. ESTIMATION PROCEDURE OF CLUSTER SAMPLING

Suppose the populations consist of *N* clusters, each of *M* elements, and that a sample of *n* clusters is drawn by the method of simple random sampling:

N =Number of clusters in the population, *n* = Number of clusters in the sample, y_{ij} = Value of the characteristic under study for the jth element ,*(j= 1,2,3.....,M)* in the *i*th cluster, $\bar{y}_i = \sum_j^M y_{ij}/M$ = Mean of per element of the *i*th cluster, $\bar{y}_n = \sum_i^n \bar{y}_i/n$ = Mean of cluster means in a sample of n clusters, $\bar{Y}_N = \sum_i^N \bar{y}_i/N$ = Mean of cluster mean in the population, $\bar{Y} = \sum_i^N \sum_j^M y_{ij}/NM$ = Mean per element in the population, $S_i^2 = \sum_j^M (y_{ij} - \bar{y}_i)^2/(M-1)$= Mean square between elements within the *ith* cluster(*i=1,2,.....,N*), $S_w^2 = \sum_i^N S_i^2/N$ = Mean square within clusters, (*w* for within), $S_b^2 = \sum_i^N (\bar{y}_i - \bar{Y}_N)^2/(N-1)$ = Mean square between cluster means in the population(b for between), $S^2 = \sum_i^N \sum_j^M (y_{ij} - \bar{Y})^2/(NM-1)$ = Mean square between elements in the population, $\rho = \frac{E(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})}{E(y_{ij}-\bar{Y})^2} = \rho = \frac{E(y_{ij}-\bar{Y})(y_{ik}-\bar{Y})}{E(y_{ij}-\bar{Y})^2}$ = Intra-cluster correlation coefficient between elements within clusters.

The variance in the cluster sampling depends on the number of clusters in the sample, the size of the cluster, the Intra-cluster correlation coefficient and the variance S^2. If the $M=1$ it gives the sampling variance of a simple random sample of Nm elements. If $M>1$ and is positive, cluster sampling will give a higher variance than the mean per element. If the is negative, there is a reason to use cluster sampling.

3. RELATIVE EFFICIENCY AND INTRA-CLASS CORRELATION COEFFICIENT:

In sampling *nM* elements from the population by simple random sampling, the variance of the sample mean is given by;

$$V(\bar{y}) = \frac{(1-f)S^2}{nM} \tag{3.1}$$

Thus, the relative efficiency of cluster sampling compared with simple random sampling is given by;

$$Relative\ Efficiency = \frac{V_{SR}(\bar{y})}{V_C(\bar{y}_n)} = \frac{S^2}{M\,S_b^2} \tag{3.2}$$

where,

$$MS_b^2 = [(NM-2)S^2 - N(M-1)S_w^2]/(N-1) \tag{3.3}$$

Therefore, relative efficiency will increase with increase in the mean square within clusters. For large *N*, the relative efficiency of cluster sampling in terms of Intra-cluster correlation coefficient is given by;

$$Relative\ Effeciency(E) = [1+(M-1)\rho]^{-1} \tag{3.4}$$

where, $\hat{\rho} = \frac{(n-1)Ms_b^2 - ns_w^2}{(n-1)Ms_b^2 + n(M-1)s_w^2}$ (3.5)

Where, $s_w^2 = \sum_i^n \sum_j^M \frac{(y_{ij}-\bar{y}_i)^2}{n(M-1)}$ (3.6)

Thus, for large *N*, an estimate of the relative efficiency of cluster sampling can be written as:

$$Est.\ Relative\ Efficiency(e) = \frac{1}{M} + \frac{(M-1)s_w^2}{M^2 s_b^2} \tag{3.7}$$

And accordingly can be estimated by,

$$\hat{\rho} = \frac{(1-e)}{(M-1)e} \tag{3.8}$$

4. ESTIMATION PROCEDURE OF SRSWOR (SIMPLE RANDOM SAMPLING WITHOUT REPLICEMENT):

Simple random sampling with replacement *(srswr)* can be regarded as sampling from an infinite population and its variance is given by ; $var(\bar{x})_{srswr} = \frac{N-1}{N}\frac{S^2}{n}$. Comparing the variance of *srswr* with *srswor*, we get $var(\bar{x})_{srswr} < var(\bar{x})_{srswor}$, *i.e., the variance of sample mean is less in srswor as compared with its variance in the case of srswr.* This implies that *srswor* provides a better (more efficient) estimate of the population mean μ relative to *srswr*. That is why in this paper only results of *srswor* are compared with cluster sampling.

5 - NUMERICAL ILLUSTRATION

The methodologies discussed in the earlier sections have been implemented on newly developed functions in R-software. One of the important feature of R- software is that it is an Open Source and freely available on website *http://cran-project.org*. Khan and Mir (2005) discuss in detail the application of R- software in agricultural data analysis. R language is essentially a functional language for all practical purposes of data analysis and graphics. However, in case some specific situations data analyst is forced to develop his own functions according to his requirements. Consequently, few functions have been developed according to the requirements of this study. For the proposed sampling scheme (cluster sampling) four functions (cluster1(x,N), cluster2(x), cluster3(x), SRSWOR(Y, N)) have been developed which are given in appendix-I.

The apple data taken from district Ganderbal at block level of Kashmir valley 2010-2011 is used. The reported 420 orchards from the block are divided into 105 clusters, 4 orchards in each cluster. 15 clusters are taken randomly as sample clusters, i.e., out of 420 orchards 60 orchards were selected randomly from all the variables, including yield, area and total number of trees. The results of cluster sampling and SRSWOR are given below:

[TABLE 1 here]

[TABLE 2 here]

6- COMPARISON OF CLUSTER SAMPLING WITH SRSWOR

[TABLE 3 here]

[TABLE 4 here]

[TABLE 5 here]

In the present data it can be seen from the tables given above, that using the same sample size (i.e., 60 orchards), even though the estimates in both the sampling schemes are almost same, but as far as their estimated variances and standard errors are concerned they are much lower for cluster sampling as compared to simple random sampling. Indicating that the clusters which were selected randomly from all the three variables (i.e., yield, area and trees) have maximum variation internally, which means that mean square within clusters is much larger than mean square between clusters (suggesting that the elements within a cluster are dispersed more than a randomly chosen group would be, which means cluster means are nearly identical implying that mean square between clusters in minimum.

7 - COMPARISION OF RELATIVE EFFICIENCIES

In the present data relative efficiencies of cluster sampling in all the three variables is above 100% , which suggests that using the same sample size (i.e., 60 orchards) cluster sampling is more efficient than simple random sampling due to the fact that the mean square within the clusters is maximum and there is a negative intra-class correlation coefficient between elements within clusters, because relative efficiency of cluster sampling increases with increase in mean square within clusters (i.e., clusters should be so formed that the variation within clusters is maximum, while variation between clusters is minimum). The relative efficiencies along with their intra class correlation coefficient of cluster sampling of all the three variables are given below;

[TABLE 6 here]

CONCLUSION

From the results of comparison it is concluded that using the same sample size the larger values of standard errors and estimated variances were found in simple random sampling as compared to cluster sampling on comparison, which suggests cluster sampling provides more efficient results than simple random sampling provided the clusters are formed in such a way that there is maximum heterogeneity within clusters and maximum homogeneity between clusters. Finally cluster sampling procedures are more attractive than its counter parts as they are less time consuming and operationally more convenient. It has been established that cluster sampling has its practical implications. It can be applied to analyze data generated in a scientific investigation. R-software package facilitates a lot in implementation of cluster sampling techniques which are very informative and applicable to sample surveys of horticultural crops which lack a proper sampling frame.

TABLES

Table1.- Results of cluster sampling

Name of Variable	Cluster mean	Cluster variance	Standard error	MSB	MSW
Yield (mt)	14.55	0.24	0.49	4.21	44.3
Area (ha)	1.66	0.004	0.06	0.07	0.51
Trees (no.)	158.18	28.84	5.37	504.79	5140.8

Table2-Results of SRSWOR

Name of Variable	Mean	Variance	Estimated Variance	Standard error
Yield (mt)	14.555	37.83	0.5404	0.735
Area (ha)	1.665	0.457	0.00653	0.080834
Trees (no.)	158.18	4400.118	62.85883	7.928356

Table3-Comparison of Cluster Sampling With SRSWOR (Yield (Mt))

Sampling Design	Sample size	Estimate	Est variance	Standard Error
SRSWOR	60	14.702	0.562	0.750
Cluster Sampling	15 clusters	14.551	0.240	0.490

Table4-Comparison of Cluster Sampling With SRSWOR (Area(ha))

Sampling Design	Sample size	Estimate	Est variance	Standard Error
SRSWOR	60	1.665	0.0065	0.0818
Cluster Sampling	15 clusters	1.673	0.0040	0.06381

Table5-Comparison of Cluster Sampling With SRSWOR (Trees)

Sampling Design	Sample size	Estimate	Est variance	Standard Error
SRSWOR	60	158.18	62.858	7.928
Cluster Sampling	15 clusters	158.18	28.845	5.3708

Table6-Relative efficiency along with Intraclass correlation coefficients

Name of variable	Relative Efficiency	Intra-class correlation coefficient
Yield (mt)	233.7%	-0.184846
Area (ha)	160.4%	-0.125558
Trees (no.)	217.9%	-0.1803681

Appendix-I

cluster1(x,N)

This is function developed in R-software. It takes the arguments x (name of the data), N (total number of clusters) and returns cluster mean (ynbar), cluster variance (vynbar) and standard error (se).The codes of the function follow

```
cluster1<-function(x,N)

{

x=data.frame(x)

#N=total no. of clusters in the population

n=nrow(x)
m=ncol(x)
yibar=apply(x,1,mean)
myibar2=sum(m*yibar^2)

si2=apply(x,1,var)

ynbar=sum(yibar)/n

yibar2=sum(yibar^2)

nybar2=ynbar^2

vynbar=((1/n)-(1/N))*(1/(n-1))*(yibar2-n*nybar2)

se=sqrt(vynbar)

list(clusterMean=ynbar,ClusterVariance=vynbar,se=se)

}

cluster2(x)
```

This is another function developed in R-software. It takes the arguments (x) i.e., the name of the data frame only. The function is of prime importance as far as cluster sampling is concerned, because this function reflects how much of the variation is present in the clusters. It returns mean square between clusters (msb2), mean square within clusters (sw2), total variance (S2) and their respective degrees of freedom.

```
> cluster2=function(x)

{

x=data.frame(x)

n=nrow(x)

m=ncol(x)

yibar=apply(x,1,mean)

si2=apply(x,1,var)

ynbar=sum(yibar)/n

yibar2=sum(yibar^2)

nybar2=ynbar^2

msb2=(1/(n-1))*(yibar2-n*nybar2)

sw2=sum(si2/n)

s2=((n-1)*m*msb2)+(n*(m-1)*sw2)

S2=s2/((n*m)-1)

dfmsb2=n-1

sw2=sum(si2/n)

dfsw2=n*(m-1)

s2=((n-1)*m*msb2)+(n*(m-1)*sw2)

S2=s2/((n*m)-1)

dfS2=n*m-1      list(DFBetweenClusters=dfmsb2,DFWithinClusters=dfsw2,DFTotal=dfS2,betweenCluSSq=msb2,WithinCluSSq=sw2,Total=S2)

}

Cluster3(x)
```

This is another function developed in R-software. It takes the same arguments of the previous function. This function returns relative efficiency (re) and intra- class correlation coefficient (rho).

```
cluster3=function(x)

{

x=data.frame(x)
```

```
n=nrow(x)
m=ncol(x)
yibar=apply(x,1,mean)
si2=apply(x,1,var)
ynbar=sum(yibar)/n
yibar2=sum(yibar^2)
nybar2=ynbar^2
msb2=(1/(n-1))*(yibar2-n*nybar2)
sw2=sum(si2/n)
s2=((n-1)*m*msb2)+(n*(m-1)*sw2)
S2=s2/((n*m)-1)
re=S2/(m*msb2)
rho=(1-re)/((m-1)*re)
list(Relative Efficiency=re, EstimatorRho=rho)
}
```

> SRSWOR1<-function (Y,N)

A function for estimation in simple random sampling without replacement. This function returns estimates with standard error and 95% confidence interval. It requires only data vector and *N*. *Y* is the data vector of responses on n units sampled. *N* is the population size

```
{
 n<-length (Y)
 Ybar<-mean (Y)
 S2<-var (Y)
 Estvar<-((N-n)/(N*n))*S2
 SeEst<-sqrt(Estvar)
 ci<-c(Ybar-1.96*SeEst,Ybar+1.96* SeEst)
 ci<-c(ci[1],Ybar,ci[2])
 names(ci)<-c("lower""estimate""upper")
 out1<-c(Estimate=Ybar,SE=SeEst)
 out1<-round(out1,3)
 out2<-ci
 out3=round(out2,3)
     list(estimates=out1,"95%cont.interval"=out3,mean=Ybar, variance=S2,estimatedvariance= Estvar, standarderror=SeEst)
```

REFERENCES

1.Hansen,M.H. and Hurwitz,W.N. (1944). On the theory of sampling from finite populations. Ann. Math. Stat., 14, 333-362. | 2.Hurtwiz, W.N. & Madow , W.G.(1953) . Sample Survey Methods and Theory. New York: John Willey and sons. | 3.Jessen, A. (1942). Purposive selection. Jour. Roy Stat. Soc., 91, 541-547. | 4.Khan ,A.A. and Mir,A.H.(2005). Applications of R-Software in Agricultural Data Analysis. SKUAST Research Journal. | 5.Mahalanobis, P.C.(1944). Report on the sample census of jute in Bengal. Indian Central Jute Committee. | 6.Pinheiro, J.C. and Bates, M.D. 2007. Mixed-Effects Models in S and S-PLUS. Springer-Verlag New York | 7.Smith, T.M.F. (1938). The foundations of survey sampling: a review. Jour. Roy. Stat. Soc. A, 139, 183-204.