Optimum Stratification in the Estimation of Pooled Variance under Proportional Allocation

P. B. Bharate[1*], P. C. Gupta[2]

[1] Associate Professor and Head, Department of Statistics, Pratap College Amalner, Dist. Jalgaon, Maharashtra, INDIA.

[2] Ex. Professor and Head, Department of Statistics, Veer Narmad South Gujarat University, Surat, Gujrat, INDIA.

[*]Corresponding address:

pbbharate@gmail.com

Research Article

Abstract: Present paper deals with the problem of optimum stratification when the parameter of interest is pooled variance. In this paper, an estimator of pooled variance is suggested and minimal equations to get optimum boundary points are obtained by minimizing the variance of the estimator under proportional allocation. The stratification is done on the study variable. Further, by assuming rectangular distribution within each stratum, optimum boundary points are obtained. Approximate boundary points are also obtained by deriving the approximate minimal equations on the line of Ekman (1958).

Key words**:** optimum stratification, study variable, minimal equations, and optimum boundary points.

1 Introduction

One of the basic design operations of stratified random sampling, construction of strata is very important. The efficiency of design can be affected by construction of strata. The construction of strata raises several questions. What should be the best characteristic for construction of strata? How should the boundaries between the strata be determined? How many strata should there be? If the knowledge of number of strata and the type of allocation is assumed the variance of the estimator of population mean (total) remains only a function of W_h and σ_h, the weight and the variance of hth stratum; h=1,2,…..L. For a given population, these two sets of parameters depend upon the choice of boundary points. The change in stratification changes the values of boundary points, which affects the variance of the estimator. Therefore, problem of optimum stratification is to choose one set of boundary points from all such sets, which give the minimum variance. A faulty choice of stratification may lead to considerable increase in the variance of the estimator. An appropriate choice of stratum boundary points is, therefore, of utmost importance. Dalenius (1950) was the first to discuss the problem of optimum stratification. He assumed the knowledge of frequency distribution of study variable and stratified on same variable for proportional and Neyman allocation. Dalenius and Gurney (1951) suggested the use of auxiliary variable for determination of optimum stratification boundary points (OSBP). Mahalanobis (1952) suggested a rule known as "Equal Aggregate Method". According the rule "an optimum or nearly optimum solution would be obtained when the expected contribution of each stratum to the total aggregate value of y is made equal to all strata"

i.e. $$W_h\mu_{hy} = \text{Constant}$$

Ayoma(1954) derived equidistance stratification by applying the mean value theorem to the equations obtained by Dalenius (1950). Kitagawa (1956) gave name to Mahalonobis's (1952) suggestion as "Principle of equipartition". Dalenius and Hodges J.L. (1959) suggested "cumulative square root method" as an approximation to OSBP under optimum allocation. In this method, the cum $\sqrt{f}$ is divided in as many equal parts as the number of strata to be formed. i. e. cum$\sqrt{f}$ = constant.

Ekman (1959) has given a method of finding AOSB for proportional allocation when the stratification is done on the study variable. Cochran (1961) observed that the cumulative square root rule used to obtain AOSB works extremely well with both theoretical as well as actual distributions. Sethi (1963) studied the problem from different angle and gave ready-made tables giving strata boundaries for some standard distributions. Des Raj (1964) obtained the equations giving the OBP for equal allocation by minimizing the variance of the stratified mean per unit estimator under equal distribution of sample to each stratum. Gupta (1970) also studied optimum stratification in case of ratio and product method of estimation, which minimizes the generalized variance of the estimates of mean of more than one character based on auxiliary character X under the proportional allocation. Singh (1971) gave the method of finding approximate OBP. Rajyaguru (1999) and Rajyaguru and Gupta (2002) have suggested an alternative aspect of optimum stratification. Instead of minimizing the variance of the stratification estimator, they minimized the weighted square co-efficient of variation under proportional allocation both when stratification variable is study variable and when it is different. Singh and Sukhatme (1969) proposed cube root of the probability function rule for determining strata boundaries. Lavallée and Hidiroglou (1988) derived an iterative procedure for stratifying skewed populations into a take-all stratum and a number of take-some strata such that the sample size is minimized for a given level of reliability. Other recent contributions include Hedlin (2000) who revisited Ekman's rule, Dorfman and Valliant (2000) who compared model based stratified sampling with balanced sampling, and Rivest (2002) who constructed a generalisation of the Lavallée and Hidiroglou algorithm by providing models accounting for the discrepancy between the stratification variable and the survey variable.

All the workers have considered the parameter of interest as mean (total). In the present paper we intend to study the problem of optimum stratification when parameter is pooled variance. We have suggested an estimator of pooled variance and minimal equations for OSB are obtained by minimizing the variance of the estimator for proportional allocation. This aspect of stratification is studied when stratification variable is study variable.

2 Theorem

Estimator Of Pooled Variance σ ^{2}

Pooled sample variance s^2 as an unbiased estimator of pooled variance σ^2.

$$E(s^2) = \sigma^2$$

where $s^2 = \sum_{h=1}^{L} W_h s_h^2$ =pooled sample variance and $\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2$ = Pooled population Variance.

Proof: Since the sample is drawn from each stratum by simple random sampling and for simple random sampling $E(s_h^2) = \sigma_h^2$

We have

$$E(s^2) = E\left[\sum_{h=1}^{L} W_h s_h^2\right]$$

$$E(s^2) = \sum_{h=1}^{L} W_h E(s_h^2)$$

$$E(s^2) = \sum_{h=1}^{L} W_h \sigma_h^2$$

$$= \boldsymbol{\sigma^2}$$

Hence, the pooled sample variance is an unbiased estimator of pooled population variance.

3 Variance Of The Estimator (s^2)

$$Var(s^2) = Var\left(\sum_{h=1}^{L} W_h s_h^{\ 2}\right)$$

$$= \sum_{h=1}^{L} W_h^2 Var(s_h^2)$$

Under the assumption of non-normality within each stratum and since sample is drawn by SRS, we have,

$$Var(s_h^2) = \frac{2\sigma_h^4}{n_h - 1}\left[1 + \frac{(n_h - 1)}{2n_h}(\beta_{2h} - 3)\right] \tag{1}$$

But $\beta_{2h} = \dfrac{\mu_{4hy}}{\sigma_h^4}$

$$\therefore Var(s^2) = \sum_{h=1}^{L} W_h^2 \frac{2\sigma_h^4}{n_h - 1}\left[1 + \frac{(n_h - 1)}{2n_h}(\beta_{2h} - 3)\right]$$

$$= 2\sum_{h=1}^{L} \frac{W_h^2 \sigma_h^4}{n_h - 1} + \sum_{h=1}^{L} \frac{W_h^2 \mu_{4h}}{n_h} - 3\sum_{h=1}^{L} \frac{W_h^2 \sigma_h^4}{n_h}$$

If we assume $\dfrac{n_h\text{-}1}{n_h} \cong 1$, then

$$Var(s^2) = \sum_{h=1}^{L} \frac{W_h^2 \mu_{4h}}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 \sigma_h^4}{n_h} \tag{2}$$

4 Mathematical Formulation of the Problem

We shall make following assumptions.

The variable Y has a continuous probability density function *f(y)* and the first four moments of Y exist.Population is infinite.Though, these assumptions will not, in general, be satisfied, yet in practice they will be approximately satisfied.

Let - ∞, y_1 , ,y_2,…….y_{h-1},y_h ,y_{h+1},……y_{L-1}, ∞ denote the boundaries of L strata, where L is fixed in advance.

In hth stratum, we define

$$W_h = \int_{y_{h-1}}^{y_h} f(y)\,dy$$

=proportion of population units in the hth stratum (3)

$$\mu_{hy} = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} yf(y)\,dy$$

=mean of the character y for hth stratum (4)

$$\sigma_{hy}^2 = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} \left(y - \mu_{hy}\right)^2 f(\mathrm{y})\,\mathrm{dy}$$

= variance of the character y for the hth stratum (5)

$$\mu_{4hy} = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} (y - \mu_{hy})^4 f(y)\,dy$$

=Fourth central moment of character y (6)

Our problem is to find strata boundary points, y_1,y_2,……y_{h-1}, y_h,y_{h+1},……y_{L-1} such that Var(s^2) is minimum. We note that as y_h changes then W_h, W_{h+1}, μ_h, μ_{h+1},σ_h, σ_{h+1}, μ_h, μ_{h+1} change. Also n_h,n_{h+1} may change**.**

Further, from (5) we have

$$\frac{\partial}{\partial y_h}\left(W_h \sigma_{hy}^2\right) = \left(y_h - \mu_{hy}\right)^2 f\left(y_h\right) \tag{7}$$

5 Stratification under Proportional Allocation

In proportional allocation $n_h=nW_h$. Thus, from the equation (2), we get

$$Var\left(s^2\right)=\frac{1}{n}\sum_{h=1}^{L}W_h\mu_{4hy}-\frac{1}{n}\sum_{h=1}^{L}W_h\sigma_{hy}^4 \tag{8}$$

Differentiating (8) partially with respect to y_h and equating to zero, we get

$$\frac{\partial}{\partial y_h}\left(W_h\mu_{4hy}\right)+\frac{\partial}{\partial y_h}\left(W_{h+1}\mu_{4h+1y}\right)$$

$$=\frac{\partial}{\partial y_h}\left(W_h\sigma_{hy}^4\right)+\frac{\partial}{\partial y_h}\left(W_{h+1}\sigma_{h+1y}^4\right) \tag{9}$$

Since terms not involving y_h are constant with respect to differentiation, these terms vanish.

From (6), we have

$$\frac{\partial}{\partial y_h}\left(W_h\mu_{4hy}\right)=\left(y_h-\mu_{hy}\right)^4 f\left(y_h\right) \tag{10}$$

And similarly

$$\frac{\partial}{\partial y_h}\left(W_{h+1}\mu_{4h+1y}\right)=-\left(y_h-\mu_{h+1}\right)^4 f\left(y_h\right) \tag{11}$$

Again

$$\frac{\partial}{\partial y_h}\left(W_h\sigma_{hy}^4\right)$$

$$=\sigma_{hy}^4\frac{\partial W_h}{\partial y_h}+4W_h\sigma_{hy}^3\frac{\partial\sigma_{hy}}{\partial y_h} \tag{12}$$

We have (7) as below

$$\frac{\partial}{\partial y_h}\left(W_h\sigma_{hy}^2\right)=\left(y_h-\mu_{hy}\right)^2 f\left(y_h\right)$$

$$\therefore\ 2W_h\sigma_{hy}\frac{\partial\sigma_{hy}}{\partial y_h}+\sigma_{hy}^2\frac{\partial W_h}{\partial y_h}=\left(y_h-\mu_{hy}\right)^2 f\left(y_h\right)$$

Since $\frac{\partial W_h}{\partial y_h}=f(y_h)$

$$2W_h\sigma_{hy}\frac{\partial\sigma_{hy}}{\partial y_h}=\left[\left(y_h-\mu_{hy}\right)^2-\sigma_{hy}^2\right]f\left(y_h\right)$$

$$\therefore\frac{\partial\sigma_{hy}}{\partial y_h}=\frac{\left[\left(y_h-\mu_{hy}\right)^2-\sigma_{hy}^2\right]}{2W_h\sigma_{hy}}f\left(y_h\right) \tag{13}$$

Putting value from (13) in (12), we get

$$\frac{\partial}{\partial y_h}\left(W_h\sigma_{hy}^4\right)=$$

$$=\sigma_{hy}^4 f(y_h)+4W_h\sigma_{hy}^3\frac{\left[\left(y_h-\mu_{hy}\right)^2-\sigma_{hy}^2\right]}{2W_h\sigma_{hy}}f\left(y_h\right)$$

$$\frac{\partial}{\partial y_h}\left(W_h\sigma_{hy}^4\right)=\sigma_{hy}^2\left[2(y_h-\mu_{hy})^2-\sigma_{hy}^2\right]f(y_h) \tag{14}$$

Similarly, $\frac{\partial}{\partial y_h}\left(W_{h+1}\sigma_{h+1y}^4\right)=$

$$=-\sigma_{h+1y}^2\left[2(y_h-\mu_{h+1y})^2-\sigma_{h+1y}^2\right]f(y_h) \tag{15}$$

Putting the values from (10), (11), (14) and (15) in (9), we get

$$(y_h-\mu_{hy})^4-(y_h-\mu_{h+1y})^4$$

$$=\sigma_{hy}^2[2(y_h-\mu_{hy})^2-\sigma_{hy}^2]\ -\sigma_{h+1y}^2[2(y_h-\mu_{h+1y})^2-\sigma_{h+1y}^2]$$

$$\therefore\ \left[(y_h-\mu_{hy})^2-\sigma_{hy}^2\right]^2=\left[(y_h-\mu_{h+1y})^2-\sigma_{h+1y}^2\right]^2$$

$$\therefore\ \left[(y_h-\mu_{hy})^2-\sigma_{hy}^2\right]=\left[(y_h-\mu_{h+1y})^2-\sigma_{h+1y}^2\right] \tag{16}$$

Equations (16) are the minimal equations by solving which the optimum boundary points can be obtained. Therefore, the boundary points, in this case, are given by following relation

$$y_h=\frac{(\sigma_{hy}^2-\sigma_{h+1}^2)+(\mu_{h+1y}^2-\mu_{hy}^2)}{2(\mu_{h+1y}-\mu_{hy})}$$

$$h=1,2,.....L\text{-}1 \tag{17}$$

Equations (17) give the solution to obtain optimum boundary points. It appears that there exists a trivial solution of equation (17) namely that we use the OBP obtained from $y_h=\frac{\mu_{hy}+\mu_{h+1y}}{2}$ under the assumption $\sigma_{hy}^2=\sigma_{h+1y}^2\forall h$, which is definitely against the logic behind stratification and as such it will always lead to poor stratification.

Remark If we assume rectangular distribution in each stratum then,

$$\mu_h=\frac{y_h+y_{h-1}}{2}\ \text{ and }\ \sigma_{hy}^2=\frac{(y_h-y_{h-1})^2}{12}$$

Then (17) reduces to

$$y_h=\frac{y_{h-1}+y_{h+1}}{2}\qquad\text{for h=1,2,…L-1} \tag{18}$$

Assumption of rectangular distribution leads to the equidistance partition.

The solutions to Equations (16) give optimum boundary points $\{y_h\}$. Here it is seen that the nature of the minimal equations is implicit. Therefore, the determination of optimum boundary points $\{y_h\}$ satisfying the equations (16) is quite difficult. It is necessary to find some approximate method of finding the OBP. In the next section, the approximate method of finding the optimum boundary points is derived.

6 Approximation

On the line of Ekman(1958), we have

$$W_h(y_h-\mu_h)=\left[\frac{(y_h-y_{h-1})^2}{2!}H'''(\mathrm{y}_{h-1})+\frac{(y_h-y_{h-1})^3}{3!}H^{(4)}(\mathrm{y}_{h-1})+\frac{(y_h-y_{h-1})^4}{4!}H^{(5)}(\xi_1)\right] \tag{19}$$

$$W_h=\left[\frac{(y_h-y_{h-1})}{1!}H'''(\mathrm{y}_{h-1})+\frac{(y_h-y_{h-1})^2}{2!}H^{(4)}(\mathrm{y}_{h-1})+\frac{(y_h-y_{h-1})^3}{3!}H^{(5)}(\xi_2)\right] \tag{20}$$

where ξ_1 and ξ_2 are points in the interval (y_{h-1},y_h)

Squaring (20) we get

$$W_h^2 = (y_h - y_{h-1})^2 H'''(\mathrm{y}_{h-1})\left[H'''(\mathrm{y}_{h-1}) + (y_h - y_{h-1})H^{(4)}(\mathrm{y}_{h-1}) + ...\right] \tag{21}$$

And squaring (19)

$$[W_h(y_h - \mu_h)]^2 = \left[\frac{(y_h - y_{h-1})^4}{4} H'''^2(\mathrm{y}_{h-1}) + \frac{(y_h - y_{h-1})^5}{3!} H'''(\mathrm{y}_{h-1})\, H^{(4)}(\mathrm{y}_{h-1}) + \mathrm{R}_1\right] \tag{22}$$

$$= \frac{(\mathrm{y}_h - y_{h-1})^4}{12}\left\{H'''(\mathrm{y}_{h-1})\left[3H'''(\mathrm{y}_{h-1}) + 2(\mathrm{y}_h - y_{h-1})H^{(4)}(\mathrm{y}_{h-1})\right] + R_1\right\}$$

$$[W_h\sigma_h]^2 =$$

$$\frac{(\mathrm{y}_h - y_{h-1})^4}{12}\left\{H'''(y_{h-1})\left[H'''(y_{h-1}) + (y_h - y_{h-1})H^{(4)}(y_{h-1})\right] + R_2\right\} \tag{23}$$

Subtracting (23) from (22) we get

$$[W_h(y_h - \mu_h)]^2 - (W_h\sigma_h)^2 =$$

$$\frac{(y_h - y_{h-1})^4}{12}\left\{H'''(y_{h-1})\left[2H'''(y_{h-1}) + (y_h - y_{h-1})H^{(4)}(\mathrm{y}_{h-1})\right]\right\} \tag{24}$$

Dividing (24) by (21), we have

$$\frac{[W_h(y_h - \mu_h)]^2 - (W_h\sigma_h)^2}{W_h^2}$$

$$\frac{(y_h - y_{h-1})^2}{12}\frac{\left\{\left[2H'''(y_{h-1}) + (y_h - y_{h-1})H^{(4)}(\mathrm{y}_{h-1})\right]\right\}}{\left[H'''(\mathrm{y}_{h-1}) + (y_h - y_{h-1})H^{(4)}(\mathrm{y}_{h-1}) + ...\right]}$$

$$= \frac{(y_h - y_{h-1})^2}{12}\cdot\frac{2f(\mathrm{y}_{h-1}) + (y_h - y_{h-1})f'(\mathrm{y}_{h-1})}{f(\mathrm{y}_{h-1}) + (y_h - y_{h-1})f'(\mathrm{y}_{h-1})}$$

$$= \frac{(y_h - y_{h-1})}{24}\cdot\frac{(y_h - y_{h-1})f(\mathrm{y}_{h-1}) + \frac{(y_h - y_{h-1})^2 f'(\mathrm{y}_{h-1})}{2}}{f(\mathrm{y}_{h-1}) + (y_h - y_{h-1})f'(\mathrm{y}_{h-1})} \tag{25}$$

In the numerator of the second factor of the right hand member of (25) we have the first two terms of the Taylor expansion of W_h about the point y= y_{h-1}, where as the denominator is partial derivative of the numerator with respect of y_h i.e .the first two terms of the expansion of $f(y_h)$ in the same point. Approximating once again by neglecting terms of greater order in the expansions of W_h and $f(y_h)$, we obtain finally

$$\frac{[W_h(\mathrm{y}_h - \mu_h)]^2 - (W_h\sigma_h)^2]}{W_h^2} \square \frac{(y_h - y_{h-1})W_h}{24f(\mathrm{y}_h)}$$

$$i.e.(\mathrm{y}_h - \mu_h)^2 - {\sigma_h}^2 \square \frac{(y_h - y_{h-1})W_h}{24f(\mathrm{y}_h)}$$

similarly,

$$(\mathrm{y}_{h+1} - \mu_{h+1})^2 - {\sigma_{h+1}}^2 \square \frac{(y_{h+1} - y_h)W_{h+1}}{24f(\mathrm{y}_h)}$$

Therefore, the minimal equations (16) reduce to

$$\frac{(y_h - y_{h-1})W_h}{24f(\mathrm{y}_h)} \square \frac{(y_{h+1} - y_h)W_{h+1}}{24f(\mathrm{y}_h)}$$

$$\therefore (y_h - y_{h-1})W_h \square (y_{h+1} - y_h)W_{h+1} \tag{26}$$

Applying (26) finally with h=1,2,…,L-1, we find an approximation to minimal equations (26) as

$$(y_h - y_{h-1})W_h = C_L \tag{27}$$

where $\mathrm{C_L}$ is a constant depending on L. The problem of determining the approximate value of $\mathrm{C_L}$ is to be resolved so that it will help to use this to determine strata boundary points.

References

1. AYOMA H. (1954): "A study of stratified random sampling" ; Ann Inst. Stat. Math., 6, 1-136.
2. DALENIUS T. (1950): "The problem of optimum stratification I", Skand. Akt. ,33, 203-213.
3. DALENIUS T. and GURNEY. M. (1951): "The Problem of optimum stratification II" Skand. Akt, 34, 139 -148.
4. DALENIUS T. and HODGES. J. L. (1957): "The choice of stratification points" Skand. Akt., 40, 198-203.
5. DALENIUS T. and HODGES. J.L. (1959): "Minimum variance stratification" JASA, 54, 88-101.
6. DES RAJ (1964): '"On formulation of strata of equal aggregate size",JASA, 59, 481-486
7. DORFMAN A.H., VALLIANT (2000):"Stratification by size revised", Journal of Official Statistics, 16, 2, 139–154.
8. EKMAN G. (1959): "An approximation useful in univariate stratification"; Ann. Math. Stat., 30, 219-229.
9. GUPTA P. C. (1970): "On some estimation problems in sampling using auxiliary in formation", Ph.D. thesis. IASRI , New Delhi.
10. HEDLIN (2000). "A Procedure for Stratification by an Extended Ekman Rule". Journal of Official Statistics, 16, 15-29.
11. KITAGAWA.T. (1956) "Some contribution to the design of sample surveys". Sankhya, 17, 1-36.
12. LAVELLEE, P. and HIDIROGLOU, M (1988),"On the Stratification of Skewed Populations",Survey Methodology, 14, 33-43.
13. MAHALANOBIS P.C. (1952) : "Some aspect of the design of sample surveys" Sankhya, 12, 1-17
14. RAJYAGURU ARTI (1999)"Some aspects of theory of optimum stratification using coefficient of variation as norm and estimation of coefficient of variation from finite population"; Ph.D. thesis ,Surat India
15. RAJYAGURU and GUPTA (2002)"On alternative aspect of optimum stratification", Guj. Stat. Review, 29 ,101-112
16. RIVEST, L.P. (2002). A Generalization of the Lavellee-Hidiroglou Algorithm for Stratification In Business Surveys, Survey Methodology, 28, 191-198.