DEMARCATION OF STRATIFICATION POINTS USING TWO STRATIFICATION VARIABLES UNDER EQUAL ALLOCATION

Faizan Danish

Ph.D. (Statistics) Scholar, Division of Statistics and Computer Science, Faculty of Basic Sciences, Main Campus, SKUAST-Jammu, danishstat@gmail.com

Abstract: In this paper, we have tried to develop the theoretical frame for determination of optimum strata boundaries on the two auxiliary variables X and Z closely related with the study variable Y. The form of regression of estimation variable on the stratification variables as also the form of conditional variance $V\left(y/x,z\right)$ is assumed to be known. While minimizing the variance of the sample mean of the study variable, minimal equations have been obtained under equal allocation. Due to implicit nature of these equations, a $Cum\left(D\left(x,z\right)\right)$ rule has been proposed for obtaining approximately optimum strata boundaries. Empirical study has also been made to illustrate the proposed method and to make its comparison with the existing methods.

Keywords: Auxiliary Information, Equal Allocation, Minimal Equation.

Introduction: When the population of N units is to be subdivided into $L \times M$ strata and the samples from each of the stratum are selected with simple random sampling, then an unbiased estimate of population mean from the variable under study (y) is given by

$$\bar{y}_{st} = \sum_{h=1}^{L}\sum_{k=1}^{M} W_{hk}\bar{y}_{hk} \qquad (1)$$

where $W_{hk} = \frac{N_{hk}}{N}$ *and* $\bar{y}_{st}$ are the proportion and sample mean of population in the $(h,k)^{th}$ stratum ,h=1,2,...,L ; k = 1,2,...,M. If the sample selected from the $(h,k)^{th}$ stratum be n_{hk} then the total sample selected from the whole p[population is denoted by n. For stratified random sampling, the sample estimate $\bar{y}_{st}$ is unbiased and its sampling variance is given below:

$$V\left(\bar{y}_{st}\right) = \sum_{h}\sum_{k}\left(1-f_{hk}\right)\frac{W_{hk}^2\sigma_{hky}^2}{n_{hk}}$$

where $f_{hk} = \frac{n_{hk}}{N_{hk}}$ denotes the sampling fraction in the $(h,k)^{th}$ stratum. However, if the finite population correction is ignored in each stratum such that the estimate in 1 has the variance as

$$V\left(\bar{y}_{st}\right) = \sum_{h}\sum_{k}\frac{W_{hk}^2\sigma_{hky}^2}{n} \qquad (2)$$

Now for the case of equal allocation method where for all h and k , $n_{hk} = \frac{n}{LM}$,then (2) can be reduced as

$$V\left(\bar{y}_{st}\right) = LM\sum_{h}\sum_{k}\frac{W_{hk}^2\sigma_{hky}^2}{n} \qquad (3)$$

The boundaries that correspond to the minimum variance for any particular allocation method are called optimum strata boundaries (OSB).The problem of optimum stratification on the study variable was first considered by Dalenius (1950),The further work in this direction is also well-known by Singh (1971) when then information about the study variable is not sufficient then they used the information given by a variable highly correlated to the study variable. Several other authors have contributed for obtaining optimum strata boundaries like Dalenius and Gurney (1951), Singh and Sukhatme (1969), Singh (1977) ,Rizvi *et al.*(2000),Danish and Rizvi (2017).In spite of all these development Danish *et al.*(2017) discussed the various developed methods for construction of stratification points. Danish et al (2017) proposed a method under Neyman allocation for using mathematical programming approach.

In this paper we are going to develop a method for a single study variable having two auxiliary variables used as the basis of stratification under the case of equal allocation.

Minimal Equations: Let us assume the regression line of the study variable Y on the auxiliary variables X and Z be linear of the form as

$$y = c(x,z) + e \qquad (4)$$

where c(x,z) is linear or non-linear function of x and z and e is the error term such that $E\left(e/_{x,z}\right) = 0$ and $V\left(e/_{x,z}\right) = \eta(x,z) > 0$, $\forall x \in (a,b) \,\&\, z \in (c,d)$ with b-a<∞ ,d-c<∞. Let the joint density function of (X,Y,Z) in the super population is f(x, y, z) and joint marginal density function of X and Z is f(x, z) .Let f(x) and f(z) be the frequency function of the auxiliary variables X and Z ,respectively, defined in the interval [a, b] and [c, d].

If the sample mean of the study variable 'Y' is estimated under the variance given above,then the problem of determining the strata boundaries is to cut up the ranges $d_x = b - a$ and $t_z = d - c$, at (L-1) and (M-1) intermediate points as $a = x_0 \le x_1 \le \ldots \le x_{L-1} \le x_L = b$ and $c = z_0 \le z_1 \le \ldots \le z_{M-1} \le z_M = d$ respectively such that the equation (1.3) is minimum.
Then we have following relations

$W_{hk} = \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} f(x,z)\,\partial x \partial z$ is the weight of the $(h,k)^{th}$ stratum.

$\mu_{hky} = \mu_{hkc} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} c(x,z) f(x,z)\,\partial x \partial z$ denotes the mean of the $(h,k)^{th}$ stratum

and

$$\sigma^2_{hky} = \sigma^2_{hkc} + \mu_{hk\eta} \qquad (5)$$

Where $(x_{h-1}, x_h, z_{k-1}, z_k)$ are the boundaries points of $(h,k)^{th}$ strata and $\mu_{hk\eta}$ is the expected value of the function $\eta(x,z)$ in the $(h,k)^{th}$ stratum and σ^2_{hkc} is given as

$$\sigma^2_{hkc} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} c^2(x,z) f(x,z)\,\partial x \partial z - (\mu_{hkc})^2$$

Under model (3) the variance reduces to

$$(L \times M) \sum_h \sum_k W^2_{hk} \left(\sigma^2_{hkc} + \mu_{hk\eta}\right) \qquad (6)$$

to obtain minimal equations in (6),it is equal to minimize

$$\sum_h \sum_k W_{hk}^2 \sigma_{hkc}^2$$

Equating to zero the partial derivative of this expression w.r.t. x_h ,we get

$$\sum_k \left[W_{hk}^2 \frac{\partial}{\partial x_h} \sigma_{hkc}^2 + \sigma_{hkc}^2 \frac{\partial}{\partial x_h} W_{hk}^2 + W_{ik}^2 \frac{\partial}{\partial x_h} \sigma_{ikc}^2 + \sigma_{ikc}^2 \frac{\partial}{\partial x_h} W_{ik}^2 \right] = 0$$

Using the values obtained in previous equations and after simplification, we get

$$\sum_k \left[W_{hk}^2 \int_{z_{k-1}}^{z_k} \frac{1}{W_{hk}} \left\{ f(x_h, z) \left[c(x_h, z) - \mu_{hkc} \right]^2 + \sigma_{hkc}^2 + \eta(x_h, z) - \mu_{hk\eta} \right\} \partial z \right]$$

$$= \sum_k \left[W_{ik}^2 \int_{z_{k-1}}^{z_k} \frac{1}{W_{ik}} \left\{ f(x_h, z) \left[c(x_h, z) - \mu_{ikc} \right]^2 + \sigma_{ikc}^2 + \eta(x_h, z) - \mu_{ik\eta} \right\} \partial z \right]$$

Similarly, equating to zero the partial derivative of the same equation w. r. t. z_k, we get

$$\sum_h \left[W_{hk}^2 \int_{x_{h-1}}^{x_h} \frac{1}{W_{hk}} \left\{ f(x, z_k) \left[c(x, z_k) - \mu_{hkc} \right]^2 + \sigma_{hkc}^2 + \eta(x, z_k) - \mu_{hk\eta} \right\} \partial x \right]$$

$$= \sum_h \left[W_{ik}^2 \int_{x_{h-1}}^{x_h} \frac{1}{W_{ik}} \left\{ f(x, z_k) \left[c(x, z_k) - \mu_{ikc} \right]^2 + \sigma_{ikc}^2 + \eta(x, z_k) - \mu_{ik\eta} \right\} \partial x \right]$$

While partially differentiating w.r.t. x_h and z_k , the same equation and equating to zero, we get

$$W_{hk} \left\{ \left[c(x_h, z_k) - \mu_{hkc} \right]^2 + \sigma_{hkc}^2 + \eta(x_h, z_k) - \mu_{hk\eta} \right\}$$

$$= W_{ij} \left\{ \left[c(x_h, z_k) - \mu_{ijc} \right]^2 + \sigma_{ijc}^2 + \eta(x_h, z_k) - \mu_{ij\eta} \right\} \qquad (7)$$

where i=h+1,j=k+1,h=1,2,...,L and k=1,2,...,M

These equations are implicit functions of the strata boundaries $[x_h, z_k]$ and their exact solutions are somewhat difficult to find. We therefore proceed to find the method of solving them at least approximately.

Minimal Equations and Their Approximate Solutions: To find approximate solutions to the minimal equations (7) we shall obtain series of expansions of the system of equations about the point (x_h, z_k),the common boundary points of $(h,k)^{th}$ and $(h+1,k+1)^{th}$ strata. These expansions for W_{hk}, μ_{hkc} and σ_{hkc}^2 about both the lower and upper boundaries of the $(h,k)^{th}$ stratum, as given by Singh and Sukhatme (1969) while doing this we shall assume the existence of all the functions and their derivatives occurring in the expansions for all x in (a,b) and z in (c,d).Thus we have for the different terms in left hand side of minimal equations (7)

$$W_{hk} = \left[f(k_h k_k) + \frac{1}{2}(f_x + f_z)(k_h k_k)^2 + \frac{1}{6}(f_{xx} + f_{zz} + f_{xz})(k_h k_k)^3 + O(k_h k_k)^4 \right]$$

$$\left[\mu_{hkc} - c(x_h, z_k) \right]^2 = \frac{(k_h k_k)^2}{4} \left[c'^2 + \frac{c'^2 (f_x + f_z) + 2 f c' c''}{3f} (k_h k_k) + O(k_h k_k)^2 \right]$$

Also

$$\eta(x_h,z_k)+\mu_{hk\eta}=\eta\left[\begin{array}{l}2+\frac{\eta'}{2\eta}(k_hk_k)-\frac{\eta'(f_x+f_z)+2f\eta''}{12f\eta}(k_hk_k)^2\\+\frac{f(f_{xx}+f_{zz}+f_{xz})\eta'+f(f_x+f_z)\eta''+f^2\eta'''-(f_x+f_z)^2\eta'}{24f^2\eta}(k_hk_k)^3+O(k_hk_k)^4\end{array}\right]$$

and $\sigma^2_{hkc}=\frac{(k_hk_k)^2}{12}\left[c'^2+c'c''(k_hk_k)+O(k_hk_k)^2\right]$

where the functions c,f,η and their derivatives are evaluated at x_h and z_k and $k_h=x_h-x_{h-1}$ $k_k=z_k-z_{k-1}$.Hence on simplifuication we get left hand side of (7) as

$$W_{hk}\left\{\left[c(x_h,z_k)-\mu_{hkc}\right]^2+\sigma^2_{hkc}+\eta(x_h,z_k)-\mu_{hk\eta}\right\}$$
$$=2\sqrt{\eta}\left[f\sqrt{\eta}(k_hk_k)-\frac{\eta'f+2f(f_x+f_z)}{4\sqrt{\eta}}(k_hk_k)^2+O(k_hk_k)^3\right] \quad (8)$$

In the similar way we can get the expansion of right hand side of (7) as

$$W_{ij}\left\{\left[c(x_h,z_k)-\mu_{ijc}\right]^2+\sigma^2_{ijc}+\eta(x_h,z_k)-\mu_{ij\eta}\right\}$$
$$=2\sqrt{\eta}\left[f\sqrt{\eta}(k_ik_j)+\frac{\eta'f+2f(f_x+f_z)}{4\sqrt{\eta}}(k_ik_j)^2+O(k_ik_j)^3\right] \quad (9)$$

where $k_i=x_{h+1}-x_h$ and $k_k=z_{k+1}-z_k$

Thus from 8 and 9, the system of minimal equations 4 can be written as

$$(k_hk_k)\left[f\sqrt{\eta}-\frac{\eta'f+2f(f_x+f_z)}{4\sqrt{\eta}}(k_hk_k)+O(k_hk_k)^2\right]$$
$$=(k_ik_j)\left[f\sqrt{\eta}+\frac{\eta'f+2f(f_x+f_z)}{4\sqrt{\eta}}(k_ik_j)+O(k_ik_j)^2\right] \quad (10)$$

which can be put as

$$\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k}\sqrt{\eta}f(t_1,t_2)\partial t_1\partial t_2\left[1+O(k_hk_k)^2\right]$$
$$=\int_{x_h}^{x_{h+1}}\int_{z_k}^{z_{k+1}}\sqrt{\eta}f(t_1,t_2)\partial t_1\partial t_2\left[1+O(k_hk_k)^2\right] \quad (11)$$

where i=h+1,h=1,2,...,L and mj=k+1.k=1,2,...,M

If we have significantly large number of strata widths k_h and k_k are small and their higher powers in the expansion can be neglected, then the system of equations (11) approximated by

$\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k}\sqrt{\eta(t_1,t_2)}f(t_1,t_2)\partial t_1\partial t_2$ = Constant

where terns of order $O(m)^3$, $m = \sup(k_h k_k)$ have been neglected on both sides of the equation (11) since $\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k}\sqrt{\eta(t_1,t_2)}f(t_1,t_2)\partial t_1\partial t_2 = O(m)$ when the function $\sqrt{\eta(x,z)}f(x,z)$ is bounded for all in (a,b) and z in (c,d).Thus, we get then following rule for finding OSB for equal allocation.

Cum $D(x,z)$ **Rule:** If the function $D(x,z)=\sqrt{\eta(x,z)}f(x,z)$ is bounded and its first derivative exists for all x in (a,b) and z in (c,d),then for a given value of L and M taking intervals on then cumulative cube root of $D(x,z)$ will give AOSB (x_h, z_k).

Remarks: Let $c(x,z)=\alpha+\beta x+\gamma z$,then by differentiating w. r. t. x and z ,we get $\eta(x,z)$ is constant. Therefore, for such a case the proposed rule reduces to Cum $\sqrt{f(x,z)}$.

Empirical Study: We shall now demonstrate empirically the effectiveness of the proposed method of findings the set of AOSB. For the sake of simplicity, the linear regression line Y on X and Z have been taken as, of the form $y=\alpha+\beta x+\gamma z+e$. For the conditional variance function $\eta(x,z)$ we have taken two forms viz. $\eta(x,z)=\alpha$ and $\eta(x,z)=\lambda xz$ where $\alpha \quad and \quad \lambda$ are constants.

The origin is deliberately excluded from the range of the auxiliary variables X and Z otherwise $\eta(x,z)=\lambda xz$ we have $m_1(x,z)=\infty$ at x=0,z=0 and the function $m_1(x,z)f(x,z)$ in that case doesn't belong to the class Ω of functions. We could have also avoided this difficulty by taking some other suitable forms to the functions. For the empirical studies under optimum allocation let us assume values of $\alpha=0.0214, \lambda=0.00437$ which are quite small so that the effect of taking $\eta(x,z)=\alpha$ and $\eta(x,z)=\lambda xz$ is negligibly small.

In order to obtain AOSB let us assume that the correlation coefficient between X and Z is denoted by ρ and is equal to 0.65 For this purpose the following density functions of the stratification variables X and Z have been considered.

Now let us assume that X follows standard normal distribution with probability density function (pdf) as

$$f(x)=\frac{1}{2\pi}e^{-\frac{x^2}{2}} \qquad ,x\geq 0$$

and Z follows Exponential distribution with pdf as

$$f(z)=e^{-z+1}, z\geq 1$$

In order to obtain the OSB when both the variables are standard normally distributed let us assume the value of regression coefficients $\beta=0.42$ and $\gamma=0.57$ .For obtaining total 16 strata, 4 along the x-axis and 3 along the z-axis using the proposed rule Cum $D(x,z)$ by solving it in Mathematica Software assuming the distribution of X and Z is truncated at x=6 and z=4 respectively, we get the stratification points as below:

Table 5.1: OSB When the Auxiliary Variables X and Z Are Having Standard Normal Distribution and Exponential Distribution For $\eta(x,z)=\alpha$

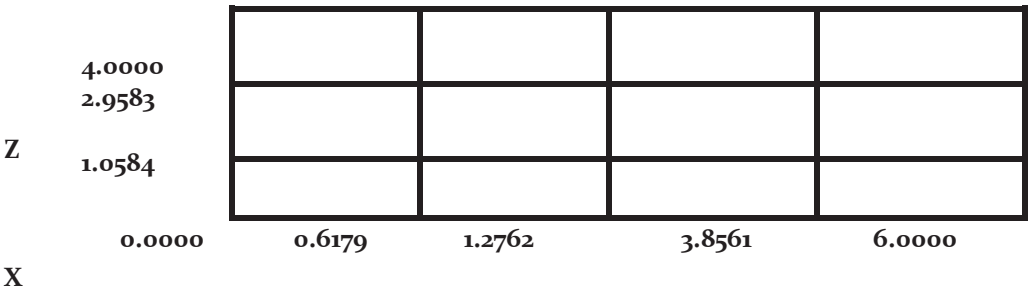

Table 5.2: OSB and Variance When the Auxiliary Variables are Having Standard Normal Distribution and Exponential Distribution for $\eta(x,z)=\alpha$

OSB (x_h, z_k)	**Variance** (Cum $D(x,z)$ Rule)	**Variance (Singh 1977)**	**% R.E.**
(0.6179,1.0584) (1.2769,1.0584) (3.8561,1.0584) (6.0000,1.0584) (0.6179,2.9583) (1.2769,2.9583) (3.8561,2.9583) (6.0000,2.9583) (0.6179,4.0000) (1.2769,4.0000) (3.8561,4.0000) (6.0000,1.0584)	0.03724698	0.08431769	226.37456

Table 5.3: OSB and Variance When the Auxiliary Variables are Having Standard Normal Distribution and Exponential Distribution for $\eta(x,z)=\lambda xz$

OSB (x_h, z_k)	**Variance** (Cum $D(x,z)$ Rule)	**Variance (Singh 1977)**	**% R.E.**
(0.5924,1.1385) (1.3591,1.1385) (3.7318,1.1385) (6.0000,1.1385) (0.5924,3.0579) (1.3591,3.0579) (3.7318,3.0579) (6.0000,3.0579) (0.5924,4.0000) (1.3591, 4.0000) (3.7318, 4.0000) (6.0000, 4.0000)	0.052841973	0.079247631	149.97099

Table 5.1 shows the stratification points when the auxiliary variables are standard normally and exponentially distributed to make 12 strata of which 4 along the x-axis and 3 along the z-axis. Furthermore Table 5.1. and 5.2 presents both OSB and variance obtained by proposed method and Singh (1977),which shows the very little effect on the boundaries by taking different form oof conditional

variance. However from both the tables it can be concluded that the proposed method is more preferable than of Singh (1977).

Conclusion: The problem of optimum stratification on the basis of auxiliary variables when the units from different strata are selected with simple random sampling was considered by several authors for the univarite case. In this investigation we have extended the same using two auxiliary variables having single study variable. A Cum D (x,z) rule for obtaining the stratification points has been proposed. It has been shown empirically that using the two stratification variables is more preferable rather than single stratification variable, A comparison of proposed method with Singh (1977) gives us the percentage of relative efficiency more than 100 that proves the superiority of the proposed method.

References:

1. Dalenius, T. The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*.(1950): 33 : 203-213.
2. Dalenius, T. and Gurney, M. The problem of optimum stratification II.*Skandinavisk Aktuarietidskrift* , (1951) :34: 133-148.
3. Danish, F. and Rizvi, S.E.H. On Optimum Stratification Using Mathematical Programming Approach. *International Research journal of Agricultural Economics and Statistics*.(2017): 8(2):435-439.
4. Danish, F. and Rizvi, S.E.H. (2017).Comparison of bi-variate versus univariate stratifying variables. Mathematical sciences international research journal.6:80-88.
5. Rizvi, S.E.H., Gupta, J.P. and Singh, R. 2000. Approximately optimum stratification for two study variables using auxiliary information. *Journal of the Indian Society of the Agricultural Statistics*.(2000): 53 (3): 287-298.
6. Singh, R. 1977. A note on equal allocation with ratio and regression methods of estimation. *Australian Journal of Statistics*.(1977): 19: 96-104.
7. Singh, R. 1971. Approximately optimum stratification on auxiliary variable. *Journal of American Statistical Association*.(1971): 66: 829-833.
8. Singh, R. and Sukhatme, B.V. 1969. Optimum stratification. *Annals of Institute of Statistical Mathematrica*.(1969):21: 515-528.

***