DETERMINATION OF OPTIMUM STRATA BOUNDARIES

RAVINDRA SINGH

Punjab Agricultural University: Ludhiana

Introduction

We shall consider the problem of estimating the population mean for the study variable y, using the stratified simple-random sampling assuming the variable y to be continuous and the population infinite. It is easy to see that an unbiased estimate of the population mean is given by

$$\bar{y}=\sum_{h=1}^{L} W_h \bar{y}_h, \tag{1.1}$$

where W_h is the proportion of the population in hth ($h=1, 2, \ldots L$) stratum and $\bar{y}_h$ is the sample mean based on n_h units drawn from that stratum so that

$$\sum_{h=1}^{L} n_h=n,$$

the total sample size.

The variance of the estimate $\bar{y}$ is given by

$$V(\bar{y})=\sum_{h=1}^{L} W_h^2 \sigma_h^2/n_h, \tag{1.2}$$

σ_h^2 being the population Variance for y in the hth stratum. Clearly the variance in (1.2) depends on

(*i*) the number of strata, L,

(*ii*) the strata boundaries, and

(*iii*) the method of allocating the sample to different strata.

In this paper we shall consider the minimisation of this variance for the optimum allocation method (taking the cost of observing y on any population unit, irrespective of its size, as same) and obtain the optimum strata boundaries corresponding to this allocation method. The variance in (1.2) with optimum allocation of the sample to different strata reduces to

$$V_o(\bar{y}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2 \qquad (1.3)$$

The problem of determining the optimum strata boundaries (OSB) on the study variable y by minimising the variance given in (1.3) was first considered by Dalenius (1950). He obtained a set of $(L\text{-}1)$ equations, called minimal equations, solutions to which gave the OSB for the L strata. Since the equations themselves involved the various strata parameters, they could not be solved exactly. With no exact solutions available, several persons tried to solve them at least approximately. Among all these workers, specially Dalenius and Gurney (1951), Mahalanobis (1952), Hansen, Hurtwitz and Madow (1953), Aoyama (1954), Dalenius and Hodges (1957), Ekman (1959) and Durbin (1959) are to be noted.

All the work referred to in the preceding paragraph was related to optimum stratification on the study variable y. But in practice we never have the required information on the study variable and the information is only available for some highly correlated auxiliary variable x. It is thus desirable to develop the theory for the optimum stratification on the auxiliary variable. Assuming the regression of y on x as linear, Dalenius (1957) gave minimal equations, solutions to which gave OSB on the auxiliary variable x. Neither exact nor approximate solutions to these equations could be obtained until (1969) when Singh and Sukhatme considered the problem in even more general form and gave various methods of finding approximate solutions to the minimal equations. As in case of stratification on y, the exactly OSB are not possible in this case.

We consider the two cases of stratification on the study variable and on the auxiliary variables in sections 2 and 5 respectively. In section 3 are given different methods of approximately solving the Dalenius (1950) equation while in section 4 an investigation into the accuracy of these methods have been given.

2. Stratification on the study variable

If $f(y)$ denotes the density for y and $[y_h]$ are the strata boundaries, then we have

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

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

and

$$\sigma_h^2=\left[\int_{y_{h-1}}^{y_h} y^2 f(y)dy/W_h\right]-\mu_h^2.$$

As mentioned earlier, to obtain the OSB the variance expression given in (1.3) is to be minimised. Minimisation of $V_o(\bar{y})$ is equivalent to the minimisation of

$$\sum_{h=1}^{L} W_h\sigma_k.$$

Thus on differentiating $\Sigma W_h\sigma_h$ partially with respect to y_h, the upper boundary of the hth stratum, we get the minimial equations as

$$\frac{\partial}{\partial y_h}\left(W_h\sigma_h\right)+\frac{\partial}{\partial y_h}\left(W_i\sigma_i\right)=O, \qquad (2.2)$$

where $i=h+1$ and $h=1, 2, \ldots, L-1$. The partial derivatives with respect to other parameters are zero because they do not depend on y_h.

It is easy to verify that

$$\frac{\partial}{\partial y_h}\left(W_h\,\sigma_h\right)=f(y_h)\left[\left(y_h-\mu_h\right)^2+\sigma_h^2\right]\Big/2\sigma_h,$$

and

$$\frac{\partial}{\partial y_h}\left(W_i\sigma_i\right)=-f(y_h)\left[\left(y_h-\mu_i\right)^2+\sigma_i^2\right]\Big/2\sigma_i$$

Using these expressions in (2.2), the minimal equations are obtained as

$$\frac{(y_h-\mu_h)^2+\sigma_h^2}{\sigma_h}=\frac{(y_h-\mu_i)^2+\sigma_i^2}{\sigma_i}; \qquad (2.3)$$

$$i=h+1,\ h=1, 2, \ldots, L-1.$$

These equations are due to Dalenius (1950). Since the equations involve the strata parameters which in turn are functions of the strata boundaries [*i.e.*, the solutions of the equations (2.3)], the exact solutions of these equations cannot be obtained. It, therefore, becomes essential to search for the approximate solutions of these equations which give approximately optimum strata boundaries (AOSB) on the study variable y.

3. The AOSB

The names of various authors who have proposed methods of finding the approximate solutions to the minimal equations (2.3) were given in section 1. The methods suggested by them are given below. It may be noted that all these rules, except the first, basically need the knowledge of the density $f(y)$ of the variable y.

3.1. Equalization of $(y_h - y_{h-1})$: This method was suggested by Aoyama (1954). According to this rule AOSB are obtained by taking equal interval on the range of the variable y, the number of intervals being equal to the number of strata desired. As this rule does not depend on the density of the variable y, much accuracy cannot be expected of it. Another drawback of the rule is that it can only be applied to variables having finite range. This method, however, has certain other desirable properties. First the knowledge of density function $f(y)$ is not required for determining AOSB, only the range of y should be known. Secondly this method gives AOSB in variety of other cases also *e.g.* *(i)* when proportional allocation is used in place of Neyman allocation, *(ii)* in case of simple random sampling, equal intervals on the range of the auxiliary variable x gives AOSB when stratification is to be done on the scale of x, *(iii)* when sampling schemes like probability proportional to size (with replacement) and estimates like ratio, regression and product are used to estimate the population total and the stratification is on the variable x. (See Singh, 1967).

3.2. Equalization of $W_h\sigma_h$: Dalenius and Gurney (1951) conjectured that the formation of strata on the basis of equalization of $W_h\sigma_h$ and giving equal allocation to the strata would lead to optimum stratification. The conjecture was later proved by Dalenius and Hodges (1957). This method is not simple as it erquires the calculation of σ_h values for different sets or stratification points.

3.3. Equalization of strata totals :

Mahalanobis (1952) and Hansen, Hurtwitz and Madow (1953) suggested that, for a given number of strata, a practical method of possible stratification is to stratify the whole population such that the strata sums are equal. The method was not supported by any theoretical justification and later Kitagawa (1956) showed that this method gave optimum stratification only for the populations in which the strata coefficients of variation were same for all possible stratifications

3.4. Equalization of cumulatives of $\sqrt{f(y)}$

Dalenius and Hodges (1957) proposed formation of strata by equalizing the cumulative of $\sqrt{f(y)}$. In obtaining this rule, it has been assumed that the distribution of y is bounded and the number of strata is large. The first assumption is generally valid in practice. Thus it is of interest to see how the rule works for small values of L. It has been shown that for some continuous distributions the rule gives a close approximation to the optimum strata boundaries even for L=2 or 3 (Dalenius and Hodges, 1959). The observation was later supported by Cochran (1961) and Sethi (1963). Serfling (1968) has advocated its use even for the optimum stratification on the auxiliary variable x when the correlation between y and x is nearly perfect.

3.5. Equalization of $W_h(y_h - y_{h-1})$

Under certain regularity conditions on the density function $f(y)$ and for a finite range of the variable y, it was shown by Ekman (1959) that the points $[y_h]$ satisfying the equalities

$$W_h(y_h - y_{h-1}) = \text{Const.} \qquad \text{for } h = 1,2,\ldots\ldots, L.$$

approximately satisfy the minimal equations (2·3). Certain modifications to the above equalities were recommended for cases where the variable y has an infinite range. With the strata so formed, the allocation of the sample to different strata has to be equal. Numerical investigations show that the method works satisfactorily even for values $L=2$ or 3 (Cochran 1961).

3.6. Equalization of cumulative of $[r(y) + f(y)]/2$

Durbin (1959) proposed the equalization of the cumulative frequencies of a distribution $g(y)$, which is mid way between the original distribution $f(y)$ and the rectangular distribution $r(y)$ over the range (y_0, y_L). Thus $r(y)$ is taken as $F(y_L)/(y_L - y_0)$ and the *AOSB* are obtained by taking equal intervals on the cumulative of the function

$$g(y) = \tfrac{1}{2}\,(r(y) + f(y)).$$

4. Relative Efficiency of Various Methods

Various authors [Ref. Cochran (1961), Des Raj (1964), Hess, Sethi and Balkrishna (1966) and Sethumadhavi (1966)] have made empirical studies about the relative efficiency of the different stratification procedures mentioned in the preceding section. The rules (3.2), (3.4) and (3.5) are usually found to be more efficient than the others. In this section the above observation is being supported theoretically.

Using the series expansions (Singh and Sukhatme, 1969) for w_h, μ_h and σ_h about the upper and lower bounaries of the hth stratum, the system of minimal equations (2·3) is found to be equivalent to

$$K_h\left[1 - \frac{f'}{4f}\cdot K_h + \frac{7ff'' - 5f'^2}{60f^2}K_h^2 + O(K_h^3)\right] \qquad (4.1)$$
$$= K_i\left[1 + \frac{f'}{4f}K_i + \frac{7ff'' - 5f'^2}{60f^2}K_i^2 + O(K_i^3)\right]$$

where $i = h+1$, $K_h = y_h - y_{h-1}$ and the function f and its derivatives are evaluated at $y = y_h$.

In order to investigate into the relative efficiency of various stratification procedures, we shall similarly find their series expansions and compare with (4·1). It is easily seen that the expansions corresponding to the rules (3.1) to (3.6) respectively are as given below.

$$K_h = K_i, \tag{4.2}$$

$$K_h\left[1-\frac{f'}{4f}\cdot K_h+\frac{44ff''-25f'^2}{480f^2}\cdot K_h^2+O(K_h^3)\right] \tag{4.3}$$

$$=K_i\left[1+\frac{f'}{4f}K_i+\frac{44ff''-25f'^2}{480f^2}K_i^2+O(K_i^3)\right]$$

$$K_h\left[1-\frac{f+f'y_h}{2fy_h}K_h+\frac{f''+2f'y_h}{6fy_h^2}K_h^2+O(K_h^3)\right] \tag{4.4}$$

$$=K_i\left[1+\frac{f+f'y_h}{2fy_h}K_i+\frac{f''+2f'y_h}{6fy_h^2}K_i^2+O(K_i^3)\right]$$

$$K_h\left[1-\frac{f'}{4f}K_h+\frac{2ff''-f'^2}{24f^2}K_h^2+O(K_h^3)\right] \tag{4.5}$$

$$=K_i\left[1+\frac{f'}{4f}K_i+\frac{2ff''-f'^2}{24f^2}K_i^2+O(K_i^3)\right]$$

$$K_h\left[1-\frac{f'}{4f}K_h+\frac{8ff''-3f'^2}{96f^2}K_h^2+O(K_h^3)\right] \tag{4.6}$$

$$=K_i\left[1+\frac{f'}{4f}K_i+\frac{8ff'-3f'^2}{96f^2}K_i^2+O(K_i^3)\right]$$

$$K_h\left[1-\frac{f'}{2f}K_h+\frac{f''}{6f}K_h^2+O(K_h^3)\right] \tag{4.7}$$

$$=K_i\left[1+\frac{f'}{2f}K_i+\frac{f''}{6f}K_i^2+O(K_i^3)\right]$$

Since $(K_h-K_i)/(y_L-y_0)=0(K_h^3)$

Comparing the expansions for different rules with (4·1) we find that the expansions for corresponding to rules (3.1), (3.3) and (3.6) agree in terms of $0(K_h)$ while the expansions for the rules (3.2), (3.4) and (3.5) agree upto the terms $0(K_h^2)$ and differ in terms of order $0(K_h^3)$. The rules (3.2), (3.4) and (3.5) should, therefore, be more accurate in giving the AOSB as compared to the rules (3.1), (3.3) and (3.6). This is also the observation made from different empirical studies.

5. Stratification on the Auxiliary Variable

The problem of determining the OSB on the auxiliary variable x has been considered by Singh and Sukhatme (1969). Assuming the regression of the study variable y on the variable x to be of the form

$$y=c(x)+e, \tag{5.1}$$

where $c(x)$ is some function of x and e is the error term such that $E(e \mid x)=0$ and $V(e \mid x)=\varphi(x)>0$ for all x in the range (a, b) of x with $(b-a)<\infty$. The functions $c(x)$ and $\varphi(x)$ are assumed to be known.

Under this set up the minimal equations, solutions to which give the OSB on the scale of x, are found to be

$$\frac{\left(c(x_h)-\mu_{hc}\right)^2+\sigma_{hc}^2+\varphi(x_h)+\mu_{h\varphi}}{\sqrt{\sigma^2_{hc}+\mu_{h\varphi}}} \tag{5.2}$$

$$=\frac{\left(c(x_h)-\mu_{ic}\right)^2+\sigma_{ic}^2+\varphi(x_h)+\mu_{i\varphi}}{\sqrt{\sigma ic^2+\mu_{i\varphi}}}$$

$$i=h+1,\ h=1,\ 2,\ \ldots\ldots\ldots,\ L-1.$$

where μ_{hc}, $\mu_{h\varphi}$ and σ_{hc}^2 are respectively the expected values of $c(x)$, $\varphi(x)$ and the variance of $c(x)$ in *hth* stratum.

If

$$g(t)=\frac{\varphi'^2(t)+4\varphi(t)c'^2(t)}{\left(\varphi(t)\right)^{\frac{3}{2}}}, \tag{5.3}$$

then under certain regularity conditions it has been shown that the approximate solutions to the minimal equations (5.2) are given by the solutions of the system of equations

$$K_h^2\int_{x_{h-1}}^{x_h} g(t)\,f(t)dt=\text{const.},\ h=1,\ 2,\ldots,\ L. \tag{5.4}$$

where $K_h=x_h-x_{h-1}$ and $f(x)$ is now the density function of x. In deriving the system of equations (5.4) the terms of order $0(m^4)$, $m=\sup_{(a,\,b)}(K_h)$, have been neglected. It has also been shown that if we have a function $Q(x_{h-1}, x_h)$ such that

$$Q(x_{h-1}, x_h)=K_h^2\int_{x_{h-1}}^{x_h} g(t)\,f(t)\,dt.\left[1+0(K_h^2)\right], \tag{5.5}$$

+.....the approximate solutions to the minimal equations (5.2) can, with the same degree of accuracy as involved in (5.4), also be obtained by solving the system of equations

$$Q(x_{h-1}, x_h)=\text{Const.},\ h=1,\ 2,\ldots,\ L. \tag{5.6}$$

By making use of this property various methods of finding AOSB on the scale of x have been proposed. One of these methods is called cum. $\sqrt[3]{g(x)f(x)}$

rule. According to this rule AOSB are obtained by taking equal intervals on the cumulative of $\sqrt[3]{g(x)f(x)}$. Numerical investigation shows that the *AOSB* so obtained give quite close approximation to the OSB.

6. Some further remarks

The problem of minimising the variance of the stratified simple random sampling estimate of the population mean for the fixed total cost has been considered by Ekman (1959, 1960). The cost of observing a unit is taken to be a function of the y value for that unit. Assuming the knowledge of $f(y)$, the approximate solutions to the minimal equations were obtained under certain regularity conditions. Four methods of finding AOSB have been given.

(*ii*) An alternative approach to tackle the problem of optimum stratification was tried by Sethi (1963). In place of finding approximations to OSB by solving some suitably chosen system of equations (other than the minimal equations), he thought of preparing ready made tables for giving stratification points for certain standard frequency distributions. In practice the actual distribution is to be replaced by some suitable standard distribution and the OSB corresponding to it are to be used.

(*iii*) So far we have discussed the problem of finding OSB for the case of stratified simple random sampling. The various methods mentioned in the preceding sections fail to give AOSB if one decides to change either the selection procedure within each stratum or the method of estimation. Singh (1967) has considered the problem of finding OSB when either the PPSWR scheme is used for the selection of the sample within each stratum or the ratio, regression or product methods of estimation are used. Various methods of finding AOSB have been given and their asymptotic properties discussed.

7. Summary

The determination of optimum strata boundaries for the optimum allocation method has been discussed in great detail.

REFERENCES

Aoyama, H. (1954) : 'A study of stratified random sampling'. Ann. Inst. Stat. Math , 6, 1-36.

Cochran, W. G. (1961) : 'Comparison of methods of determining strata boundaries'. Bull. Int. Stat. Inst. 38, 345-358.

Dalenius, T (1950) : 'The problem of optimum stratification' Skand. Akt., 33, 203-13.

Dalenius, T. and Gurney, M. (1951) : 'The problem of optimum stratification II' Skand. Akt., 34, 133-148.

Dalenius, T. (1957) : 'Sampling in Sweden' Almqvist and Wicksell, Stockholm.

Dalenius, T. and Hodges, J. L. (1957) : 'The choice of stratification points' Skand. Akt., 40, 198-203.

.....(1959) : 'Minimum variance stratification' Jour Amer. Stat. Assoc., 54, 88-101.

Des Raj (1964) : 'On forming the strata of equal aggregate size' Jour. Amer Stat. Assoc., 59, 481-486.

Durbin, J. (1959) : 'Review of 'Sampling in Sweden' Jour. Roy Sta'. Soc. (A), 246-248.

Ekman, G. (1959) : 'An approximation useful in Univariate stratification'. Ann. Math. Stat. 30, 219-229.

.....(1959) ; 'A limit theorem in connection with stratified sampling. Part I' Skand. Akt. 42, 208-23.

.....(1960) : 'A limit theorem in connection with stratified sampling. part II' Skand. Akt. 43, 1-16.

Hansen, M. H , Hurwitz. W. N. and Madow, W. G. (1953) : 'Sample survey methods and theory'. Vol. I and II. John Wiley and Sons, New York.

Hess, I, Sethi, V. K. and Balkrishna, T. R. (1966) : 'Stratification : A practical investigation' Jour. Amer. Stat. Assoc. 61, 74-90.

Kitagawa, T. (1956) : 'Some contributions to the design of sample surveys' Sankhya, 17, 1-36.

Mahalanobis, P. C. (1952) : 'Some aspects of the design of sample surveys.' Sankhya. 12, 1-7.

Serfling, R.J. (1968) : Approximately optimum stratification Jour. Amer. Stat. Assoc., 63, 1298-1309.

Sethi, V. K. (1963) : A note on optimum stratification of populations for estimating the population means'. Aust. Jour. Stat., 5, 20-33.

Sethumadhavi, R. (1966) : 'Stratification in surveys on fruit crops' Unpublished M. Sc. thesis submitted to I.A.R.I., New Delhi.

Singh. R. (1967) : 'Some contributions to the theory of construction of strata' Unpublished Ph.D. thesis submitted to I.A.R.I., New Delhi.

Singh, R. and Sukhatme, B. V. (1969) : "Optimum stratification' Ann. Inst. Stat. Math. 21, 515-528.