J. Ind. Soc. Agril. Statist.
59(2), 2005 : 146–150

Optimum Stratification for Exponential Study Variable under Neyman Allocation

M.G.M. Khan, Najmussehar[1] and M.J. Ahsan[1]
Department of Mathematics and Computing Science
The University of the South Pacific, Suva, Fiji
(Received : May, 2005)

SUMMARY

For stratified sampling to be efficient the strata should be as homogeneous as possible with respect to the main study variable. In other words, the stratum boundaries are so chosen that the stratum variances are as small as possible. This could be done effectively when the frequency distribution of the main study variable is known. Usually this frequency distribution is unknown but it is possible to approximate it from the past experience and prior knowledge about the population. In the present paper the problem of optimum stratification is studied and formulated as a Mathematical Programming Problem (MPP) assuming exponential frequency distribution of the main study variable. The stratum boundaries are optimum in the sense that they minimize the sampling variance of the stratified sample mean under Neyman allocation. The formulated MPP is separable with respect to the decision variables and is treated as a multistage decision problem. A solution procedure is developed using dynamic programming technique. A numerical example is also given to show the computational efficiency of the procedure.

Key words : Optimum stratification, Exponential study variable, Mathematical programming problem, Dynamic programming.

1. INTRODUCTION

The use of stratified sampling involves the solution of four carefully formulated optimization problems according to the objective and the available resources to the sample survey. These four optimization problems are related to the optimum choice of the

(i) stratification variable
(ii) number of strata
(iii) stratum boundaries, and
(iv) sample size allocation

The basic consideration involved in the formation of strata is that they should be internally as homogenous as possible, that is the stratum variances S_h^2 should be as small as possible. If the distribution of the study variable is available, the strata should be created by cutting this distribution at suitable points.

The problem of determining the optimum strata boundaries (OSB), when the study variable itself is used as stratification variable, was first discussed by Dalenius (1950) who obtained the minimal equations that gave the OSB as their solutions. Unfortunately the exact solutions of these equations are not possible because of their implicit nature. Several attempts have been made to find out approximate solutions. Some of these are due to Dalenius and Gruney (1951), Mahalanobis (1952), Aoyama (1954), Dalenius and Hodge (1959), Ekman (1959), Sethi (1963), Serfling (1968) and Singh (1975b). Cochran (1961) and Hess *et al.* (1966) had made empirical comparisons of these methods.

Usually the frequency distribution of the study variable is unknown. Thus using the study variable as stratification variable is not practically feasible. Taga (1967), Singh and Sukhatme (1969, 1972, 1973), Singh and Prakash (1975) and Singh (1975a, 1975c,

1. Aligarh Muslim University, Aligarh

1977) and many others used an auxiliary stratification variable to determine the OSB under different allocations.

Unnithan (1978) used the modified Newton-Raphson method for determining the OSB that leads only to a local minimum of the objective function. Later on Unnithan and Nair (1995) gave a method of selecting an appropriate starting point for modified Newton-Raphson method that may lead to a global minimum of the objective function.

Yadava and Singh (1984) obtained the approximate OSB for allocations proportional to strata totals. Mehta *et al.* (1996) extended the above work for the use of ratio, regression and product estimators of the population mean.

Buhler and Deutler (1975) formulated the problem of determining the OSB as an optimization problem that can be solved by dynamic programming technique. Khan *et al.* (2002) formulated the problem as an MPP when the number of strata is fixed in advance. By suitable transformation they converted the problem into a multistage decision problem in which at each stage the value of a single decision variable is worked out using dynamic programming technique.

The present paper is a sequel to that of Khan *et al.* (2002). Here we consider the problem of determining the OSB for an exponential study variable under Neyman allocation.

2. FORMULATION OF THE PROBLEM

Let the population under study be stratified into L strata and the estimation of the population mean is of interest. Let x_0 and x_L be the smallest and largest values of the study variable X in the population. Then the problem of optimum stratification can be described as to find the intermediate stratum boundaries $x_1 \le x_2 \le \ldots \le x_{L-1}$ such that the variance of the stratified sample mean $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ under Neyman allocation, that is

$$V(\bar{x}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{N} \qquad (2.1)$$

is minimum, where $\bar{x}_h$ is the sample mean based on n_h units drawn from the h^{th} stratum and W_h is the proportion of population units falling in that stratum.

If the finite population correction is ignored, minimizing the expression of the right hand side of (2.1) is equivalent to minimize

$$\sum_{h=1}^{L} W_h \sigma_h \qquad (2.2)$$

Let f(x) be the frequency function of study variable X. The problem of determining the OSB is then equivalent to finding the L-1 intermediate points $x_1 \le x_2 \le \ldots \le x_{L-1}$ in the interval $[x_0, x_L]$ such that (2.2) is minimum.

Let

$$x_L - x_0 = d \qquad (2.3)$$

The values of W_h and σ_h in (2.2) are obtained by

$$W_h = \int_{x_{h-1}}^{x_h} f(x)\,dx \qquad (2.4)$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)\,dx - \mu_h^2 \qquad (2.5)$$

$$\text{where } \mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x)\,dx \qquad (2.6)$$

and (x_{h-1}, x_h) are the boundaries of h^{th} stratum.

When the frequency function f (x) is known, using (2.4), (2.5) and (2.6), $W_h \sigma_h$ in (2.2) could be expressed as a function of x_h and x_{h-1} only.

Let $f_h(x_h, x_{h-1}) = W_h \sigma_h$

Then the problem of determining the OSB can be expressed as

"Find $x_1, x_2, \ldots, x_{L-1}$ which minimize $\sum_{h=1}^{L} f_h(x_h, x_{h-1})$, subject to the constraint (2.3)".

Define

$$y_h = x_h - x_{h-1};\ h = 1,2,\ldots,L$$

where $y_h \ge 0$ denotes the width of the h^{th} stratum.

With the above definition of y_h, (2.3) is expressed as

$$\sum_{h=1}^{L} y_h = \sum_{h=1}^{L} (x_h - x_{h-1}) = x_L - x_0 = d$$

The k^{th} stratification piont x_k; k=1, 2, . . ., L-1 is then expressed as $x_k = x_0 + y_1 + y_2 + \ldots + y_k$.

Then the problem of determining OSB can now be considered as the problem of determining optimum strata widths as the following Mathematical Programming Problem (MPP).

$$\text{Minimize} \quad \sum_{h=1}^{L} f_h(y_h, x_{n-1})$$

$$\text{subject to} \quad \sum_{h=1}^{L} y_h = d \qquad (2.7)$$

and

$$y_h \geq 0; h = 1,2,\ldots,L$$

For h = 1 the term $f_1(y_1, x_0)$ in the objective function of (2.7) is a function of y_1 alone, as x_0 is known. Similarly, for h = 2 the term $f_2(y_2, x_1) = f_2(y_2, x_0 + y_1)$ will become a function of y_2 alone once y_1 is known. Thus, stating the objective function as a function of y_h alone we may rewrite the MPP (2.7) as

$$\text{Minimize} \quad \sum_{h=1}^{L} f_h(y_h)$$

$$\text{subject to} \quad \sum_{h=1}^{L} y_h = d \qquad (2.8)$$

and $y_h \geq 0; h = 1,2,\ldots,L$

Let the stratification variable X follows the exponential distribution with parameter $\lambda > 0$, that is

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x \leq \infty$$
$$= 0, \text{ elsewhere}$$

In practice the actual populations are often finite, so assuming the largest value of x in the population as D, the above frequency function can be approximated as

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \leq x \leq D \qquad (2.9)$$
$$= 0, \text{ elsewhere}$$

Note that we have here $x_0 = 0$ and $x_L = D$. If D is sufficiently large, (2.9) can be considered as an approximate exponential density otherwise the truncated exponential density is to be used and the expressions (2.10) - (2.12) are to be worked out accordingly.

Using (2.4), (2.5) and (2.6) we obtain

$$W_h = e^{-x_{h-1}/\lambda}\left(1 - e^{-y_h/\lambda}\right) \qquad (2.10)$$

$$\mu_h = \frac{(\lambda + x_{h-1})\left(1 - e^{-y_h/\lambda}\right) - y_h e^{-y_h/\lambda}}{1 - e^{-y_h/\lambda}} \qquad (2.11)$$

and

$$\sigma_h^2 = \frac{\lambda^2\left(1 - e^{-y_h/\lambda}\right)^2 - y_h^2 e^{-y_h/\lambda}}{\left(1 - e^{-y_h/\lambda}\right)^2} \qquad (2.12)$$

Using (2.10), (2.11) and (2.12), the problem of determining optimum strata boundaries, when the frequency of the main study variable X is given by (2.9), may be expressed as

Minimize

$$\sum_{h=1}^{L} e^{-x_{h-1}/\lambda} \sqrt{\lambda^2\left(1 - e^{y_h/\lambda}\right)^2 - y_h^2 e^{-y_h/\lambda}}$$

$$\text{subject to} \quad \sum_{h=1}^{L} y_h = d \qquad (2.13)$$

and $y_h \geq 0; h = 1,2,\ldots,L$

where d is obtained by (2.3) with $x_0 = 0$ and $x_L = D$.

3. THE SOLUTION

Consider the following subproblem of (2.8) for first k (< L) strata.

$$\text{Minimize} \quad \sum_{h=1}^{k} f_h(y_h)$$

$$\text{subject to} \quad \sum_{h=1}^{k} y_h = d_k \qquad (3.1)$$

and $y_h \geq 0; h = 1,2,\ldots,k$

where $d_k < d$ is the total width available for division into k strata.

Note that $d_k = d$ for $k = L$

Also $d_k = y_1 + y_2 + \ldots\ldots + y_k$

$d_{k-1} = y_1 + y_2 + \ldots\ldots + y_{k-1} = d_k + y_k$

$d_{k-2} = y_1 + y_2 + \ldots\ldots + y_{k-2} = d_{k-1} - y_{k-1}$

. .

. .

. .

$d_2 = y_1 + y_2 = d_3 - y_3$

and $d_1 = y_1 = d_2 - y_2$

If $f(k, d_k)$ denotes the minimum value of the objective function of (3.1), then

$$f(k,d_k)=\min\Big[\sum_{h=1}^{k}f_h(y_h)\Big/\sum_{h=1}^{k}y_h=d_k$$

$$\text{and } y_h\geq 0;h=1,2,....k\Big]$$

With the above definition of $f(k, d_k)$ the recurrence relations of the dynamic programming takes the form

$$f(k,d_k)=\min_{0\leq y_k\leq d_k}\left(f_k(y_k)+f(k-1,d_k-y_k)\right),\ k\geq 2 \quad (3.2)$$

For the first stage (i.e. k = 1)

$$f(1,d_1)=f_1(d_1) \Rightarrow y_1=d_1 \quad (3.3)$$

From f (L, d) the optimum width of L^{th} stratum, y_L, is obtained from $f(L-1, d-y_L)$ the optimum width of $(L-1)^{th}$ stratum, y_{L-1}, is obtained and so on until y_1 is obtained.

Using (3.3) and (3.2) the recurrence relations for MPP (2.13) are as given

For first stage (k = 1)

$$f(1,d_1)=\sqrt{\lambda^2\left(1-e^{-d_1/\lambda}\right)^2-d_1^2e^{-d_1/\lambda}}$$

$$\text{at } y_1=d_1 \quad (3.4)$$

because $x_{k-1}=x_0=0$, when k = 1.

For the stage k, where $k\geq 2$

$$f(k,d_k)=\min_{0\leq y_k\leq d_k}\Big[e^{(-d_k-y_k)/\lambda}\sqrt{\lambda^2\left(1-e^{-y_k/\lambda}\right)^2-y_k^2e^{-y_k/\lambda}}+f(k-1,d_k-y_k)\Big] \quad (3.5)$$

because $x_{k-1}=x_0+y_1+...+y_{k-1}=d_k-y_k$

4. A NUMERICAL EXAMPLE

Executing a computer program the recurrence relations (3.4) and (3.5) are solved to seek optimum stratum widths y_k; (k = 1,2,....., L) for the exponential study variable with density function given in (2.9), with D = 20 and $\lambda = 1$.

Table 1 gives the optimum values of y_h, x_h and $\sum_{h=1}^{L}f_h(y_h)$ for L = 2,3,4, and 5.

Table 1

No. of strata L	Strata widths y_h^*	Strata boundary points $x_h^*=x_{h-1}^*+y_h^*$	Optimum values of the objective function $\sum_{h=1}^{L}f_h(y_h)=\sum_{h=1}^{L}W_h\sigma_h$
2	$y_1^*=1.2610$ $y_2^*=18.7390$	$x_1^*=x_0+y_1^*=1.2610$	0.5341
3	$y_1^*=0.7678$ $y_2^*=1.2501$ $y_3^*=17.9821$	$x_1^*=x_0+y_1^*=0.7678$ $x_2^*=x_1^*+y_2^*=2.0179$	0.3648
4	$y_1^*=0.5509$ $y_2^*=0.7638$ $y_3^*=1.2513$ $y_4^*=17.4340$	$x_1^*=x_0+y_1^*=0.5509$ $x_2^*=x_1^*+y_2^*=1.3147$ $x_3^*=x_2^*+y_3^*=2.5650$	0.2770
5	$y_1^*=0.4393$ $y_2^*=0.5610$ $y_3^*=0.7569$ $y_4^*=1.2688$ $y_5^*=16.9740$	$x_1^*=x_0^*+y_1^*=0.4393$ $x_2^*=x_1^*+y_2^*=1.0003$ $x_3^*=x_2^*+y_3^*=1.7572$ $x_4^*=x_3^*+y_4^*=2.0260$	0.2233

The total width availabe for cutting stratum boundaries is taken as 20 units, that is the largest population value $X_L = D = 20$, because the area to the right of $X = 20$ for exponential distribution is almost zero, when $\lambda = 1$.

REFERENCES

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

Buhler, W. and Deutler, T. (1975). Optimum stratification and grouping by dynamic programming. *Metrika*, **22**, 121-175.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Statist. Inst.* **38(2)**, 345-358.

Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuartidskrift*, **33**, 203-213.

Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification-II. *Skand. Akt.*, **34**, 133-148.

Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. *J. Amer. Statist. Assoc.*, **54**, 88-101.

Ekman, G. (1959). Approximate expression for the conditional mean and variance over small intervals of a continuous distribution. *Ann. Math. Statist.*, **30**, 1131-1134.

Hess, I., Sethi, V.K. and Balakrishnan, T.R. (1966). Stratification : A practical investigation. *J. Amer. Statist. Assoc.*, **61**, 74-90.

Khan, E.A., Khan, M.G.M. and Ahsan, M.J. (2002). Optimum Stratification: A Mathematical Programming Approach. *Cal. Stat. Assoc. Bull.*, **52 (Special Volume)**, 323-333.

Mahalanobis, P.C. (1952). Some aspect of the design of sample surveys. *Sankhya*, **12**, 1-17.

Mehta, S., Singh, R. and Kishore, L. (1996). On optimum stratification for allocation proportional to strata totals. *J. Ind. Statist. Assoc.*, **34**, 9-19.

Serfling, R. J. (1968). Approximately optimal stratification. *J. Amer. Statist. Assoc.*, **63**, 1298-1309.

Sethi, V.K. (1963). A note on optimum stratification of population for estimating the population mean. *Austr. J. Statist.*, **5**, 20-23.

Singh, R. (1975a). An alternative method of stratification on the auxiliary variable. *Sankhya*, **C37**, 100-108.

Singh, R. (1975b). On optimum stratification for proportional allocation. *Sankhya*, C37, 109-115.

Singh, R. (1975c). A note on optimum stratification in sampling with varying probabilities. *Austr. J. Statist.*, **27**, 12-21.

Singh, R. (1977). A note on optimum stratification for equal allocation with ratio and regression methods of estimation. *Austr. J. Statist.*, **19(2)**, 69-104.

Singh, R. and Prakash, D. (1975). Optimum allocation for equal allocation. *Ann. Inst. Statist. Math.*, **27**, 273-280.

Singh, R. and Sukhatme, B.V. (1969). Optimum stratification. *Ann. Inst. Statist. Math.*, **21**, 515-528.

Singh, R. and Sukhatme, B.V. (1972). Optimum stratification in sampling with varying probabilities. *Ann. Inst. Statist. Math.*, **24**, 485-494.

Singh, R. and Sukhatme, B.V. (1973). Optimum stratification with ratio and regression methods of estimation. *Ann. Inst. Statist. Math.*, **25**, 627-633.

Taga, Y. (1967). On optimum stratification for the objective variable based on concomitant variable using prior information. *Ann. Inst. Statist. Math.*, **19**, 101-129.

Unnithan, V.K.G. (1978). The minimum variance boundary points of stratification. *Sankhya*, **40C**, 60-72.

Unnithan, V.K.G. and Nair, U. (1995). Minimum variance stratification. *Comm. Stat.-Theory Methods*, **24(1)**, 275-284.

Yadava, S.S. and Singh, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Comm. Stat. - Theory Methods*, **13(22)**, 2793-2806.