Approximation method in finding optimum stratum depending on Neyman allocation applied on Beta distribution

AHMAD MAHIR, R., AL-KHAZALEH A. M. H. and AL-KASSAB, M. M.T
School of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D. E.
MALAYSIA

Abstract:- The problem of optimum stratification on the auxiliary variable y for Neyman allocation has been considered. A Taylor Expansion rule of finding approximately optimum strata boundaries and general formulas for stratum weight, stratum mean and variance mean depending on Neyman Allocation have been suggested for Beta distribution. A numerical investigation into the boundaries stratum and variance mean of Neyman allocation with respect to optimum allocation has also been made for different values to the parameter Beta distribution.

Key-Words: Optimum Stratification, Optimum Strata Boundaries, Neyman Allocation, Taylor Expansion, Beta Distribution

1 Introduction

For a given method of allocation, the variance is clearly a function of the strata boundaries. Dalenius [1950] first considered the problem of determining optimum strata boundaries, i.e., optimum stratification. By minimizing the variance of the estimate, $\overline{y}_{st}$ sets of equations were obtained, solutions to which gave optimum strata boundaries for Neyman allocations. These equations involved population parameters, which were functions of the optimum strata boundaries. Subsequently, various authors gave methods of obtaining approximations to the exact solutions of the minimal equations. For an excellent account of these investigations, reference may be made to Cochran [1977]. He has given a set of sufficient conditions under which the optimum strata boundaries for the general case reduce to the optimum strata boundaries for the case when the strata are necessarily intervals. In most of these investigations of the problem of optimum stratification, both the estimation and the stratification variables are taken to be the same. Since the distribution of the estimation variable y is rarely known in practice, we consider the problem of optimum stratification on the variable y. Since these equations cannot be solved easily, various methods of finding approximations to the exact solutions have been given. The results for Neyman allocation come out as a special case of the results. The paper concludes by comparing approximate solutions with the exact solutions for certain populations also $Cumf^{5/6}$. Dalenius and Hodges (1959) suggested $Cumf^{1/2}$ which gives the equation $V_{Ney}\left(\overline{y}_{st}\right)=\dfrac{k^{4}(y)}{12nl^{2}}$ where $k(y)=cum.f^{1/2}=\int_{-\infty}^{\infty} f^{1/2}(y)dy$. Al-Kassab and Al-Taay (1994) suggested method $Cum.f^{2/3}$, which gives $V_{Ney}\left(\overline{y}_{st}\right)=\dfrac{M^{3}(y)}{12nl^{2}}$ where $M(y)=cum.f^{2/3}=\int_{-\infty}^{\infty} f^{2/3}(y)dy$. Al-Haso (1996) suggested the methods $cumf^{1/3}$ and $cumf^{5/6}$, which give these equations respectively:

$$V_{Ney}\left(\overline{y}_{st}\right)=\frac{H^{6}(y)}{12nl^{2}}$$ where

$$H(y)=cum.f^{1/3}=\int_{-\infty}^{\infty} f^{1/3}(y)dy$$ and

$$V_{Ney}\left(\overline{y}_{st}\right)=\frac{C^{12/5}(y)}{12nl^{2}}$$

with $C(y) = cum.f^{5/6} = \int_{-\infty}^{\infty} f^{5/6}(y)dy$

2 Procedures

The suffix h denotes the stratum and i the unit within the stratum. The following symbols all refer to stratum h.

N_h Total number of units in stratum h.

n_h Number of units in sample from stratum h.

y_{hi} Value obtained for the ith unit from stratum h.

$W_h = \frac{N_h}{N}$ Stratum weight which is in proportion to the population size attributable to stratum h.

$f_h = \frac{n_h}{N}$ Sampling fraction in the stratum h.

$\overline{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$ True mean of stratum h.

$\overline{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ Sample mean of stratum h.

When a population of N units is to be stratified into L strata and the samples from each stratum are selected with simple random sampling,
An unbiased estimate of population mean for the estimation variable y, is given by

(1) $$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{\mu}_h$$

where in the $h(h = 1, 2, ..., L)$ th stratum $\overline{y}_h$ is the sample mean based where $\sum_{h=1}^{L} n_h = n$ units and W_h (stratum weight) on n_h is the proportion of population units falling in that stratum. If the finite population corrections are neglected in each stratum, the estimate in (1) has the variance

(2) $$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^{\,2} \sigma_h^{\,2} / n_h$$

where $\sigma_h^{\,2}$ is the variance of y in the hth stratum.

For the Neyman allocation method where for all h

(3) $$n_h = n \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h},$$

$h = 1, 2, ..., L$

The variance in (2) reduces to

(4) $$V_{Ney} = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2$$

Let $a = y_{\circ}$ and $b = y_L$ be the smallest and largest values of y in the population. By differentiating with respect to the stratum boundaries. We get the minimum variance. $V_{Ney}(\overline{y}_{st})$ with respect to y_h since y_h appears in the sum only in the terms $W_h \sigma_h$ and $W_{h+1}\sigma_{h+1}$.

Hence we have the formula for finding the optimum stratum [Cochran, 1977].

(5) $$\frac{(y_h - \mu_h)^2 + \sigma_h^{\,2}}{\sigma_h} = \frac{(y_h - \mu_{h+1})^2 + \sigma_{h+1}^{\,2}}{\sigma_{h+1}}$$

Where $h = 1, 2, 3, ..., L-1$

These equations are difficult to solve, since μ_h and σ_h depend on y_h .so we must use the iterative method to solve them by using computer program such as C++.

3 The construction of strata:

The choice of a sample size depends on the size of the population in stratum h, the variance of population in stratum h and the cost of taking the sample in the strata h denoted by C_h .The simplest cost function is the form Total cost $= C = C_0 + \sum_{h=1}^{L} C_h n_h$

The question of how strata are made will now be considered. The basic consideration involved in the formation of strata is that the strata should be internally homogeneous. For example, if units are to be selected at random from within strata, strata variances for the character under estimation should be as homogenous as possible. This can be achieved by allocating units believed to be similar to the same stratum. The ideal situation is that in which the distribution of y is available. Then the strata would be created by cutting this distribution at suitable points. Let the distribution of y be continuous with the density function, $f(y), a < y < b$ in order to make L strata, the range of y is to be cut up at points $y_1 < y_2 < y_3 < ... < y_{L-1}$. The relative frequency W_h, the mean μ_h and the variance σ_h^2 of the hth stratum are given by

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

(7) $$\mu_h = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} yf(y)dy$$

(8) $$\sigma_h^2 = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 f(y)dy - \mu_h^2$$

The mean of population as the following

(9) $$\mu = \sum_{h=1}^{L} W_h \mu_h = \int_a^b yf(y)dy$$

4-Aproximated using Taylor expansion

Taylor Expansion defined as the following

(10) $$f(x) = \sum_{k=0}^{\infty} f^{(k)}(c)\frac{(x-c)^k}{k!}$$

Where $c \in [a,b]$ and $f^{(k)}(c)$ denoted for k derivative where k =0, 1, 2,..

Beta distribution defined as the following

(11) $$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}; x \in (0,1)$$

Where α, β are the parameters of beta function.

Let the distribution of y be continuous with the density function, $f(y), a < y < b$. In order to make L strata, we expand the function using the Taylor expansion to approximate the function $f(y)$. The range of y is to be cut up at points $y_1 < y_2 < y_3 < ... < y_{L-1}$. The relative frequency W_h of the hth stratum are given by

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

therefore

$$W_h = \int_{y_{h-1}}^{y_h} \{f^{(0)}(c) + f^{(1)}(c)(y-c) + \frac{f^{(2)}(c)(y-c)^2}{2!} + \frac{f^{(3)}(c)(y-c)^3}{3!} + ...\}dy$$

We found the derivatives $f^{(0)}, f^{(1)}, f^{(2)}, f^{(3)}, ...$ for beta distribution, and then substitute in the expression and simplified until, we got the general formula for W_h as the following

(13)

$$W_h = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left\{\sum_{n=0}^{D-k-1}\sum_{k=0}^{D-1}\sum_{t=0}^{D-1}\left\{\frac{(-1)(-1)}{t+1}\binom{\alpha-1}{n}\binom{\beta-1}{k}\binom{n+k}{t} c^{\alpha+k-t-1}(1-c)^{\beta-k-1}(y_h^{t+1} - y_{h-1}^{t+1})\right\}\right\}$$

$\alpha \,\&\, \beta \in n, \alpha \,\&\, \beta \geq 1$

Where D is the number of boundaries cut.

Similarly we got the W_{h+1}

(14)

$$W_{h+1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left\{\sum_{n=0}^{D-k-1}\sum_{k=0}^{D-1}\sum_{t=0}^{D-1}\left\{\frac{(-1)(-1)}{t+1}\binom{\alpha-1}{n}\binom{\beta-1}{k}\binom{n+k}{t} c^{\alpha+k-t-1}(1-c)^{\beta-k-1}(y_{h+1}^{t+1} - y_h^{t+1})\right\}\right\}$$

$\alpha \,\&\, \beta \in n, \alpha \,\&\, \beta \geq 1$

We got the mean μ_h by using the formula

(15) $$\mu_h = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} yf(y)dy$$

After expanding the function $f(y)$ it become

$$\mu_h = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} \{f^{(0)}(c)x + f^{(1)}(c)x(x-c) + \frac{f^{(2)}(c)x(x-c)^2}{2!} + \frac{f^{(3)}(c)x(x-c)^3}{3!} + ...\}dx$$

After simplified and substitute the derivatives, we got the general formula of the mean stratum for beta function

(16)

$$\mu_h = \frac{\overline{)\alpha+\beta}}{\overline{)\alpha)\beta}} \frac{1}{W_h} \left\{ \sum_{n=0}^{D-k-1} \sum_{k=0}^{D-1} \sum_{t=0}^{D-1} \{ \frac{(-1)^t (-1)^n}{t+2} \binom{\alpha-1}{n} \binom{\beta-1}{k} \binom{n+k}{t} c^{\alpha+k-t-1} (1-c)^{\beta-k-1} (y_h{}^{t+2} - y_{h-1}{}^{t+2}) \} \right\}$$

$$\alpha \,\&\, \beta \in n \,, \alpha \,\&\, \beta \geq 1$$

Similarly we got the μ_{h+1}

(17)

$$\mu_{h+1} = \frac{\overline{)\alpha+\beta}}{\overline{)\alpha)\beta}} \frac{1}{W_{h+1}} \left\{ \sum_{n=0}^{D-k-1} \sum_{k=0}^{D-1} \sum_{t=0}^{D-1} \{ \frac{(-1)^t (-1)^n}{t+2} \binom{\alpha-1}{n} \binom{\beta-1}{k} \binom{n+k}{t} c^{\alpha+k-t-1} (1-c)^{\beta-k-1} (y_{h+1}{}^{t+2} - y_h{}^{t+2}) \} \right\}$$

$$\alpha \,\&\, \beta \in n \,, \alpha \,\&\, \beta \geq 1$$

In addition, we got the mean variance as the following

(18) $$\sigma_h{}^2 = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y^2 f(y) dy - \mu_h{}^2$$

$$\sigma_h{}^2 = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} \{ f^{(0)}(c) y^2 + f^{(1)}(c) y^2 (y-c) + \frac{f^{(2)}(c) y^2 (y-c)^2}{2!} + \ldots \} dy - \mu_h{}^2$$

After simplified and substitute the derivatives, we got the general formula of the variance mean for stratum h :

(19)

$$\sigma_h{}^2 = \frac{\overline{)\alpha+\beta}}{\overline{)\alpha)\beta}} \frac{1}{W_h} \left\{ \sum_{n=0}^{D-k-1} \sum_{k=0}^{D-1} \sum_{t=0}^{D-1} \{ \frac{(-1)^t (-1)^n}{t+3} \binom{\alpha-1}{n} \binom{\beta-1}{k} \binom{n+k}{t} c^{\alpha+k-t-1} (1-c)^{\beta-k-1} (y_h{}^{t+3} - y_{h-1}{}^{t+3}) \} \right\} - \mu_h{}^2$$

$$\alpha \,\&\, \beta \in n \,, \alpha \,\&\, \beta \geq 1$$

Similarly we got the general formula for the variance mean for stratum $h+1$:

(20)

$$\sigma_{h+1}{}^2 = \frac{\overline{)\alpha+\beta}}{\overline{)\alpha)\beta}} \frac{1}{W_{h+1}} \left\{ \sum_{n=0}^{D-k-1} \sum_{k=0}^{D-1} \sum_{t=0}^{D-1} \{ \frac{(-1)^t (-1)^n}{t+3} \binom{\alpha-1}{n} \binom{\beta-1}{k} \binom{n+k}{t} c^{\alpha+k-t-1} (1-c)^{\beta-k-1} (y_{h+1}{}^{t+3} - y_h{}^{t+3}) \} \right\} - \mu_{h+1}{}^2$$

$$\alpha \,\&\, \beta \in n \,, \alpha \,\&\, \beta \geq 1$$

4 Result and discussion

We compare the variance mean between the three methods of exact, $Cumf\ {}^5\!/_6$ and alternative method using Taylor expansion. We reach to the general formulas, which we can calculate, the boundaries stratum by using the iterative method when we use equation (5). To get the variance mean we compensate the $W_h, \sigma_h{}^2$ in the equation (4). The following tables obtain the values of the variance mean for the three methods. The exact method, $Cumf\ {}^5\!/_6$ and the alternative method by approximation using Taylor expansion.

Table 1

Simulation result comparing the variance mean between the three methods, exact, $Cumf\ {}^5\!/_6$ and alternative method using Taylor expansion for different values α, β, c, D

$\alpha = 1, \beta = 2$			
$nV_{Ney}(\bar{x}_{st})$			
L	Exact	$Cum.f\ {}^5\!/_6$	Taylor Expansion
2	0.0150	1.95×10^{-2}	0.0150
3	0.0069	8.65×10^{-3}	0.0069
4	0.0039	4.86×10^{-3}	0.0039
5	0.0025	3.11×10^{-3}	0.0025
6	0.0018	2.16×10^{-3}	0.0018
7	0.0013	1.59×10^{-3}	0.0013
8	0.0010	1.22×10^{-3}	0.0010
$\alpha = 2, \beta = 1$			
$nV_{Ney}(\bar{x}_{st})$			
L	Exact	$Cum.f\ {}^5\!/_6$	Taylor Expansion
2	0.0150	1.95×10^{-2}	0.0150
3	0.0069	8.65×10^{-3}	0.0069
4	0.0039	4.86×10^{-3}	0.0039
5	0.0025	3.11×10^{-3}	0.0025
6	0.0018	2.16×10^{-3}	0.0018
7	0.0013	1.59×10^{-3}	0.0013
8	0.0010	1.22×10^{-3}	0.0010

$\alpha = \beta = 2$				
$nV_{Ney}(\bar{x}_{st})$				
L	Exact	$Cum.f^{5/6}$	Taylor Expansion	
			D=2 C=0.5	D=3
2	0.0150	1.99×10^{-2}	0.0469	0.0150
3	0.0069	8.86×10^{-3}	0.0208	0.0069
4	0.0039	4.98×10^{-3}	0.0117	0.0039
5	0.0025	3.12×10^{-3}	0.0075	0.0025
6	0.0018	2.21×10^{-3}	0.0052	0.0019
7	0.0013	1.63×10^{-3}	0.0038	0.0014
8	0.0010	1.25×10^{-3}	0.0029	0.0011

From the previous tables we deduce the following. When we increase the number of stratum, we got the minimum mean variance in the three methods. The value of variance mean for exact method less than, the approximated $Cum.f^{5/6}$ method for all the values of α,β .Also the values of α,β increase and the number of boundaries cut decrease the mean variance is decrease for the suggested method by using Taylor expansion. When $(\alpha = 2, \beta = 1)$ and $(\alpha = 1, \beta = 2)$ the $nV_{ney}(\overline{y}_{st})$ are equal but, the boundaries stratum varies, since the function is linear equation and when we approximate any linear equation by using Taylor expansion since $\beta(\alpha,\beta) = \beta(\beta,\alpha)$ (Beta function).It gives the same function so the approximate did not depend on the constant C. We observed that $nV_{ney}(\overline{y}_{st})$ by using Taylor expansion is more less from both methods exact and approximate $Cumf^{5/6}$ when the α,β are increase and D is decrease. The $nV_{ney}(\overline{y}_{st})$ are equal for exact method and approximated method using Taylor expansion when $\alpha + \beta \le D + 2$ for any values α,β .Finally we choose a suitable value for C to give minimum variance.

References:

[1] Al-Haso, Azhar, Method in finding optimum stratification using NeymanAllocation. Almousel University, Iraq, 1996.

[2] Al-Kassab, M. M.T. & Al-Taay, H., Approximately optimal stratification using Neyman Allocation, *J. of Tanmiat Al-Rafidain*, Vol. 19, Issue 50, pp.31-37, 1997.

[3] Chaudhary, F.S. & Kumar, Anil, A sampling strategy for Mean Based on Three-Way stratification, *J Sankhya: The Indian Journal of statistics*, Vol. 50, Series B, Pt. 1, p.p. 69-77,1988.

[4] Cochran, W.G., Comparison of Method for Determing Stratum Boundaries, *Bulletin of the International Statistical Institute*, Vol. 38, p.p. 345-348 ,1961.

[5] Cochran, W.G., *Sampling Technique*, 3rd Edition, John Wiley, New York, 1977.

[6] Dalenius, T. (1950). The problem of stratification, *skand.Aktuarietidskr.*,33,203-213.

[7] Dalenius, T. & Hodges, J.L., Minimum Variance Stratification, *Journal of American Statistical Association*, Vol. 54, p.p.88-101, 1959.

[8] Serfling, R.S, Approximately Optimal Stratification, *Journal of American Statistical Association*, Vol. 63,p.p. 1298-1309, 1968.

[9] Sethi, V.K., A note on Optimum Stratification for Estimation the Population means, *Australian Journal of Statistics*, Vol. 5, p.p. 20-33.1963.

[10] Singh, R., Approximately Optimal Stratification on the Auxiliary Variable, *Journal of American Statistical Association*, Vol. 66,p.p. 829-833, 1971.