A NOTE ON OPTIMUM STRATIFICATION

AVINDRA SINGH
Punjab Agricultural University Ludhiana
AND
B. V. SUKHATME
Iowa State University Ames

(*Received in Nov. 1970*)

INTRODUCTION

Let $f(x)$ denote the probability density function of the auxiliary variable x and let the regression of y on x be given by

$$y = c\,(x) + e, \qquad \ldots \quad (1.1)$$

where $c(x)$ is a function of x and e is such that $E(e/x)=0$ and $V(e/x)=\varphi(x)>0$ for all x in the range (a, b) with $(b-a)<\infty$. Under this set up, Singh and Sukhatme (1969) have shown that in case of stratified simple random sampling the solutions to the systems of minimal equations, *viz.*

$$\frac{[c(x_h)-\mu_{hc}]^2+\sigma_{hc}{}^2+\varphi(x_h)+\mu_{h\varphi}}{\sqrt{\sigma_{he}{}^2+\mu_{h\varphi}}}$$

$$=\frac{[C(x_h)-\mu_{ic}]^2+\sigma_{ic}{}^2+\varphi(x_h)+\mu_{i\varphi}}{\sqrt{\sigma_{ic}{}^2+\mu_{i\varphi}}} \qquad \ldots (1.2)$$

and $$x_h=\frac{\mu_{hc}+\mu_{ic}}{2}, \qquad \ldots (1.3)$$

where in the *h-th* stratum,

μ_{hc}=Expected value of $c(x)$,

$\sigma_{hc}{}^2$=Variance of $c(x)$,

$\mu_h\varphi$=Expected value of $\varphi(x)$,

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

give optimum strata boundaries (in the sense of min. variance) on the auxiliary variable x under Neyman and proportional allocations respectively, provided that the functions $p(x)=f(x)$. $[4\varphi(x)c'^2(x)+\varphi'^2(x)]$. $(\varphi(x))^{-3/2}$ and $c'^2(x)f(x)$ are bounded and possess first two derivatives for all x in (a, b).

The system of equations (1.2) has been shown to be equivalent to the system

$$K^2_h \int_{x_{h-1}}^{x_h} p(t)dt. \ [1+0(K^2_h)]=K^2_i \int_{x_h}^{x_{h+1}} p(t)dt. \ [1+0(K^2_i)]$$

$$i=h+1, \ h=1, 2, \ldots, L-1 \qquad \ldots(1.4)$$

where for the *h th* stratum

$K_h=x_h-x_{h-1}$,

x_h=Upper boundary,

x_{h-1}= lower boundary, and $0(K_h^2)$ approaches zero faster than K_h^2.

If the number of strata L is taken to be large so that the strata widths K_h are small and the terms of $0(m^4)$, $m=\underset{(a,\,b)}{\text{Sup}}\,(K_h)$, are negligible, the system (1.4) can be approximated by the system

$$K_h^2 \int_{x_{h-1}}^{x_h} p(t)dt=\text{constant} \qquad \ldots \ (1.5)$$

or equivalently by

$$Q(x_{h-1}, x_h)=\text{constant}, \qquad \ldots(1.6)$$

where

$$Q(x_{h-1}, \ x_h)=K_h^2 \int_{x_{h-1}}^{x_h} p(t)dt. \ [1+0(K_h^2)] \qquad \ldots(1.7)$$

Now one can obtain various functions $Q(x_{h-1}, x_h)$ as defined in (1.7). Each such function gives a system of equations, the solutions to which provide approximately optimum strata boundaries (AOSB) on the variable x. In theorem 2.1 of this note we give certain asymptotic properties of all such AOSB.

2. Properties of the AOSB

As mentioned in the previous section, various systems of equations can be developed from (1.5) and (1.6) and let (x_{hi}) be the solution to the *i-th* such system. Define

$$F(x_{h-1}, x_h) = K_h^2 \int_{x_{h-1}}^{x_h} p(t)\, dt. \; [1 + 0(K_h^2)] \qquad \ldots(2.1)$$

and

$$E_L\,[x_h] \sum_{h=1}^{L} F(x_{h-1},\; x_h), \qquad ..\,(2.2)$$

so that the variance $V(\bar{y}_{st})_N$ of the stratified estimate $\bar{y}_{st}$ of population mean, with Neyman allocation of the sample to the strata, can be obtained from

$$\sqrt{nV(\bar{y}_{st})_N} = \int_a^b \sqrt{\varphi(t)}\; .\; f(t)dt + \frac{E_L\,[x_h]}{96} \qquad ..\,(2.3)$$

where n is the total sample size.

Also let $[x_h]$ denote the exact solutions to the minimal equations (1.2). Then, we have the following theorem :

Theorem 2.1 If the function $p(x)$ is bounded away from zero and possesses first two derivatives for all x in (a, b), with $(b - a) < \infty$, then as $L \to \infty$

(A) $\sup\limits_{(a,\, b)} (x_{hi} - x(_{h-1})i) \to 0$

(B) $\sup\limits_{(a,\, b)} \mid x_{hi} - x_h \mid = 0 \left[\sup\limits_{(a,\, b)} (x_{hi} - x(_{h-1})_i) \right]^2 \to 0$

(C) $1 - E_L\,[x_h]/E_L\,[x_{hi}] = 0 \left[\sup\limits_{(a,\, b)} (x_{hi} - x_{(h-1)})_i \right]^2 \to 0$

(D) $\lim E_L\; [x_{hi}] = \lim\; E_L\,[x_h] = \left[\int_a^b \sqrt[3]{p(t)}\;\; dt. \right]^3 \Big/ L^2$

Proof :– (A) If $K_{hi} = x_{hi} - x(_{h-1})_i$, then as $L \to \infty$, $\inf\,(K_{hi}) \to 0$. Suppose it is not true and $\lim\limits_{L \to \infty} \inf\,(K_{hi}) = \epsilon > 0$, then $\sum\limits_{h=1}^{L} K_{hi} > L\epsilon$

and a value L_0 of L can be found such that for all $L>L_0$, $L\epsilon>(b-a)$, which is impossible. Hence we have $\lim\inf_{L\to\infty} (K_{hi})=0$

Now let $j(i)$ be the stratum for which $K_{ji}=\inf (K_{hi})$. As inf. $(K_{hi})=K_{ji}\to 0$, σ^2_{jix}, K^2_{ji} and C_i, where C_i is the constant of the *i-th* system, tend to zero. Since the equation,

$$K^2_{hi}\int_{x_{(h-1)i}}^{x_{hi}} p(t)dt=C_i$$

(of the *i-th* system) holds for all h and $0<p(x)<\infty$ for all x in (a,b),

$$\sup_{(a,\,b)} (K_{hi})\to 0 \text{ as } L\to\infty.$$

(*B*) The system of equations (1.4) can also be put as

$$A(x_{h-1},\, x_h)=B(x_h,\, X_{h+1}) \qquad \ldots(2.4)$$

where $A(x_h,\, x_{h+1})$ and $B(x_h,\, x_{h+1})$ differ in respect of the terms $0(K^2_{h+1})$ in $[1+0(K^2_{h+1})]$. On expanding the two sides of (2.4) with the help of Taylor's theorem about the points $(x_{(h-1)i},\, x_{hi})$ and $(x_{hi},\, x_{(h+1)i})$ respectively, we get the equation (2.4) as

$$K^2_{hi}\int_{x_{(h-1)i}}^{x_{hi}} p(t)dt.\ [1+0(K^2_{hi})]+z_{(h-1)i}\frac{\partial A(U_{(h-1)i},\, U_{hi})}{\partial U_{(h-1)i}}$$

$$+z_{hi}\frac{\partial A(U_{(h-1)i},\, U_{hi})}{\partial U_{hi}}$$

$$=K^2_{(h+1)i}\int_{x_{hi}}^{x_{(h+1)i}} p(t)dt.\ [1+0(K^2_{(h+1)i}]+z_{hi}\frac{\partial B(U_{hi},\, U_{(h+1)i})}{\partial U_{hi}}$$

$$+z_{(h+1)i}\frac{\partial B(U_{hi},\, U_{(h+1)i})}{\partial U_{(h+1)i}} \qquad \ldots(2.5)$$

where $z_{hi}=x_h-x_{hi}$, $U_{hi}=x_{hi}+\theta_h z_{hi}$, $0\le \theta_h<1$.

Since

$$K^2_{hi}\int_{x_{(h-1)i}}^{x_{hi}} p(t)dt=K^2_{(h+1)i}\int_{x_{ki}}^{x_{(h+1)i}} p(t)dt.=C_i\ ,$$

the equations (2.5) can be put as

$$-W_{(h-1)i}+2W_{hi}-W_{(h+1)i}=R_{hi} \qquad \ldots(2.6)$$

where

$$\left.\begin{aligned} W_{(h-1)i}&=-z_{(h-1)i}\,.\,\frac{\partial A(U_{(h-1)i},\,U_{hi})}{\partial U_{(h-i)}}\\ W_{hi}&=z_{hi}\left[\frac{\partial A(_{(h-1)i},\,U_{hi})}{\partial U_{hi}}-\frac{\partial B(U_{hi},\,U_{(h+1)i})}{\partial U_{hi}}\right]\\ W_{(h+1)i}&=z_{(h+1)i}\,.\,\frac{\partial B(U_{hi},\,U_{(h+1)i})}{\partial U_{(h+1)i}}\end{aligned}\right\} \ldots(2.7)$$

and

$$R_{hi}=K^2{}_{(h+1)i}\int_{x_{hi}}^{x'_{(h+1)i}} p(t)dt.\ 0(K^2{}_{(h+4)i})-K^2{}_{hi}\int_{x_{(h-1)i}}^{xh_i} p(t)dt.\ 0(K^2{}_{hi})$$

It can be easily seen that the solution

$$W_{ji}=L^{-1}\left[(L-j)\sum_{h=1}^{j} hR_{hi}+j\sum_{h=j+i}^{L-1}(L-h)R_{hi}\right], j=h-1,\ h,\ h+1$$

satisfies the equation (2.6).

Now we have

$$K_{ji}/K_{hi}=[C_i/p(x_{(j-1)i}+\theta_jK_{ji})]^{1/3}/[C_i/p(x_{(h-1)i}+\theta_hK_{hi})]^{1/3}=0(1),$$

where $0\leq\theta_j,\ \theta_h<1$ and also $C_i^{1/3}=\mathrm{C}(K_{hi})$ for any h.

Also

$$\begin{aligned}K_{(h+1)}-K_{hi}&=C_i^{1/3}[p^{-1/3}(X_{hi}+\theta_{(h+1)}K_{(h+1)i}-p^{-1/3}\\ &\qquad(x_{hi}-(1-\theta_h)K_{hi})]\\ &=0(K^2{}_{hi})+0(K^2{}_{(h+1)i})\end{aligned}$$

Therefore,

$$R_{hi}=C_i^{1/3}\ 0(K_h{}^5{}_i).$$

Writing $m_i=\sup\limits_{(a,b)}(K_{hi})$ and taking $K=$constant $<\infty$,

$$|\,W_{ji}\,|=|\,C_i^{1/3}.0(L.\sum_{h=1}^{L}K_h{}^5{}_i)\,|\leq KLm_i^4.(b-a)C^{1/3}=0(m_i^4) \quad \ldots(2.8)$$

Since

$$\inf.\ K_{hi}\leq L^{-1}(b-a)\leq \sup_{(a,b)}(K_{hi})=m_i \text{ so that } L=0\,[m_i^{-1}(b-a)].$$

As the coefficients of z_{ji} ($j=h-1, h, h+1$) in (2.7) are of order $0(m_i^2)$, we find from (2.8),

$$W_{ji}=z_{ji}[\text{Coefficient of } 0(m_i^2)]=0(m_i^4).$$

Hence $\qquad z_{ji}=0(m_i^2)$ ...(2.9)

which establishes (*B*).

(*C*) Let $\triangle_i$ denote the difference $E_L\,(x_{hi})-E_L\,[x_{hi}]=\sum_{h=1}^{L} F(x_{(h-1)i}, x_{hi})$

$$-\sum_{h=1}^{L} F(x(_{h-1}, x_h), \text{ so that } \triangle_i \geqslant 0.$$

Now to prove this part we expand $F(x_{(h-1)i}, x_{hi})$ about (x_{h-1}, x_{yi}) with the help of Taylor's theorem. Thus

$$\sum_{h=1}^{L} F(x_{(h-1)i}, x_{hi})=\sum_{h=1}^{L} F(x_{h-1}, x_h)+\tfrac{1}{2}\sum_{h=1}^{L-1}$$

$$z_{hi}^2\left[\frac{\partial^2 F(v_{(h-1)i}, v_{hi})}{\partial^2 v_{hi}}+\frac{\partial^2 F(v_{hi}, v_{(h+1)i})}{\partial^2 v_{hi}}\right]+\sum_{h=1}^{L-1} z_{hi}\, z_{(h-1)i}$$

$$\frac{\partial^2 F(v_{(h-1)i}, v_{hi})}{\partial v_{(h-1)i}\partial v_{hi}} \qquad \text{.. (2.10)}$$

where

$$z_{hi}=x_{hi}--x_h,\ v_{ji}=v_j-\theta z_{ji} \text{ and } 0<\theta<1.$$

The terms involving first partial derivatives of $F(x_{(h-1)i}, x_h)$ cancel out in the light of the *i-th* system of equation of which $[x_{hi}]$ are the solutions.

From (2.9), $z_{hi}=0\ (m^2_i)$. Since it is easy to verify that the partial derivatives in (2.10) are of order $0(m_i)$, it is clear that

$$\triangle_i=\sum_{h=1}^{L-1} \iota(m_i^4).\ 0(m_i)+\sum_{h=1}^{L-1} 0(m_i^4).\ (m_i),$$

$$=L.0(m^5_i),$$

$$=0(m_i^4). \qquad \text{...(2.11)}$$

As

$$F(x_{(h-1)i}, x_{hi}) = K^2_{hi} \int_{x(h-1)_i}^{x_{hi}} p(t)\, dt. \left[1+0(K^2_{hi})\right]$$

and $0<p(x)<\propto$, we get $F(x_{(h-1)i}, x_{hi})=0(m_i^3)$.
Thus

$$\sum_{h=1}^{L} F(x_{(h-1)i}, x_{hi}) = E_L\,[x_{hj}] = 0(m_i^3),$$

which gives us

$$(1-E_L\,[x_h]/E_L\,(x_{hi}]) = \frac{\Delta_i}{E_L\,[x_{hi}]} = \frac{0(m_i^4)}{0(m_i^2)} = 0(m_i^2). \qquad \ldots(2.12)$$

(*D*) As can be easily seen, we have

$$K^2_{hi} \int_{x(h-1)_i}^{x_{hi}} p(t)\, dt.\ [1+0(K^2_{hi})] = \left[\int_{x(h-1)_i}^{x_{hi}} p(t)\, dt.\right]^3 . \left[1+0(K_{hi}^2)\right]$$

$$= C_i\left[1+0(K_{hi}^2)\right], \qquad \ldots(2.13)$$

where C_1 is the constant of the system of equations

$$\left[\int_{x(h-1)}^{x_h} \sqrt[3]{p(t)}\, dt.\right]^3 = C_1 \qquad \ldots(2.14)$$

Thus

$$\int_{x(h-1)_i}^{x_{hi}} \sqrt[3]{p(t)}\, dt. = C_1^{1/3} \left[1+0(K^2_{hi})\right]$$

adding over all the strata

$$\sum_{h=1}^{L} \int_{x(h-1)_i}^{x_{hi}} \sqrt[3]{p(t)}\, dt. = \int_a^b \sqrt[3]{p(t)}\, dt. = L.C_1^{1/3}[1+0(m_i^2)]_2$$

which gives

$$C_1 = \left[\int_a^b \sqrt[3]{p(t)}\, dt./L\right]^3 [1+0(m_i^2)]. \qquad \ldots(2.15)$$

But

$$E_L\,[x_{hi}] = \sum_{h=1}^{L} K_{hi}^2 \int_{x(h-1)_i}^{x_{hi}} p(t)\, dt.\ [1+0(K_{hi}^2)] = L.C_1\,[1+0(m_i^2)],$$

which from (2.15) is equal to

$$\left[\left\{ \int_a^b \sqrt[3]{p(t)}\, dt\right\}^3 /L^2 \right] . \left[1+0(m^2{}_i)\right].$$

Therefore,

$$\lim_{L\to\infty} E_L \left[x_{hi}\right] = \left[\int_a^b \sqrt[3]{p(t)}\, dt\right]^3 \Big/ L^2. \qquad \ldots(2.16)$$

From (C) and (2.16) one obtains

$$\mathop{\text{Lim}}_{L\to\infty} E_L\, [x_{hi}] = \lim_{L\to\infty} E_L\, [x_h] = \left[\int_a \sqrt[3]{p(t)}\, dt. \right]^3 \Big/ L^2,$$

which proves (*D*).

This proves the asymptotic equivalence of the approximate solutions $[x_{hi}]$ and the exact solutions $[x_h]$. Since the system of equations (1.3) for proportional allocation is a particular case of the system (1.2), it is not considered separately.

3. Summary

Singh and Sukhatme (1969) have considered the problem of optimum stratifiation on an auxiliary variable x when the form of the regression of estimation variable y on the auxiliary variable x as also the form of the conditional variance function $V(y/x)$ are known. Minimal equations, solutions to which are the optimum strata boundaries, have been obtained for Neyman and proportional allocations. Since the minimal equations cannot be solved exactly, various methods of finding the approximate solutions have been suggested. The present note gives certain asymptotic properties of the approximately optimum strata boundaries that can be obtained by following the methods suggested in Ekman (1959).

References

Ekman (1959)	: An approximation useful in univariate stratification. Ann. Math. Stat. 30, 219-229.
Ekman (1959)	: Approximate expression for the conditional mean and variance over small intervals of a continuous distri bution, Ann. Math. Stat., 30, 1131-1134.
Ravindra Singh and B.V, Sukhatme (1969)	: 'Optimum Stratification' Ann. Inst. Stat. Math., 21, 515-28.