CHAPTER III :

ON LINEAR INVARIANT UNBIASED ESTIMATORS

3.0 SUMMARY

In this chapter we obtain necessary and sufficient conditions on the sampling design for the existence of a linear invariant unbiased estimator of the total of a finite population. The same conditions are shown to imply the existence of a linear invariant unbiased estimator which is admissible in the class of all linear unbiased estimators. We also introduce the concept of connectedness in sampling designs and prove that connected designs satisfy the above conditions and hence admit a linear invariant, unbiased estimator which is admissible.

3.1 INTRODUCTION

Let $D = (S,P)$ be a sampling design given in (1.2.3). The problem of unbiased estimation of the population total $Y = \sum_{i=1}^{N} Y_i$ is considered here.

A homogeneous linear estimator $t(s,\underset{\sim}{Y})$ has the form given in (1.2.6).

The conditions for $t(s,\underset{\sim}{Y})$ to be unbiased for Y are

$$\sum_{s \ni i} b(s,i)\, p(s) = 1 \qquad i = 1,2,\ldots,N. \qquad \ldots(3.1.1)$$

The conditions for $t(s,\underset{\sim}{Y})$ to be invariant are (1.2.11).

Let $\pi_i = \sum_{s \ni i} p(s)$ denote the inclusion probability for the unit i. It is easy to verify that (3.1.1) can be fulfilled by taking $b(s,i) = 1/\pi_i$ and conditions (1.2.11) can obviously be satisfies by taking $b(s,i) = N/n(s)$, where $n(s)$ denotes number of units in sample s. However it is not clear whether both (3.1.1) and (1.2.11) can be simultaneously satisfied by a suitable choice of $b(s,i)$. We now obtain a necessary condition on the design for the existence of $b(s,i)$ satisfying both (3.1.1) and (1.2.11). Next we show that this condition is also sufficient. In fact, using the condition, we construct a linear invariant unbiased estimator which is admissible among all linear unbiased estimators. The condition involves certain connectedness properties of the sampling design.

3.2 CONNECTEDNESS AND THE C-MATRIX IN SAMPLING DESIGNS

Definition 3.2.1 : For a sampling design $D=(S,P)$ for $\mathcal{U}$, we say that a unit i and a unit j are connected if we can find samples $s_1, \ldots, s_n$ in S and units $i_1, \ldots, i_{n-1}$ such that s_1 contains both i and i_1, s_2 contains both i_1 and i_2, ..., s_n contains both i_{n-1} and j.

It is clear that the relation of "being connected" is an equivalence relation. Therefore $\mathcal{U}$ splits into equivalence classes $\mathcal{U}_1, \ldots, \mathcal{U}_k$ which will be called components of $\mathcal{U}$ under the design D.

Definition 3.2.2 : The design $D = (S,P)$ is called a connected design for $\mathcal{U}$, if $\mathcal{U}$ itself is an equivalence class or equivalently if every pair of units of $\mathcal{U}$ is connected.

Let $$C_{ii} = \pi_i = \sum_{s \ni i} [p(s)/n(s)] \quad ; \text{ and}$$

$$C_{ij} = - \sum_{s \supset \{i,j\}} [p(s)/n(s)]$$

for $i,j = 1,2, \ldots, N$ and $i \neq j$. ...(3.2.1)

Then the N x N matrix $C = (C_{ij})$ will be called the C-matrix of the sampling design. It is easy to see that C is symmetric.

Also

$$C_{ij} \leq 0 \quad \text{for } i \neq j \qquad \ldots(3.2.2)$$

$$\sum_{j=1}^{N} C_{ij} = 0 \quad \text{for } i = 1,2,\ldots, N. \qquad \ldots(3.2.3)$$

These three properties are also associated with the C-matrix of an incomplete block design. The main difference between the C-matrix of an incomplete block design and the C-matrix of a sampling design is that C_{ij}'s in the first case are rational, while in the latter case they may be irrational. However, these three properties enable us to prove the theorem 3.2.1.

Theorem 3.2.1 : A Sampling design is connected if, and only if, the rank of its C-matrix is N-1. The rank of the C-matrix of a sampling design D equals N-k, where k is the number of components of $\mathcal{U}$ under D.

Proof : (a) Suppose the design is connected.

Let $\underset{\sim}{e}$ denote the column vector all of whose N entries are equal to 1.

Equation (3.2.3) shows that $C\underset{\sim}{e} = \underset{\sim}{0}$. Suppose now that $\underset{\sim}{x} = (x_1, \ldots, x_N)'$ is a column vector such that $C\underset{\sim}{x} = \underset{\sim}{0}$. We

show that $\underset{\sim}{x}$ is a multiple of $\underset{\sim}{e}$. This would imply that rank of $C = N-1$.

Let M = Maximum $(x_1, \ldots, x_N)$ and m = minimum $(x_1, \ldots, x_N)$. If possible suppose $M > m$. Let $A = \{ j : x_j = M \}$. Then A is a proper non-empty subset of $\mathcal{U}$. Since the design is connected there is an $i \in A$ and a $j_o \in \mathcal{U} - A$, such that i and j_o belong to some sample $s \in S$. This implies that

$$C_{ij_o} < 0 \qquad \text{by (3.2.1).}$$

Now

$$j \in A \Rightarrow x_j = M \Rightarrow C_{ij} x_j = MC_{ij}$$

and

$$j \in \mathcal{U} - A \Rightarrow x_j < M \Rightarrow C_{ij} x_j > M C_{ij},$$

with strict inequality for $j = j_o$. Therefore

$$\sum_{j=1}^{N} C_{ij} x_j > M \sum_{j=1}^{N} C_{ij} = 0. \text{ Thus } C\underset{\sim}{x} \neq \underset{\sim}{0}.$$

This contradiction proves $M = m$.

In otherwords $\underset{\sim}{x}$ is a multiple of $\underset{\sim}{e}$ and thus rank $C = (N-1)$.

(b) Suppose that the design is not connected. Let $\mathcal{U}_1, \ldots, \mathcal{U}_k$ be the components of $\mathcal{U}$ under the design $D=(S,P)$ Then S can be written as the disjoint union of $S_1, \ldots, S_k$ such that S_r contains samples whose units come exclusively

from $\mathcal{U}_r$. Further the union of all samples in S_r must be $\mathcal{U}_r$. Let N_r be the size of $\mathcal{U}_r$ and $\alpha_r = \sum_{s \in S_r} p(s)$. If we write $P_r = \{p(s)/\alpha_r;\ s \in S_r\}$ then (S_r, P_r) is a connected sampling design for $\mathcal{U}_r$. Now the matrix C of the design D is the direct sum of matrices $\alpha_1 C_1, \ldots, \alpha_k C_k$ where C_r is the C-matrix of a connected sampling design for $\mathcal{U}_r$. From part (a) of the proof, rank $C_r = N_r - 1$. Hence rank

$$C = \sum_{r=1}^{k} (N_r - 1) = (N-k).$$

This completes the proof of the theorem.

3.3 A NECESSARY CONDITION FOR THE EXISTENCE OF A LINEAR INVARIANT UNBIASED ESTIMATOR

We will again use the quantities $\mathcal{U}_r, S_r, N_r$ and α_r of part (b) of the proof of Theorem 3.2.1. Suppose that for estimating the population total $Y = \sum_{i=1}^{N} Y_i$ there is a linear invariant unbiased estimator

$$t(s, \underset{\sim}{Y}) = \sum_{i \in s} b(s,i)\, Y_i. \qquad \ldots(3.3.1)$$

Then the quantities b(s,i) satisfy the conditions (1.2.11) and (3.1.1).

From (3.1.1) we get

$$\sum_{i \in \mathcal{U}_r} \sum_{s \ni i} b(s,i)\, p(s) = N_r \quad r=1,2,\ldots,k. \quad \ldots(3.3.2)$$

Similarly from (1.2.11) we get

$$\sum_{s \in S_r} \sum_{i \in s} b(s,i)\, p(s) = N \alpha_r \quad r=1,2,\ldots,k. \quad \ldots(3.3.3)$$

However, the left sides of (3.3.2) and (3.3.3) are the same.

Therefore $N\alpha_r = N_r \quad r = 1,2, \ldots, k. \quad \ldots(3.3.4)$

Thus we have proved the following theorem.

Theorem 3.3.1 : A necessary condition for the existence of a linear invariant unbiased estimator for $\sum_{i=1}^{N} Y_i$ is that $N\alpha_r = N_r$ for all $r = 1,2,\ldots,k$. This condition always holds for a connected design.

3.4 THE EXISTENCE OF AN ADMISSIBLE LINEAR INVARIANT UNBIASED ESTIMATOR

We now prove that the necessary condition (3.3.4) of theorem 3.3.1 is also sufficient. In fact, we show that the resulting estimator is also admissible in the class of all linear unbiased estimators. We will follow Godambe [3]

and try to find an estimator which minimises the average of the variances at the points $\underset{\sim}{Y}_1 = (1,0,\ldots 0)$; $\underset{\sim}{Y}_2 = (0,1,0,\ldots 0)$; ...; $\underset{\sim}{Y}_N = (0,0 \ldots, 0,1)$ within the class of all linear invariant unbiased estimator of $Y = \sum_{i=1}^{N} Y_i$.

Let $t(s,\underset{\sim}{Y})$ have the form (3.3.1) and let V_i denote the variance of $t(s,\underset{\sim}{Y})$ at $\underset{\sim}{Y}_i$. Then

$$1 + V_i = \sum_{s \ni i} b^2(s,i)p(s). \qquad \ldots(3.4.1)$$

To minimise $\sum_{i=1}^{N} V_i$, subject to conditions (3.1.1) and (1.2.11), we consider

$$\phi = \sum_{i=1}^{N} \sum_{s \ni i} b^2(s,i)\, p(s) - 2 \sum_{i=1}^{N} \lambda_i \sum_{s \ni i} b(s,i)p(s)$$

$$-2 \sum_{s \in S} \mu_s \sum_{i \in s} b(s,i). \qquad \ldots(3.4.2)$$

where λ_i and μ_s are Lagrange's multipliers. Differentiating ϕ w.r.t. $b(s,i)$ and equating the derivatives to zero we get

$$b(s,i)\, p(s) = \lambda_i\, p(s) + \mu_s \quad \text{or}$$

$$b(s,i) = \lambda_i + a_s \qquad \ldots(3.4.3)$$

where $a_s = \mu_s/p(s)$. Write $\bar{\lambda}_s = \sum_{i \in s} [\lambda_i/n(s)]$.

Then (3.4.3) and (1.2.11) give

$$N = \sum_{i \in s} b(s,i) = n(s)\left[\bar{\lambda}_s + a_s\right] \quad \text{or}$$

$$a_s = \frac{N}{n(s)} - \bar{\lambda}_s .$$

Therefore (3.4.3) can be written as

$$b(s,i) = (\lambda_i - \bar{\lambda}_s) + \left[N/n(s)\right] . \qquad \ldots(3.4.4)$$

If we substitute for b(s,i) in (3.1.1) from (3.4.4), we get

$$\sum_{s \ni i} \left[(\lambda_i - \bar{\lambda}_s) + N/n(s)\right] p(s) = 1 \qquad i=1,2,\ldots,N. \qquad \ldots(3.4.5)$$

After some simplification (3.4.5) can be written as

$$\sum_{j=1}^{N} C_{ij} \lambda_j = d_i \qquad i = 1,2, \ldots, N$$

where C_{ij} are given by (3.2.1)

and $d_i = 1 - N \sum_{s \ni i} \left[p(s)/n(s)\right]$ $\qquad i=1,2, \ldots, N.$ $\qquad \ldots(3.4.6)$

Thus we have to solve the system

$$C \underset{\sim}{\lambda} = \underset{\sim}{d} . \qquad \ldots(3.4.7)$$

Theorem 3.4.1 : Suppose the sampling design is such that $N\alpha_r = N_\gamma$ for $r = 1,2, \ldots, k$. Then the system (3.4.7) is consistent and the resulting estimator given by (3.4.4) is unique. The estimator is linear invariant and unbiased for Y and is admissible within the class of all unbiased linear estimators.

Proof : The matrix C is the direct sum of the matrices $\alpha_1 C_1, \ldots, \alpha_k C_k$, where C_r is the C-matrix of a connected sampling design for the component $\mathcal{U}_r$. Therefore the system (3.4.7) is equivalent to

$$\alpha_r C_r \underset{\sim}{\lambda}_r = \underset{\sim}{d}_r \qquad r=1,2, \ldots, k. \qquad \ldots(3.4.8)$$

where $\underset{\sim}{\lambda}_r$ and $\underset{\sim}{d}_r$ have obvious meanings. Since C_r corresponds to a connected sampling design for $\mathcal{U}_r$, theorem 3.2.1 shows that the rank of C_r is (N_r-1). Therefore the only linear restriction satisfied by the row vectors of $\alpha_r C_r$ is that their sum is the zero vector. Therefore (3.4.8) is consistent whenever the sum of the entries in $\underset{\sim}{d}_r$ is zero. That is, we must have

$$\begin{aligned} 0 = \sum_{i \in \mathcal{U}_r} d_i &= N_r - N \sum_{i \in \mathcal{U}_r} \sum_{s \ni i} \left[p(s)/n(s)\right] \\ &= N_r - N \sum_{s \in S_r} \sum_{i \in s} \left[p(s)/n(s)\right] \\ &= N_r - N \sum_{s \in S_r} p(s) = N_r - N\alpha_r. \end{aligned}$$

But this condition holds by hypothesis. Thus (3.4.8) is consistent.

Now if $\underset{\sim}{x}_r$ is a column vector such that $C_r \underset{\sim}{x}_r = \underset{\sim}{0}$, then the entries of $\underset{\sim}{x}_r$ must be all the same because C_r is a C-matrix of rank $N_r - 1$. Therefore if $\underset{\sim}{\lambda}_r$ and $\underset{\sim}{\lambda}_r^*$ are two solutions of (3.4.8), then the entries of $\underset{\sim}{\lambda}_r - \underset{\sim}{\lambda}_r^*$ must be all the same. But then it is clear that the quantities $b(s,i)$ given by (3.4.4) is the same whether it is computed from $\underset{\sim}{\lambda}_r$ or from $\underset{\sim}{\lambda}_r^*$. It follows that the resulting estimator $t(s,\underset{\sim}{Y})$ is the unique estimator which minimises the average of the variances at $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_N$ within the class of those linear unbiased estimators which attain zero variance at $(1,1,\ldots,1)$. Therefore $t(s,\underset{\sim}{Y})$ is admissible.

Remark 1 : It follows from theorem 3.4.1 that a connected design always admits an admissible linear invariant unbiased estimator of Y. In practice, however, the estimator of theorem 3.4.1 is quite difficult to obtain.

Remark 2 : Roy-Chakravarti [17] called a design balanced if $\sum_{s \ni i} [p(s)/n(s)] = N^{-1}$ for $i = 1,2, \ldots, N$. It is clear that this condition is equivalent to $d_i = 0$ for all i. For

such a design the estimator of theorem 3.4.1 reduces to

$$t(s,\underset{\sim}{Y}) = \frac{N}{n(s)} \sum_{i \in s} Y_i .$$

Remark 3 : For a unicluster design, the only unbiased linear estimator is the Horvitz-Thompson estimator. This estimator is linear invariant if, and only if,

$$p(s) = n(s)/N \qquad \text{for all } s \in S.$$

We now give a simple example illustrating the use of theorem 3.4.1.

Suppose we take samples of 2 units from a population of 3 units. Suppose that the three samples $s_1 = \{1,2\}$; $s_2 = \{2,3\}$; $s_3 = \{3,1\}$ have probabilities p_1, p_2, p_3 respectively. The entries in the C-matrix and the vector $\underset{\sim}{d}$ are easily calculated. The system $C\underset{\sim}{\lambda} = \underset{\sim}{d}$ can be written as

$$\begin{aligned}
(p_1+p_3)\,\lambda_1 - p_1\lambda_2 - p_3\lambda_3 &= 3p_2-1, \\
-p_1\lambda_1 + (p_2+p_1)\,\lambda_2 - p_2\lambda_3 &= 3p_3-1, \\
-p_3\lambda_1 - p_2\lambda_2 + (p_3+p_2)\,\lambda_3 &= 3p_1-1. \qquad \ldots(3.4.9)
\end{aligned}$$

If we eliminate λ_3 from the first two equations of (3.4.9) we get

$$(p_1p_2 + p_2p_3 + p_3p_1)(\lambda_1 - \lambda_2) = p_2(3p_2-1) - p_3(3p_3-1). \quad \ldots(3.4.10)$$

Equation (3.4.10) suggests that we try the solution

$$\lambda_1 = \frac{p_2(3p_2-1)}{\Delta}; \quad \lambda_2 = \frac{p_3(3p_3-1)}{\Delta}; \quad \lambda_3 = \frac{p_1(3p_1-1)}{\Delta} \quad \ldots(3.4.11)$$

where $\Delta = p_1p_2 + p_2p_3 + p_3p_1$. It is simple to check that the solution given by (3.4.11) satisfies (3.4.9). Therefore the estimator of theorem 3.4.1 is given by

$$b(s_i,i) = \frac{3}{2} + \frac{\lambda_i - \lambda_{i+1}}{2}, \quad b(s_i,i+1) = \frac{3}{2} + \frac{\lambda_{i+1} - \lambda_i}{2}$$

where $i = 1,2,3$ and $(i+1)$ is interpreted as 1 when $i=3$.

This estimator seems to be new for this simple situation.