Chapter-VI

An Efficient Estimator of Population Mean Using a Special Case of PBIB Design under Controlled Simple Random Sampling without Replacement

6.1 Introduction

In chapter-II and chapter-III, we have proposed estimators under simple random sampling design but in hilly area, villages are generally widely spread and approachability to some villages is also difficult which results more variability between them. So, it becomes highly costly as well as time consuming to collect information on some represented sampled villages. In such situation, it will be convenient to use either stratified random sampling procedure or controlled sampling procedure instead of conventional sampling procedure under the given budget constraints. Suppose, we are interested to select a sample of size $n$ out of the population size $N$, under the given budget then there will be $\binom{N}{n}$ possible samples. Out of these samples, variability and the cost of collecting information on some samples will be very high because of administrative difficulties such as traveling, quality of data due to non-response and investigator bias etc. Such samples are termed as non-preferred samples under the given budget constraints and remaining samples as preferred samples by Goodman and Kish (1950). So, preferred sample is generally selected under the given constraints.
Stratified random sampling design is one of the procedures that minimize the selection of non-preferred samples to some extent which may not be always sufficient even after fully exploiting the mechanism of stratification. So, it may still be necessary to further control the selection of samples within each stratum. For example, suppose we have two strata consisting of first stage units from each of which one unit is to be selected. Further suppose that in stratum-I, units B, C & F lie adjacent to ocean or other major water way and in stratum-II, unit d is similarly located and that all other units are located inland. Assume the units and the probabilities of selection assigned to each of them are as follows:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Probability</th>
<th>Unit</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.10</td>
<td>a</td>
<td>0.15</td>
</tr>
<tr>
<td>B</td>
<td>0.15</td>
<td>b</td>
<td>0.30</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>c</td>
<td>0.10</td>
</tr>
<tr>
<td>D</td>
<td>0.20</td>
<td>d</td>
<td>0.20</td>
</tr>
<tr>
<td>E</td>
<td>0.25</td>
<td>e</td>
<td>0.25</td>
</tr>
<tr>
<td>F</td>
<td>0.20</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
\[
\begin{align*}
P_{\text{Stratum-I}}(\text{inland}) & = 0.55 \\
P_{\text{Stratum-I}}(\text{coastal}) & = 0.45 \\
P_{\text{Stratum-II}}(\text{inland}) & = 0.80 \\
P_{\text{Stratum-II}}(\text{coastal}) & = 0.20
\end{align*}
\]

(6.1.1)

It is desired to select one inland and one coastal unit and undesirable to select two coastal units. If we use the usual stratified random sampling, then probabilities of different combinations of units getting selected are:

\[
\begin{align*}
P(\text{one inland and one coastal unit}) & = 0.55\times 0.20 + 0.45\times 0.80 = 0.47 \\
P(\text{two inland units}) & = 0.55\times 0.80 = 0.44 \\
P(\text{two coastal units}) & = 0.45\times 0.20 = 0.09
\end{align*}
\]

(6.1.2)

An alternative procedure is to redefine the selection procedure. Rearrange the units in stratum-I by listing B, C, F first followed by A, D & E. Rearrange the units in stratum-II by shifting unit d to the end. Draw a random no. from 1 to 100. let this one no. determine the selection in both the strata. If the random no. is 45 or less, a coastal unit will be selected from stratum-I and inland unit from stratum-II.
If the no. is 46 to 80, an inland unit will be selected from both the strata. If the no. is greater than 80, an inland unit will be selected from stratum-I and a coastal unit from stratum-II. In this new sampling procedure, the probabilities of different combinations of units getting selected are:

$$P(\text{one inland and one coastal unit}) = 0.45 + 0.20 = 0.65$$
$$P(\text{two inland units}) = 1 - 0.65 = 0.35$$

The probability of selection of desired sampling units one from inland and one from coastal is more in case of alternative selection procedure. Goodman and Kish (1950) showed that this procedure retains equal probability of selection within strata to permit unbiased estimator of population mean.

In the present chapter, we use controlled sampling to further minimize the selection of non-preferred samples by using balanced incomplete block design to construct sampling frame in the population which retains the fundamental principals of random sampling technique. The expressions for the bias and the variance of conventional mean per unit estimator under the given sampling design have been obtained. It has been shown that efficiency of the estimator under the given design remains the same as in the case of simple random sampling design but the probability of selection of non-preferred samples becomes less in the proposed sampling design as compared to the conventional simple random sampling design.
6.2 Proposed Controlled Sampling Design and its Probabilities

For selecting a sample of size \( n \) from a population of size \( N \), a sampling frame is prepared in such a way that every unit of population replicates \( r \) times and satisfies the parametric relationships of balanced incomplete block design i.e.

1. \( vr = bk \)
2. \( \lambda(v - 1) = r(k - 1) \)

Where \( r, b \) and \( \lambda \) denote the number of replications, number of blocks and number of times a pair of units occur together in the same block respectively. This implies that total population units are divided into \( b \) blocks each of size \( k \).

Here \( N = v \)
\( n = k \)

Assuming that blocks are equally likely, we select a block of size \( k \) out of \( b \) blocks randomly, then

\[
P(\text{inclusion of } i^{th} \text{ unit in the sample}) = \frac{r}{b}
\]
\[
= \frac{k}{v} = \frac{n}{N} \quad ; \quad i = 1, 2, \ldots, v \quad (6.2.1)
\]

\[
P(\text{inclusion of } i^{th} \text{ and } j^{th} \text{ units in the sample}) = \frac{\lambda}{b}
\]
\[
= \frac{n(n - 1)}{N(N - 1)} \quad ; \quad i, j = 1, 2, \ldots, v
\]
From (6.2.1) & (6.2.2), we see that inclusion probabilities of units in the sample under the above sampling design is same as for the case of simple random sampling without replacement sampling design, but the probability of selection of preferred samples will be higher under the proposed sampling design as compared to the conventional simple random sampling design.

For example, suppose a sample of 3 villages is to be selected out of 7 villages. Suppose that 7 villages are located in hilly area as under:

```
<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
```

Then there are $7C_3 = 35$ possible combinations of size 3. These are:

```
(1 2 3)*  (1 2 4)  (1 2 5)  (1 2 6)*  (1 2 7)  
(1 3 4)  (1 3 5)  (1 3 6)*  (1 3 7)*  (1 4 5)  
(1 4 6)*  (1 4 7)*  (1 5 6)  (1 5 7)  (1 6 7)*  
(2 3 4)*  (2 3 5)  (2 3 6)*  (2 3 7)*  (2 4 5)  
(2 4 6)*  (2 4 7)*  (2 5 6)  (2 5 7)  (2 6 7)  
(3 4 5)  (3 4 6)  (3 4 7)*  (3 5 6)  (3 5 7)  
(3 6 7)  (4 5 6)  (4 5 7)  (4 6 7)*  (5 6 7)  
```
There are 14 samples in which distance between the villages is large marked with asterisk which are termed as non-preferred samples. If all the 35 samples are considered equally likely, then the probability of selection a non-preferred sample is clearly $\frac{14}{35} = 0.40$. Under the proposed scheme with $v = 7$, $b = 7$, $r = 3$, $k = 3$, $\lambda = 1$, we have prepared a sampling frame as:

<table>
<thead>
<tr>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>b₁</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

As per the original locations of the villages, we observe that block ‘b₇’ is not preferred one, so from the remaining preferred blocks, we select one block as a sample at random. It shows that the probability of obtaining a non-preferred sample is $\frac{1}{7} = 0.14$ which is much less than in simple random sampling if no control is employed.

### 6.3 Estimator and its Results

Let $Y_i; i = 1, 2, \ldots, N$ be the observation on the $i^{th}$ unit of the population. Suppose a random sample (block) of size $n (k)$ units is selected under proposed controlled sampling design defined
in section 6.2. Let \( y_1, y_2, \ldots, y_k \) be observations of study variable \( y \) made in the sample. The estimator of population mean \( \bar{Y} \) under the proposed sampling design is defined as:

\[
\bar{Y}_H = \frac{1}{k} \sum_{i=1}^{k} y_i
\]  
(6.3.1)

The estimator \( \bar{Y}_H \) is similar to the conventional mean per unit estimator \( \bar{y} \) under simple random sampling design, where

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]  
(6.3.2)

To find the bias and variance of the estimator \( \bar{Y}_H \) under the given design, we proceed as:

Defining

\[
\alpha_i = \begin{cases} 
1 & \text{if } i^{th} \text{ unit in the population is included in the sample} \\
0 & \text{Otherwise}
\end{cases}
\]

Here \( \alpha_i \) is the binomial variable taking values 1 and 0 with probability \( r/b \) and \( 1-r/b \) respectively. So,

\[
E(\alpha_i) = 1 \cdot P(\alpha_i = 1) + 0 \cdot P(\alpha_i = 0) = \frac{r}{b}
\]  
(6.3.3)
\[ E(a, a_j) = 1 \times P(a_i = 1 \cap a_j = 1) + 0 \times P(a_i = 0 \cap a_j = 0) = \frac{\lambda}{b} \quad (6.3.4) \]

Now the estimator \( \bar{y}_H \) can be put as:

\[ \bar{y}_H = \frac{1}{k} \sum_{i=1}^{\nu} a_i Y_i \quad (6.3.5) \]

Taking expectation on both sides of (6.3.5), we get

\[ E(\bar{y}_H) = E\left( \frac{1}{k} \sum_{i=1}^{\nu} a_i Y_i \right) = \frac{1}{k} \sum_{i=1}^{\nu} E(a_i) Y_i \]
\[ = \frac{r}{bk} \sum_{i=1}^{\nu} Y_i = \frac{rv}{bk} \bar{Y} \]
\[ = \bar{Y} \quad (\because vr = bk) \]

Hence, we have

\[ Bias(\bar{y}_H) = 0 \quad (6.3.6) \]

Therefore, proposed estimator under the given sampling scheme is unbiased.

The variance of the estimator \( \bar{y}_H \) is obtained as:

\[ V(\bar{y}_H) = \frac{1}{k^2} \left[ \sum_{i=1}^{\nu} Y_i^2 V(a_i) + \sum_{i=1, j=1, i \neq j}^{\nu} Y_i Y_j Cov(a_i, a_j) \right] \quad (6.3.7) \]
Using (6.3.3) and (6.3.4), we have following results:

\[ V(a_i) = E(a_i^2) - E^2(a_i) \]
\[ = \frac{r}{b} - \frac{r^2}{b^2} = \frac{r(b - r)}{b^2} \]  \hfill (6.3.8)

\[ Cov(a_i, a_j) = E(a_i a_j) - E(a_i)E(a_j) \]
\[ = \frac{\lambda}{b} - \frac{r^2}{b^2} = \frac{b \lambda - r^2}{b^2} \]  \hfill (6.3.9)

Substituting the results of (6.3.8) and (6.3.9) in (6.3.7), we get

\[ V(\bar{y}_H) = \frac{1}{k^2} \left[ \frac{r}{b} \sum_{i=1}^{k} Y_i^2 - \frac{r^2}{b^2} \sum_{i=1}^{k} Y_i^2 + \frac{\lambda}{b} \sum_{i=1}^{k} \sum_{j \neq i} Y_i Y_j - \frac{r^2}{b^2} \sum_{i=1}^{k} \sum_{j \neq i} Y_i Y_j \right] \]
\[ = \frac{1}{k^2} \left[ \frac{k}{v} \sum_{i=1}^{v} Y_i^2 - \frac{k^2}{v^2} \sum_{i=1}^{v} Y_i^2 + \frac{k(k - 1)}{v(v - 1)} \sum_{i=1}^{v} \sum_{j \neq i} Y_i Y_j - \frac{k^2}{v^2} \sum_{i=1}^{v} \sum_{j \neq i} Y_i Y_j \right] \]
\[ = \frac{v - k}{v^2 k} \left[ \sum_{i=1}^{v} Y_i^2 - \frac{1}{v - 1} \sum_{i=1}^{v} \sum_{j \neq i} Y_i Y_j \right] \]
\[ = \frac{v - k}{vk} \left[ \sum_{i=1}^{v} Y_i^2 - v \bar{Y}^2 \right] \]
\[ = \frac{v - k}{vk} \left[ \sum_{i=1}^{v} (Y_i - \bar{Y})^2 \right] \]

Hence, variance of the estimator \( \bar{y}_H \) is
\[ V(\bar{y}_H) = \left( \frac{1}{k} - \frac{1}{v} \right) S_y^2, \quad \text{where} \quad S_y^2 = \frac{1}{v-1} \sum_{i=1}^{v} (Y_i - \bar{Y})^2 \]  

(6.3.10)

From (6.3.10), the variance of \( \bar{y}_H \) is same as that of the variance of sample mean \( \bar{y} \) in case of simple random sampling without replacement. Hence, we conclude that without violating the fundamental properties of simple random sampling, the efficiency of the purposed sampling strategy remains same as that of sampling strategy under simple random sampling but cost of selecting sample reduces drastically in the sense that it reduces the probability of selection of non-preferred samples. Finally, we can say that under proposed sampling design more accurate information at smaller cost regarding the population on the basis of sample selected as compared to conventional simple random sampling design in the situation where the units of the populations are haphazardly located and it is very difficult to take measurement over some of the units.

**Summary**

For estimating the population mean, efficient sampling strategy has been proposed under controlled sampling design. Balanced incomplete block design is used for preparing the sampling frame which gives quite high probability of selection of preferred samples as compared to non-preferred sample. The expressions for bias and variance of proposed sampling strategy have been obtained. It has been shown that efficiency of the estimator under the given design remains the same as in the case of simple random sampling design but the probability of selection of non-preferred samples becomes less in the proposed sampling design as compared to the conventional simple random sampling design resulting low cost of survey sampling.