

APPLICATIONS OF PBIB DESIGNS IN CLUSTER SAMPLING

by D. RAGHAVARAO*, *Department of Mathematics, University of North Carolina at Chapel Hill, USA and*

RAJINDER SINGH, *Punjab Agricultural University, Ludhiana, Punjab.*

(Communicated by Prof. P. K. Menon, F.N.A.)

(Received 9 May 1974; after revision 6 January 1975)

The relationship between sampling techniques and experimental designs was first of all studied by Chakrabarti (1963) followed by Mohanty (1971). They discussed the relationship of some incomplete block designs with some sampling schemes. The results of Chakrabarti lend themselves to some logical generalizations. Some of these are discussed in the present paper.

In general, in heterogeneous populations it is sometimes advantageous to study relatively homogeneous groups of units, in the population, called clusters. The use of group divisible (GD) partially balanced incomplete block designs and balanced incomplete block (BIB) designs in cluster sampling is discussed. Sometimes the clustering of units may be in two directions with respect to two different extraneous factors. Sampling in such situations may be referred to as two dimensional cluster sampling. Partially balanced incomplete block (PBIB) designs of L_2 and rectangular type provide interesting applications in such cases, and these results are also discussed in the present paper. It is also shown that semi regular (SA) group divisible designs can also be used in stratified sampling.

INTRODUCTION AND SUMMARY

The close relationship between sampling techniques and experimental designs was only partly explored in literature (cf. Chakrabarti 1963; and Mohanty 1971). PBIB designs provide interesting applications to cluster sampling and in this paper we discuss the applications of PBIB designs to cluster sampling.

Definition 1: A Balanced Incomplete Block (BIB) design is an arrangement of v symbols into b sets each of k ($< v$) symbols, satisfying the following conditions :

- (i) Every symbol occurs at most once in each set;
 - (ii) Every symbol occurs in exactly r sets; and
 - (iii) Every pair of symbols occurs together in exactly λ sets, where
- v, b, r, k and λ are called the parameters of the BIB design and they satisfy

$$vr = bk, \lambda(v-1) = r(k-1) \quad \dots(1)$$

Partially balanced incomplete block (PBIB) designs are generalizations of BIB designs and to define them, we need the concept of association scheme as given in Definition 2.

*On leave from Punjab Agricultural University. Part of this work of this author was supported by NSF grants GU-2059 and GU-19568. Presently at Temple University, Philadelphia.

Definition 2: Given symbols $1, 2, \dots, v$ a relation satisfying the following conditions is said to be an association scheme with m classes :

(i) Any two symbols are either 1st, 2nd, \dots , or m -th associates, the relation of association being symmetrical; that is, if the symbol α is the i -th associate of β , then β is the i -th associate of α .

(ii) Each symbol α has n_i i -th associates, the number n_i being independent of α .

(iii) If any two symbols α and β are i -th associates, then the number of symbols that are j -th associates of α , and k -th associates of β , is p_{jk}^i and is independent of the pair of i -th associates α and β .

Given an association scheme for the v symbols, we define a PBIB design as follows :

Definition 3: If we have an association scheme with m classes and given parameters, we get a PBIB design with m associate classes if the v symbols are arranged into b sets of size k ($< v$) such that :

(i) Every symbol occurs at most once in a set;

(ii) Every symbol occurs in exactly r sets; and

(iii) If two symbols α and β are i -th associates, then they occur together in λ_i sets, the number λ_i being independent of the particular pair of i -th associates α and β .

The numbers n_i, p_{jk}^i, v are called the parameters of the association scheme and v, b, r, k, λ_i are called the parameters of the design. These parameters satisfy

$$vr = bk, \sum_{i=1}^m n_i = v - 1, \sum_{i=1}^m n_i \lambda_i = r(k - 1), \quad (2)$$

$$\sum_{k=1}^m p_{jk}^i = n_j - \delta_{ij}, n_i p_{jk}^i = n_j p_{ki}^i = n_k p_{ij}^k$$

where δ_{ij} is the Kronecker delta taking the value 1 or 0 according as $i = j$ or not.

In this paper we use PBIB designs with group divisible, L_2 and rectangular association schemes and for completeness, we introduce these association schemes in the following definitions :

Definition 4: A group divisible association scheme has $v = mn$ symbols divided into m groups of n symbols each, such that any two symbols of the same group are first associates and two symbols from different groups are second associates.

The PBIB designs with group divisible association scheme are called group divisible designs and are classified as

(a) Singular (S) if $r - \lambda_1 = 0$

(b) Semi-regular (SR) if $rk - v\lambda_2 = 0$, and $r - \lambda_1 > 0$

(c) Regular (r) if $rk - v\lambda_2 > 0$, and $r - \lambda_1 > 0$.

Definition 5: An L_2 association scheme has $v = s^2$ symbols arranged in a $s \times s$ square array such that symbols in the same row or column are first associates

while other pairs of symbols are second associates. PBIB designs with L_2 association scheme are called L_2 designs.

Definition 6: A rectangular association scheme is a three-associate-class association scheme with $v = mn$ symbols arranged in a rectangle with m rows and n columns. With respect to each symbol, the first associates are the other $n-1$ symbols of the same row, the second associates are the $m-1$ symbols of the same column, and the remaining $(m-1)(n-1)$ symbols are third associates, PBIB designs with rectangular association scheme are called rectangular designs.

For more details of these association schemes, we refer to Raghavarao (1971, Ch. 8).

In sampling from finite populations, it is often advantageous to form suitable clusters of units and surveying all the units or a fraction of units from selected clusters (cf. Murthy 1967, Ch. 8, 9). We discuss the problem of estimating population total or mean when the clusters are of equal size in the next Section and postpone the discussion on clusters of unequal sizes to the third Section. The clusters could be formed based on two extraneous factors A and B taking s and t levels respectively; there being in all st clusters. The problem of estimating population total when $s=t$ could be tackled through L_2 designs and this will be discussed in the Section titled "Use of L_2 designs in cluster sampling", and when $s \neq t$, we get required results with the help of rectangular designs as given in Section under, "Use of rectangular designs in cluster sampling". Finally we indicate some possible generalization of our results in the last Section.

USE OF GD DESIGNS IN CLUSTER SAMPLING WITH CLUSTERS OF EQUAL SIZE

Let the $v = MN$ population units be divided into M clusters each of N units and we are interested to take a sample of size n . Let y_{ij} be the value of the study variable on the j -th unit of the i -th cluster for $j = 1, 2, \dots, N; i = 1, 2, \dots, M$, Let

$$Y = \sum_{i=1}^M \sum_{j=1}^N y_{ij}, \bar{Y} = Y/MN, S^2 = \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - \bar{Y})^2 / (MN - 1).$$

$$\bar{Y}_i = \sum_{j=1}^N y_{ij} / N, S_i^2 = \sum_{j=1}^N (y_{ij} - \bar{Y}_i)^2 / (N - 1), \quad i = 1, 2, \dots, M. \quad (3)$$

Let there exist a GD design with parameters $v = MN, b, r, k = n, \lambda_1, \lambda_2$. Let the population units be identified with the symbols of the design such that the clusters coincide with the groups in the association scheme. The sampling procedure we follow will be to select a set of the GD design with equal probability and sample the units corresponding to the symbols of the selected set. Let \bar{y}_G be the sample mean and s_i^2 the variance of the study variable for the selected units of the i -th cluster ($i = 1, 2, \dots, M$).

Then, analogous to the results of Chakrabarti (1963) the following results can be easily established :

Theorem 1: \bar{y}_G is an unbiased estimator of \bar{Y} , that is, $\hat{\bar{y}} = \bar{y}_G$.

Corollary 1: $v\bar{y}_G$ is an unbiased estimator of Y , that is $\hat{Y} = v\bar{y}_G$.

Theorem 2: If $V(\cdot)$ denotes the variance of the estimator in parenthesis, we have

$$V(\bar{y}_G) = (vrn)^{-1} \{N(N-1)(\lambda_2 - \lambda_1) \sum_{i=1}^M S_i^2 + (v-1)(rn - v\lambda_2) S^2\} \quad (4)$$

Corollary 2:

$$V(v\bar{y}_G) = v(rn)^{-1} \{N(N-1)(\lambda_2 - \lambda_1) \sum_{i=1}^M S_i^2 + (v-1)(rn - v\lambda_2) S^2\}. \quad (5)$$

The $V(\bar{y}_G)$ given in (4) can be reduced in some cases by choosing the GD design to be semi-regular because the coefficient of S^2 becomes zero and in that case we will have the following results.

Theorem 3: If the sample is selected as one of the sets of a SRGD design, we have

$$V(\bar{y}_G) = \frac{v-n}{vMn} \sum_{i=1}^M S_i^2 \quad \dots \quad \dots \quad (6)$$

clearly

Theorem 4: If the sample is selected as one of the sets of a SRGD and if $n/M \geq 2$, then the estimated variance $\hat{V}(\bar{y}_G)$ of $V(\bar{y}_G)$ is

$$\hat{V}(\bar{y}_G) = \frac{v-n}{vMn} \sum_{i=1}^M s_i^2 \quad \dots \quad \dots \quad (7)$$

Corollary 3:

$$\hat{V}(v\bar{y}_G) = \frac{v(v-n)}{Mn} \sum_{i=1}^M s_i^2 \quad \dots \quad \dots \quad (8)$$

We easily observe that the method described in this section can be used in stratified sampling and the estimate \bar{y}_G and its variance by selecting the sample as a set of SRGD design are identical with stratified sample mean and its variance under proportional allocation.

SAMPLING FROM CLUSTERS OF UNEQUAL SIZES-USE OF BIB DESIGNS

Let there be M clusters, the i -th cluster having M_i population units and let

$\sum_{i=1}^M M_i = N$. As in the last section, let y_{ij} be the value on the study variable of the j -th unit of the i -th cluster

($j = 1, 2, \dots, M_i; i = 1, 2, \dots, M$). Let

$$Y_i = \sum_{j=1}^{M_i} y_{ij}, \bar{Y}_i = Y_i/M_i, Y = \sum_{i=1}^M M_i \bar{Y}_i, \bar{Y} = Y/N, \bar{Y} = Y/M,$$

$$S_{wi}^2 = \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2 / (M_i - 1), S_b^2 = \sum_{i=1}^M (Y_i - \bar{Y})^2 / (M - 1). \tag{9}$$

Let there exist a BIB design with parameters $v = M, b, r, k$ and λ . Let the symbols of the design correspond to the clusters. We select a set of the BIB design with equal probability in the first stage. If (i_1, i_2, \dots, i_k) is the selected set of the BIB design, then the clusters numbered i_1, i_2, \dots, i_k enter the sample. If we determine to have a sample of size n , then from i_α -th cluster we randomly select without replacement.

$$n_{i_\alpha} = \frac{M_{i_\alpha} n}{\sum_{\alpha=1}^k M_{i_\alpha}}$$

units for $\alpha = 1, 2, \dots, k$. The sampling from different clusters will be independently made.

Let

$$\bar{y}_B = \frac{M}{kN} \sum_{\alpha=1}^k M_{i_\alpha} \bar{y}_{i_\alpha}, \dots \dots \dots \tag{10}$$

where \bar{y}_{i_α} is the sample mean of the units of i_α -th cluster.

With a little algebra, analogous to the results of the previous Section, we have the following results :

Theorem 5: \bar{y}_B is an unbiased estimator of \bar{Y} .

Theorem 6: If B_i denotes the sum of the cluster sizes of the clusters in the i -th set of the BIB design and P_α denotes the sum of the B_i 's of the sets in which the α -th symbol occurs, we have

$$V(\bar{y}_B) = \frac{M}{rKN^2} \sum_{i=1}^M M_i \left(\frac{P_i}{n} - r \right) S_{wi}^2 + \frac{M(r - \lambda)(M - 1)}{rkN^2} S_b^2$$

Let $s'^2_{w_{i\alpha}}$ be the sample variance of the variable for the units of i_α -th cluster.

Then

Theorem 7: The estimated $V(\bar{y}_B)$ is given by

$$\begin{aligned} V(\bar{y}_B) = & \frac{M}{kN^2} \sum_{\alpha=1}^k \frac{M_{i_\alpha}^2 (M_{i_\alpha} - n_{i_\alpha})}{M_{i_\alpha} n_{i_\alpha}} s'^2_{w_{i\alpha}} \\ & + \frac{r - \lambda}{\lambda N^2} \left[\frac{M}{k} \sum_{\alpha=1}^k (M_{i_\alpha} \bar{y}_{i_\alpha})^2 - \frac{N^2}{M} \bar{y}^2 \right] \quad \dots \quad \dots \quad (11) \end{aligned}$$

USE OF L_2 DESIGNS IN CLUSTER SAMPLING

Let the population units be formed into clusters with the help of two factors A and B each at s levels. Thus there are in all s^2 clusters and each cluster can be designated by (ij) , where i denotes the level of A factor and j denotes the level of B factor. Let there be M_{ij} population units in the ij -th cluster. Let $y_{ij\alpha}$ be the observation of the study variable on the α -th unit of ij -th cluster ($\alpha = 1, 2, \dots, M_{ij}$; $i = 1, 2, \dots, s$; $j = 1, 2, \dots, s$).

Let

$$N_i = \sum_{j=1}^s M_{ij}, \quad N_j = \sum_{i=1}^s M_{ij}, \quad N = \sum_{i=1}^s \sum_{j=1}^s M_{ij}$$

$$Y_{ij} = \sum_{\alpha=1}^{M_{ij}} y_{ij\alpha}; \quad \bar{Y}_{ij} = Y_{ij}/M_{ij}, \quad Y_{i.} = \sum_{j=1}^s Y_{ij}, \quad Y_{.j} = \sum_{i=1}^s Y_{ij},$$

$$Y = \sum_{i=1}^s \sum_{j=1}^s Y_{ij}, \quad \bar{Y} = Y/N, \quad \bar{Y} = Y/s,$$

$$(M_{ij} - 1) S'^2_{w_{ij}} = \sum_{\alpha=1}^{M_{ij}} (y_{ij\alpha} - \bar{Y}_{ij})^2,$$

$$(s - 1) S'^2_{bA} = \sum_{i=1}^s (Y_{i.} - \bar{Y})^2, \quad (s - 1) S'^2_{bB} = \sum_{j=1}^s (Y_{.j} - \bar{Y})^2,$$

$$(s^2 - 1) S'^2 = \sum_{i=1}^s \sum_{j=1}^s \left(Y_{ij} - \frac{\bar{Y}}{s} \right)^2. \quad \dots \quad \dots \quad (12)$$

Let there exist a L_2 design with parameters $v = s^2, b, r, k, \lambda_1$ and λ_2 and let the symbols of this design be identified with the s^2 clusters. Let a set of the design be selected with equal probability and if T is the selected set, then the sample consists of the clusters $ij \in T$. If the sample size is fixed to be n , then from the selected ij -th cluster a simple random sample without replacement will be taken of size $n_{ij} = nM_{ij}/(\sum_{ij \in T} M_{ij})$. Sampling from different clusters will be independently made.

Let \bar{y}_{ij} be the sample mean of the selected ij -th cluster. Let $\bar{y}_L = \frac{s^2}{kN} \sum_{ij \in T} M_{ij} \bar{y}_{ij}$.

Analogous to the results of previous sections we will have

Theorem 8: \bar{y}_L is an unbiased estimator of \bar{Y}

Theorem 9:

$$V(\bar{y}_L) = \frac{b}{r^2 N^2} \left[(r - 2\lambda_1 + \lambda_2) (s^2 - 1) S'^2 + (\lambda_1 - \lambda_2) (s - 1) (S'^2_{bA} + S'^2_{bB}) + \sum_{i=1}^s \sum_{j=1}^s \left(\frac{P_{ij}}{n} - r \right) M_{ij} S'^2_{wij} \right] \dots \dots \quad (13)$$

where P_{ij} is the sum of the cluster sizes of all the clusters in the sets of L_2 design where symbol ij occurs.

$V(\bar{y}_L)$ can be reduced in some cases by choosing L_2 design to satisfy $r - 2\lambda_1 + \lambda_2 = 0$, as the coefficient of S'^2 becomes zero. The estimate of $V(\bar{y})$ can be easily determined through the standard techniques.

USE OF RECTANGULAR DESIGNS IN CLUSTER SAMPLING

As in the previous section we form st clusters of population units with the help of two factors A and B at s and t levels respectively. Let the ij -th cluster have M_{ij} elements. Let $y_{ij\alpha}$ be the observation of the study variable on the α -th unit of ij -th cluster ($\alpha = 1, 2, \dots, M_{ij}; i = 1, 2, \dots, s; j = 1, 2, \dots, t$). Let

$$\begin{aligned} N_{i.} &= \sum_{j=1}^t M_{ij}, N_{.j} = \sum_{i=1}^s M_{ij}, N = \sum_{i=1}^s \sum_{j=1}^t M_{ij}, \\ Y_{ij} &= \sum_{\alpha=1}^{M_{ij}} y_{ij\alpha}, \bar{Y}_{ij} = Y_{ij}/M_{ij}, Y_{i.} = \sum_{j=1}^t Y_{ij}, Y_{.j} = \sum_{i=1}^s Y_{ij}, \\ Y &= \sum_{i=1}^s \sum_{j=1}^t Y_{ij}, \bar{Y} = Y/N, (M_{ij} - 1) S'^2_{wij} = \sum_{\alpha=1}^{M_{ij}} (y_{ij\alpha} - \bar{Y}_{ij})^2, \\ (s - 1) S'^2_{bA} &= \sum_{i=1}^s \left(Y_{i.} - \frac{Y}{s} \right)^2, (t - 1) S'^2_{bB} = \sum_{j=1}^t \left(Y_{.j} - \frac{Y}{t} \right)^2, \\ (st - 1) S'^2 &= \sum_{i=1}^s \sum_{j=1}^t \left(Y_{ij} - \frac{Y}{st} \right)^2. \end{aligned} \quad (14)$$

Let there exist a rectangular design with parameters $v = st, b, r, k, \lambda_1, \lambda_2,$ and λ_3 and let the symbols of this design be identified with the st clusters such that the rows and columns of the association scheme coincide with the groups of clusters corresponding to constant level of factor A and factor B respectively. We follow a similar procedure of drawing the sample as described in the foregoing section. Let \bar{y}_{ij} be the sample mean of the selected ij -th cluster and let $\bar{\bar{y}}_R = \frac{st}{kN} \sum M_{ij} \bar{y}_{ij}$, where the summation is over the selected clusters. Then

Theorem 10: $\bar{\bar{y}}_R$ is an unbiased estimator of $\bar{\bar{Y}}$.

Theorem 11:

$$V(\bar{\bar{y}}_R) = \frac{b}{r^2 N^2} \left[(r - \lambda_1 - \lambda_2 + \lambda_3) (st - 1) S'^2 + (\lambda_1 - \lambda_3) (t - 1) S'^2_{bB} \right. \\ \left. + (\lambda_2 - \lambda_3) (s - 1) S'^2_{bA} + \sum_{i=1}^s \sum_{j=1}^t \left(\binom{P_{ij}}{n} - r \right) M_{ij} S'^2_{w_{ij}} \right], \quad (15)$$

where P_{ij} is the sum of the cluster sizes of all the clusters in the sets of design where symbol ij occurs.

$V(\bar{\bar{y}}_R)$ can be reduced in some cases by choosing rectangular design to satisfy $r - \lambda_1 - \lambda_2 + \lambda_3 = 0$ as the coefficient of S'^2 vanishes. The estimate of $V(\bar{\bar{y}}_R)$ can be easily obtained by text-book procedures.

CONCLUDING REMARKS

Hypercubic designs and extended group divisible designs (cf. Raghavarao 1971) could be effectively used to develop sampling schemes when the clusters are formed based on more than two factors and results in this direction will be discussed in a future communication. The relative efficiency of using different designs in a given sampling situation is also under study and we expect to discuss these results in a subsequent paper.

REFERENCES

- Chakrabarti, M. C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *J. Indian. statist. Assoc.*, **1**, 78-85.
- Mohanty (1971). Ph.D. thesis, *Unpublished. Inst. agric. Res. Statist., New Delhi.*
- Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, India.
- Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. John Wiley, New York.