*Part of this work of this author was supported by NSF grants GU-2059 and GU-19568.

[1]On Leave from Punjab Agricultural University (India).

Paper prepared in connection with the symposium on "Combinatorial Mathematics at New Delhi (India) from December 22 to December 27, 1972.

APPLICATIONS OF PBIB DESIGNS IN CLUSTER SAMPLING*

D. Raghavarao[1] and Rajinder Singh

Department of Statistics
University of North Carolina at Chapel Hill
and Punjab Agricultural University

Institute of Statistics Mimeo Series No. 855

December, 1972

APPLICATIONS OF PBIB DESIGNS IN CLUSTER SAMPLING

by

D. Raghavarao[1] and Rajinder Singh

University of North Carolina at Chapel Hill
and Punjab Agricultural University

1. INTRODUCTION AND SUMMARY. The close relationship between sampling techniques and experimental designs was only partly explored in literature (see Chakrabarti (1963), 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 1.1. A Balanced Incomplete Block (BIB) design which is an arrangement of v symbols into b sets each of k ($<v$) symbols, satisfying the following conditions:

1. Every symbol occurs at most once in each set.

2. Every symbol occurs in exactly r sets.

3. Every pair of symbols occurs together in exactly λ sets.

v, b, r, k and λ are called the parameters of the BIB design and they satisfy

$$vr = bk, \quad \lambda(v-1) = r(k-1) \quad \ldots (1.1).$$

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 1.2.

---

[1]On leave from Punjab Agricultural University. Part of this work of this author was supported by NSF grants GU-2059 and GU-19568.

Definition 1.2. Given symbols 1, 2, ..., v a relation satisfying the following conditions is said to be an association scheme with m classes:

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

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

3. If any two symbols α and β are i-th asociates, 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 1.3. 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

1. Every symbol occurs at most once in a set.

2. Every symbol occurs in exactly r sets.

3. 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

$$vr = bk\ ,\quad \sum_{i=1}^{m} n_i = v-1\ ,\quad \sum_{i=1}^{m} n_i \lambda_i = r(k-1)\ ,$$
$$\sum_{k=1}^{m} p_{jk}^{i} = n_j - \delta_{ij},\ n_i p_{jk}^{i} = n_j p_{ki}^{j} = n_k p_{ij}^{k}\ , \tag{1.2}$$

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 defintions:

Definition 1.4. 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 1.5. An L_2 asociation 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 1.6. 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 other $m-1$ 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 cluster of units and surveying all the units or a fraction of units from selected clusters (see Murthy (1967), Ch. 8, 9). We discuss the problem of estimating population total or mean when the clusters are of equal size in Section 2 and postpone the discussion on clusters of unequal sizes to Section 4. 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 Section 4 and when $s \neq t$, we get required results with the help of rectangular designs as given in Section 5. Finally we indicate some possible generalizations of our results in Section 6.

2. USE OF GD DESIGNS IN CLUSTER SAMPLING WHEN THE CLUSTER ARE 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 $j = 1, 2, \ldots, N$; $i = 1, 2, \ldots, M$. Let

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

$$\bar{Y}_i = \sum_{j=1}^{N} y_{ij}/N\ ,\ S;\ S_i^2 = \sum_{j=1}^{N} (y_{ij} - \bar{Y}_i)^2/(N-1)\ ,\quad i=1, 2, \ldots, N \qquad (2.1)$$

Let there exist a GD design with parameters $v = MN$, b, r, $k = n$, λ_1, λ_2. Let ther population units be identified with the symbols of the design. 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 ($i = 1, 2, \ldots, M$).

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

Theorem 2.1. $\bar{y}_G$ is an unbiased estimator of $\bar{Y}$, that is, $\hat{\bar{Y}} = \bar{y}_G$.

Corollary 2.1.1. $v\bar{y}_G$ is an unbiased estimator of Y, that is, $\hat{Y} = v\bar{y}_G$.

Theorem 2.2. If $V(.)$ denotes the variance of the estimator in parenthesis, we have

$$V(\bar{y}_G) = (vrn)^{-1} \{N(N-1)\ (\lambda_2-\lambda_1) \sum_{i=1}^{M} S_i^2 + (v-1)\ (rn-v\lambda_2)\ S^2\}. \qquad (2.2)$$

Corollary 2.2.1.

$$V(v\bar{y}_G) = v(rn)^{-1} \{N(N-1)\ (\lambda_2-\lambda_1) \sum_{i=1}^{M} S_i^2 + (v-1)\ (rn-v\lambda_2)\ S^2\}. \qquad (2.3)$$

The $V(\bar{y}_G)$ given in (2.2) can be reduced by choosing the GD design to be semi-regular and in that case we will have the following results.

Theorem 2.3. If the sample is selected as one of the sets of a SRGD design, we have

$$V(\bar{y}_G) = \frac{v-n}{vMn} \sum_{i=1}^{M} S_i \qquad (2.4)$$

Clearly

Theorem 2.4. If the sample is selected as one of the sets of a SRGD and if $N/M \geq 2$, then the estimated variance $\hat{V}(\hat{\bar{y}}_G)$ of $V(\bar{y}_G)$ is

$$\hat{V}(\bar{y}_G) = \frac{v-n}{vMn} \sum_{i=1}^{M} s_i^2 \qquad (2.5)$$

Corollary 2.4.1.

$$\hat{V}(v\bar{y}_G) = \frac{v(v-n)}{Mn} \sum_{i=1}^{M} s_i^2 \qquad (2.6)$$

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 statified

sample mean and its variance under proportional allocation.

3. 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 of the study variable of the j-th unit of the i-th cluster ($5 = 1, 2, \ldots, M_i$; $i = 1, 2, \ldots, M$). Let

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

$$\bar{\bar{Y}} = Y/N \ , \ \bar{Y} = Y/M \ , \ S'^2_{wi} = \sum_{j=1}^{Mi} (y_{ij}-\bar{Y}_i)^2/(M_i-1) \ ,$$

$$S'^2_b = \sum_{i=1}^{M} (Y_i-\bar{\bar{Y}})^2/(M-1) \qquad (3.1)$$

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 $(i_1, i_2, \ldots, i_k)$ is the selected set of the BIB design, then the clusters numbered $i_1, i_2, \ldots, i_k$ enter the sample. If we determine to have a sample size n, then from i_α the cluster we randomly select without replacement $n_{i_\alpha} = \dfrac{M_{i_\alpha} n}{\sum_{\alpha=1}^{k} M_{i_\alpha}}$ units for $\alpha = 1, 2, \ldots, k$. The sampling from different clusters will be independently made.

Let

$$\bar{\bar{y}}_B = \frac{M}{k N} \sum_{\alpha=1}^{M} M_{i_\alpha} \bar{y}_{i_\alpha} \qquad (3.2)$$

where $\bar{y}_{i_\alpha}$ is the sample mean of the units of i_α cluster.

With a little algebra, analogous to the results of Section 2, we have the following results:

Theorem 3.1. $\bar{\bar{y}}_B$ is an unbiased estimator of $\bar{\bar{Y}}$.

Theorem 3.2. If B_i denotes the sum of the cluster sizes of the clusters in the i 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

$$V(\bar{y}_B) = \frac{M}{rkN^2} \sum_{i=1}^{M} M_i \left(\frac{P_i}{n} - r\right) S_{wi}'^2 + \frac{M(r-\lambda)(M-1)}{rkN^2} S_b'^2 \qquad (3.3)$$

Let $s_{wi_\alpha}'^2$ be the sample variance of the variable for the units of i_α cluster. Then

Theorem 3.3. The estimated $V(\bar{y}_B)$ is given by

$$\hat{V}(\bar{y}_B) = \frac{M}{kN^2} \sum_{\alpha=1}^{k} \frac{M_{i_\alpha}^2 (M_{i_\alpha} - n_{i_\alpha})}{M_{i_\alpha} N_{i_\alpha}} s_{wi_\alpha}'^2 + \frac{r-\lambda}{\lambda N^2} \left[\frac{M}{k} \sum_{\alpha=1}^{k} (M_{i_\alpha} \bar{y}_{i_\alpha})^2 - \frac{N^2}{M} \bar{\bar{y}}^2\right] \qquad (3.4)$$

4. USE OF L_2 DESIGNS IN CLUSTER SAMPLING. Let the population units be formed 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, the cluster represents. 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 cluster ($\alpha = 1, 2, \ldots, M_{ij}$; $i = 1, 2, \ldots, s$; $j = 1, 2, \ldots, s$). Let

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

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

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

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

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

$$(s^2-1)\, S'^2 = \sum_{i=1}^{s}\sum_{j=1}^{s} (Y_{ij} - \frac{\bar{\bar{Y}}}{s})^2. \qquad (4.1)$$

Let there exist a L_2 design with parameters $v = s^2$, b, r, k, λ_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 $n_{ij} = n\, M_{ij}/(\sum_{ij\in T} M_{ij})$. Sampling from different clusters will be independently made. Let $\bar{y}_{ij}$ be the sample mean of the selected ij-th cluster. Let $\bar{\bar{y}}_L = \frac{s^2}{kN} \sum_{ij\in T} M_{ij}\, \bar{y}_{ij}$.

Analogous to the results of previous sections we will have

Theorem 4.1. $\bar{\bar{y}}_L$ is an unbiased estimator of $\bar{\bar{Y}}$.

Theorem 4.2.

$$V(\bar{\bar{y}}_L) = \frac{b}{r^2N^2}\left[(r-2\lambda_1+\lambda_2)(s^2-1)\, S'^2 + (\lambda_1-\lambda_2)(s-1)(S'^2_{bA} + S'^2_{bB}) + \sum_{i=1}^{s}\sum_{j=1}^{s} (\frac{P_{ij}}{n} - r)\, M_{ij}\, S'^2_{wij}\right] \qquad (4.2)$$

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.

$V(\bar{\bar{y}}_L)$ can be reduced by choosing the L_2 design to satisfy $r - 2\lambda_1 + \lambda_2 = 0$. The estimate of $V(\bar{\bar{y}})$ can be easily determined through the standard techniques.

5. 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 $(.\alpha = 1, 2, \ldots, M_{ij}$; $i = 1, 2, \ldots, s$; $j = 1, 2, \ldots, t)$. Let

$$N_{i\cdot} = \sum_{j=1}^{t} M_{ij},\ N_{\cdot j} = \sum_{i=1}^{s} M_{ij},\ N = \sum_{i=1}^{s}\sum_{j=1}^{t} M_{ij},$$

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

$$Y = \sum_{i=1}^{s}\sum_{j=1}^{t} Y_{ij},\ \bar{\bar{Y}} = Y/N,\ (M_{ij}-1)\, S'^2_{wij} = \sum_{\alpha=1}^{M_{ij}} (y_{ij\alpha}-\bar{Y}_{ij})^2,$$

$$(s-1)\, S'^2_{bA} = \sum_{i=1}^{s} (Y_{i\cdot} - \frac{Y}{s})^2,\ (t-1)\, S'^2_{bB} = \sum_{j=1}^{t} (Y_{\cdot j} - \frac{Y}{t})^2,$$

$$(st-1)\, S'^2 = \sum_{i=1}^{s}\sum_{j=1}^{t} (Y_{ij} - \frac{Y}{st})^2. \qquad (5.1)$$

Let there exist a rectangular design with parameters $v = st$, b, r, k, λ_1, λ_2, and λ_3 and let the symbols of this design be identified with the st clusters. 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 5.1. $\bar{\bar{y}}_R$ is an unbiased estimator of $\bar{\bar{Y}}$.

Theorem 5.2.

$$V(\bar{\bar{y}}_R) = \frac{b}{r^2N^2}\Big[(r-\lambda_1-\lambda_2+\lambda_3)\,(st-1)\, S'^2 + (\lambda_1-\lambda_3)(t-1)\, S'^2_{bB} + (\lambda_2-\lambda_3)\,(s-1)\, S'^2_{bA} + \sum_{i=1}^{s}\sum_{j=1}^{t} (\frac{P_{ij}}{n} - r)\, M_{ij}\, S'^2_{wij}\Big] \qquad (5.2)$$

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 by choosing the rectangular design to satisfy $r-\lambda_1-\lambda_2+\lambda_3 = 0$. The estimate of $V(\bar{\bar{y}})$ can be easily obtained by text-book procedures.

6. CONCLUDING REMARKS. Hypercubic designs and extended group divisible designs (see Raghavarao (1971)) could be effectively used to develop sampling schemes when the clusters are formed based on more than 2 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 understudy and we expect to discuss these results in a subsequent paper.

REFERENCES

[1] Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. J. Indian Statist. Assoc., 1, 78-85.

[2] Mohanty (1971). Unpublished Ph.D thesis, Institute of Agricultural. Research Statistics, New Delhi.

[3] Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, India.

[4] Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. John Wiley, New York.