APPLICATIONS OF HIGHER ASSOCIATE CLASS PBIB DESIGNS IN MULTIDIMENSIONAL CLUSTER SAMPLING

Rajinder Singh[1], D. Raghavarao[2] and W. T. Federer
Punjab Agricultural University, Guelph University
and Cornell University

BU-463-M June, 197~~4~~3

ABSTRACT

This paper is a continuation of the authors' (Raghavarao, D. and Rajinder Singh [1972]) previous article "Applications of PBIB designs in cluster sampling" and aims at showing the applications of GRA, EGRA, Hypercubic, EGD, GD m-associate designs in drawing samples from finite populations.

1. INTRODUCTION

While samples can be randomly drawn from populations by using Balanced Incomplete Block (BIB) designs and certain PBIB designs, a controlled selection of the population units could be made and the estimates obtained therein are identical to the classical estimators (cf. Chakrabarti [1963], Raghavarao and R. Singh [1972]). BIB designs are also helpful in eliciting information on delicate questions from the respondents as described by Raghavarao and Federer [1973]. The use of designs is expected to throw more light on finite sampling and in this paper we discuss the applications of higher associate class Partially Balanced Incomplete Block (PBIB) designs in multidimensional cluster sampling.

[1] The work of this author was financially supported by an ICAR fellowship.

[2] On leave from Punjab Agricultural University and his work was supported by an NIH grant 5-R01-GM-05900 while he was at Cornell University.

The familiarity of design and sampling concepts on the part of the reader is assumed. We follow the notation and definitions of designs as given in Chapter 8 of Raghavarao [1971].

While one and two-dimensional cluster sampling is discussed in the literature (cf. Des Raj [1968]), the study of drawing samples from three and higher dimensional clusters in the population has not drawn attention so far. This problem can be effectively solved by using PBIB designs with higher associate classes. Due to the non-availability of material on sampling from multi-dimensional clustered populations, our methods cannot be compared with other methods. However, some comparison of our methods with known methods of deep stratification will be made in the last section of this paper.

2. SOME GENERAL RESULTS

Let the N population units be divided into v clusters. Let the i^{th} cluster contain M_i population units $(i=1,2,\cdots,v)$. Clearly, $\sum_{i=1}^{v} M_i = N$. Let Y_{it} be the response of the t^{th} unit of the i^{th} cluster. Let

$$Y_{i.} = \sum_{t=1}^{M_i} Y_{it},\ Y_{..} = \sum_{i=1}^{v} Y_{i.},\ \bar{Y}_i = Y_{i.}/M_i\ ,$$

(2.1)

$$S^2_{wi} = \sum_{t=1}^{M_i} \left(Y_{it}-\bar{Y}_i\right)^2 / \left(M_i-1\right)\ .$$

Let an m-associate class scheme exist on v symbols and let there exist a PBIB design with parameters $v,b,r,k,\ \lambda_1,\lambda_2,\cdots,\lambda_m$. We identify the symbols of the design with the v clusters. To draw a sample of size n , we proceed in two steps. In the first stage, we select a set of the PBIB design with equal probability, and if T is the selected set all the clusters i_α will be included in the sample if and only if $i_\alpha \epsilon T$. Then at the second stage, we select a simple random sample of $n_{i_\alpha} = nM_{i_\alpha} / \sum_{i_\alpha \epsilon T} M_{i_\alpha}$ units without replacement from the i_αth selected cluster. Further the sampling from different selected clusters will be independently made. Let $\bar{y}_{i_\alpha}$ be the sample mean and $s^2_{wi_\alpha}$ be the sample variance of the study variable in the i_αth selected cluster. Let

$$\hat{Y}_P = \frac{v}{k} \sum_{i_\alpha \epsilon T} M_{i_\alpha} \bar{y}_{i_\alpha} \quad . \tag{2.2}$$

The following results can then be deduced analogous to the results contained in Raghavarao and Rajinder Singh [1972]:

Theorem 2.1. *$\hat{Y}_P$ is an unbiased estimator of $Y_{..}$* .

Theorem 2.2. *If B_β denotes the sum of cluster sizes of the clusters in the βth set of the design and if P_i denotes the sum of B_β's for the sets in which the ith symbol occurs, then*

$$V\left(\hat{Y}_P\right) = \frac{v}{rk}\left\{ \sum_{i=1}^{v} \left(\frac{P_i}{n} - r\right) M_i S^2_{wi} + r \sum_{i=1}^{v} Y^2_{i.} + \left(\sum_{i=1}^{v} Y_{i.}\right)\left(\sum_{j=1}^{m} \lambda_j A_{ij}\right) - \frac{rk}{v} Y^2_{..}\right\} , \tag{2.3}$$

*where A_{ij} is the sum of Y's for all j*th *associates of the i*th *symbol, and V(·) denotes the variance of the estimator in parenthesis.*

Theorem 2.3. An unbiased estimator $\hat{V}(\hat{Y}_P)$ of $V(\hat{Y}_P)$ is

$$\hat{V}(\hat{Y}_P) = \frac{v}{k} \sum_{i_\alpha \epsilon T} \frac{M_{i_\alpha}\left(M_{i_\alpha} - n_{i_\alpha}\right)}{n_{i_\alpha}} s^2_{wi_\alpha} + \frac{v(v-k)}{k^2} \sum_{i_\alpha \epsilon T} \left(M_{i_\alpha} \bar{y}_{i_\alpha}\right)^2$$

$$- \sum_{\substack{i_\alpha, i_{\alpha'} \epsilon T \\ i_\alpha \neq i_{\alpha'}}} M_{i_\alpha} M_{i_{\alpha'}} \bar{y}_{i_\alpha} \bar{y}_{i_{\alpha'}} \left(\frac{b}{\lambda_{i_\alpha, i_{\alpha'}}} - \frac{v^2}{k^2}\right),$$

where $\lambda_{i_\alpha, i_{\alpha'}}$ denotes the number of times i_α th and $i_{\alpha'}$ th symbols occur together in the design.

3. THREE DIMENSIONAL CLUSTER SAMPLING

3.1. Three Dimensional Cluster Sampling with Unequal Number of Clusters in Different Directions—Use of GRA Designs

Let the N population units be divided into $v = N_1N_2N_3$ clusters with respect to three factors x_1, x_2, x_3 and let the clusters be denoted by $i_1, i_2, i_3)$ where $i_1 = 1, 2, \cdots, N_1$; $i_2 = 1, 2, \cdots, N_2$; $i_3 = 1, 2, \cdots, N_3$. Let the $(i_1, i_2, i_3)^{th}$ cluster contain $M_{i_1i_2i_3}$ population units, so that $\underset{i_1, i_2, i_3}{\Sigma} M_{i_1i_2i_3} = N$. Let $Y_{i_1i_2i_3t}$ be the measurement of the study variable on the t^{th} unit of the $(i_1, i_2, i_3)^{th}$ cluster for $t = 1, 2, \cdots, M_{i_1i_2i_3}$. Let

$$Y_{i_1i_2i_3} = \sum_t Y_{i_1i_2i_3t}, \quad \bar{Y}_{i_1i_2i_3} = Y_{i_1i_2i_3} / M_{i_1i_2i_3},$$

$$Y_{\cdots} = \sum_{i_1, i_2, i_3} Y_{i_1i_2i_3}, \quad \bar{Y} = Y_{\cdots}/N,$$

$$S^2_{i_1i_2i_3} = \sum_t \left(Y_{i_1i_2i_3t} - \bar{Y}_{i_1i_2i_3}\right)^2 \Big/ \left(M_{i_1i_2i_3}-1\right) ,$$

$$N_2N_3\sigma^2_1 = \sum_{i_1} Y^2_{i_1\cdot\cdot} - Y^2_{\cdot\cdot\cdot}/N_1 ,$$

$$N_1N_3\sigma^2_2 = \sum_{i_2} Y^2_{\cdot i_2\cdot} - Y^2_{\cdot\cdot\cdot}/N_2 ,$$

$$N_1N_2\sigma^2_3 = \sum_{i_3} Y^2_{\cdot\cdot i_3} - Y^2_{\cdot\cdot\cdot}/N_3 ,$$

$$N_3\sigma^2_{12} = \sum_{i_1,i_2} Y^2_{i_1i_2\cdot} - Y^2_{\cdot\cdot\cdot}/N_1N_2 ,$$

$$N_2\sigma^2_{13} = \sum_{i_1,i_3} Y^2_{i_1\cdot i_3} - Y^2_{\cdot\cdot\cdot}/N_1N_3 ,$$

$$N_1\sigma^2_{23} = \sum_{i_2,i_3} Y^2_{\cdot i_2i_3} - Y^2_{\cdot\cdot\cdot}/N_2N_3 ,$$

$$\sigma^2_{123} = \sum_{i_1,i_2,i_3} Y^2_{i_1i_2i_3} - Y^2_{\cdot\cdot\cdot}/N_1N_2N_3 ,$$

where the usual dot notation is used so that, for example, $Y_{i_1\cdot i_3} = \sum_{i_2} Y_{i_1i_2i_3}$.

Let there exist a GRA design with parameters $v = N_1N_2N_3, b, r, k, \lambda_i (i=1,2,3,4)$. By following the procedure described in Section 2, we draw our sample and form the following estimator

$$\hat{Y}_{GRA} = v \sum_{(i_1,i_2,i_3)\,\epsilon T} M_{i_1i_2i_3}\bar{y}_{i_1i_2i_3}\Big/k \enspace .$$

Applying Theorems 2.1, 2.2 and 2.3, we get the following:

Theorem 3.1. $\hat{Y}_{GRA}$ is an unbiased estimator of $Y_{\ldots}$.

Theorem 3.2. If B_α denotes the sum of cluster sizes of the clusters in the α^{th} set of the design and if $P_{i_1i_2i_3}$ denotes the sum of B_α's for the sets in which the $(i_1,i_2,i_3)^{th}$ symbol occurs, then

$$\begin{aligned} V\left(\hat{Y}_{GRA}\right) = v(rk)^{-1}\Big\{ & \sum_{i_1,i_2,i_3} M_{i_1i_2i_3}\left(\frac{P_{i_1i_2i_3}}{n} - r\right)S^2_{i_1i_2i_3} \\ & + N_3(\lambda_1-\lambda_2-\lambda_3+\lambda_4)\sigma^2_{12} + N_2N_3(\lambda_2-\lambda_4)\sigma^2_1 \\ & + N_1N_3(\lambda_3-\lambda_4)\sigma^2_2 + (r-\lambda_1)\sigma^2_{123}\Big\} \enspace . \end{aligned}$$

Theorem 3.3. If (i_1,i_2,i_3) and (i_1',i_2',i_3') occur together in $\lambda_{i_1i_2i_3,i_1'i_2'i_3'}$ sets, then an unbiased estimator $\hat{V}\left(\hat{Y}_{GRA}\right)$ of $V\left(\hat{Y}_{GRA}\right)$ is given by

$$\begin{aligned} \hat{V}\left(\hat{Y}_{GRA}\right) = vk^{-1} & \sum_{(i_1,i_2,i_3)\,\epsilon T} M_{i_1i_2i_3}\left(M_{i_1i_2i_3} - n_{i_1i_2i_3}\right)n^{-1}_{i_1i_2i_3} \\ & \times s^2_{i_1i_2i_3} + v(v-k)k^{-2} \sum_{(i_1,i_2,i_3)\,\epsilon T} \left(M_{i_1i_2i_3}\bar{y}_{i_1i_2i_3}\right)^2 \\ & - \sum_{\substack{(i_1,i_2,i_3)\,\epsilon T \\ (i_1',i_2',i_3')\,\epsilon T \\ (i_1,i_2,i_3) \neq (i_1',i_2',i_3')}} \left(M_{i_1i_2i_3}\bar{y}_{i_1i_2i_3}\right)\left(M_{i_1'i_2'i_3'}\bar{y}_{i_1'i_2'i_3'}\right) \\ & \times \left(b\,\lambda^{-1}_{i_1i_2i_3,i_1'i_2'i_3'} - v^2 k^{-2}\right) \enspace . \end{aligned}$$

3.2. Three Dimensional Cluster Sampling with Equal Number of Clusters in Different Directions—Use of Cubic Designs

We assume similar notation as Subsection 3.1. for population units and population parameters subject to the further restriction that $N_1 = N_2 = N_3 = p$.

Let there exist a cubic design with parameters $v = p^3$, b, r, k, $\lambda_i (i=1,2,3)$. Analogous to the procedure described in Section 2, we draw our sample and form the estimator

$$\hat{Y}_c = vk^{-1} \sum_{(i_1,i_2,i_3)\ \epsilon T} M_{i_1 i_2 i_3} \tilde{y}_{i_1 i_2 i_3} \quad .$$

Then the following holds:

Theorem 3.4. $\hat{Y}_c$ is an unbiased estimator of $Y_{...}$.

Theorem 3.5.

$$V\left(\hat{Y}_c\right) = p^3 (rk)^{-1} \Big\{ \sum_{i_1,i_2,i_3} M_{i_1 i_2 i_3} \left(\frac{P_{i_1 i_2 i_3}}{n} - r\right) S^2_{i_1 i_2 i_3}$$

$$+ p^2(\lambda_2 - \lambda_3)(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)$$

$$+ p(\lambda_1 - 2\lambda_2 + \lambda_3)(\sigma_{12}^2 + \sigma_{13}^2 + \sigma_{23}^2)$$

$$+ (r - 3\lambda_1 + 3\lambda_2 - \lambda_3)\sigma_{123}^2 \Big\} \quad .$$

Theorem 3.6.

$$\hat{V}\left(\hat{Y}_c\right) = vk^{-1} \sum_{(i_1,i_2,i_3)\ \epsilon T} M_{i_1 i_2 i_3} \left(M_{i_1 i_2 i_3} - n_{i_1 i_2 i_3}\right) n^{-1}_{i_1 i_2 i_3} s^2_{i_1 i_2 i_3}$$

$$+ (v-k)k^{-2} \sum_{(i_1,i_2,i_3)\ \epsilon T} \left(M_{i_1 i_2 i_3} \tilde{y}_{i_1 i_2 i_3}\right)^2$$

$$- \sum_{\substack{(i_1,i_2,i_3)\,\epsilon T\\(i_1',i_2',i_3')\,\epsilon T\\(i_1,i_2,i_3)\neq(i_1',i_2',i_3')}} \left(M_{i_1i_2i_3}\bar{y}_{i_1i_2i_3}\right)\left(M_{i_1'i_2'i_3'}\bar{y}_{i_1'i_2'i_3'}\right)$$

$$\times \left(b\,\lambda^{-1}_{i_1i_2i_3,i_1'i_2'i_3'} - v^2k^{-2}\right) .$$

4. FOUR DIMENSIONAL CLUSTER SAMPLING

In this case, when there are an unequal number of clusters in different directions, EGRA designs could be used and if the number of clusters in different directions are the same, hypercubic four-associate class designs could be used. Using the notation analogous to the one in Section 3, the estimators, their variances and the estimated variances could be deduced from Theorems 2.1, 2.2 and 2.3.

5. MULTIDIMENSIONAL CLUSTER SAMPLING

EGD designs could be used in multidimensional cluster sampling when the clusters in different directions are unequal. However, due to the large number of associate classes we do not recommend their use in 3- or 4-dimensional cluster sampling. In problems of more than 4-dimensional cluster sampling, due to the non-availability of other designs, EGD designs could be used and the estimator of population mean could be obtained and its standard error obtained in a similar way as in 3- and 4-dimensional situations. Due to the tediously involved expressions, we do not discuss these results here.

Hypercubic designs could be used in multidimensional cluster sampling when the clusters in different directions are equal in an obvious manner generalizing our result in subsection 3.2.

6. USE OF GD m-ASSOCIATE DESIGNS IN DEEP STRATIFICATION

Let the N population units be divided into $v = N_1 N_2 \cdots N_m$ strata by the method of deep stratification, based on m stratification variables. Let the $(i_1, i_2, \cdots, i_m)^{th}$ stratum contain $M_{i_1 i_2 \cdots i_m}$ population units, so that $\sum_{i_1, i_2, \cdots, i_m} M_{i_1 i_2 \cdots i_m} = N$. Let $Y_{i_1 i_2 \cdots i_m t}$ denote the measurement on the t^{th} unit of the $(i_1, i_2, \cdots, i_m)^{th}$ stratum. Let

$$Y_{i_1 i_2 \cdots i_m} = \sum_t Y_{i_1 i_2 \cdots i_m t}, \quad \bar{Y}_{i_1 i_2 \cdots i_m} = Y_{i_1 i_2 \cdots i_m} / M_{i_1 i_2 \cdots i_m}$$

$$Y = \sum_{i_1, i_2, \cdots, i_m} Y_{i_1 i_2 \cdots i_m}, \quad \bar{Y} = Y/N$$

$$S^2_{i_1 i_2 \cdots i_m} = \sum_t \left(Y_{i_1 i_2 \cdots i_m t} - \bar{Y}_{i_1 i_2 \cdots i_m}\right)^2 \Big/ \left(M_{i_1 i_2 \cdots i_m} - 1\right)$$

$$N_{\alpha+1} N_{\alpha+2} \cdots N_m \sigma^2_{12 \cdots \alpha} = \sum_{i_1, i_2, \cdots, i_\alpha} Y^2_{i_1 i_2 \cdots i_\alpha} - \frac{Y^2}{N_1 N_2 \cdots N_\alpha},$$

$$\alpha = 1, 2, \cdots, m \ .$$

Let a GD m-associate design with parameters $v = N_1 \cdots N_m$, b, r, k, λ_i $(i=1,2,\cdots,m)$ exist. Let the strata be identified with the v symbols of the GD m-associate design. Then to draw a sample of n units, we first select a set of the GD m-associate design with equal probability. If T is the selected set of the design, then the sample consists of the strata $(i_1, i_2, \cdots, i_m)$ if and only if $(i_1, i_2, \cdots, i_m) \epsilon T$. Then at the second stage, we select a simple random sample of $n_{i_1 i_2 \cdots i_m} = nM_{i_1 i_2 \cdots i_m} \Big/ \sum_{(i_1, i_2, \cdots, i_m) \epsilon T} M_{i_1 i_2 \cdots i_m}$ units without replacement from the $(i_1, i_2, \cdots, i_m)^{th}$ selected stratum. Sampling from different strata will be made independently. Let $\bar{y}_{i_1 i_2 \cdots i_m}$ be the sample mean and $s^2_{i_1 i_2 \cdots i_m}$ be the sample variance of the study variable in the $(i_1, i_2, \cdots, i_m)^{th}$ selected stratum. Let

$$\hat{Y}_{GD-m} = vk^{-1} \sum_{(i_1, i_2, \cdots, i_m)\, ET} M_{i_1 i_2 \cdots i_m} \bar{y}_{i_1 i_2 \cdots i_m} .$$

Then the following can be easily established:

<u>Theorem 6.1</u>. $\hat{Y}_{GD-m}$ <u>is an unbiased estimator of</u> Y .

<u>Theorem 6.2</u>.

$$V\left(\hat{Y}_{GD-m}\right) = v(rk)^{-1} \left\{ \sum_{i_1, i_2, \cdots, i_m} M_{i_1 i_2 \cdots i_m} \left(\frac{P_{i_1 i_2 \cdots i_m}}{n} - r \right) S^2_{i_1 i_2 \cdots i_m} \right.$$

$$+ N_m\left(\lambda_1 - \lambda_2\right)\sigma^2_{12\cdots(m-1)} + N_m N_{m-1}\left(\lambda_2 - \lambda_3\right)\sigma^2_{12\cdots(m-2)}$$

$$\left. + \cdots + N_m N_{m-1} \cdots N_2\left(\lambda_{m-1} - \lambda_m\right)\sigma^2_1 + (r - \lambda_1)\sigma^2_{12\cdots m} \right\} .$$

<u>Theorem 6.3</u>.

$$\hat{V}\left(\hat{Y}_{GD-m}\right) = vk^{-1} \sum_{(i_1, i_2, \cdots i_m)\, \epsilon T} M_{i_1 i_2 \cdots i_m} \left(M_{i_1 i_2 \cdots i_m} - n_{i_1 i_2 \cdots i_m}\right) n^{-1}_{i_1 i_2 \cdots i_m}$$

$$\times\, s^2_{i_1 i_2 \cdots i_m} + v(v-k)k^{-2} \sum_{(i_1, i_2, \cdots, i_m)\, \epsilon T} \left(M_{i_1 i_2 \cdots i_m} \bar{y}_{i_1 i_2 \cdots i_m}\right)^2$$

$$- \sum_{\substack{(i_1, i_2, \cdots, i_m)\, \epsilon T \\ (i_1', i_2', \cdots, i_m')\, \epsilon T \\ (i_1, i_2, \cdots, i_m) \neq (i_1', i_2', \cdots, i_m')}} \left(M_{i_1 i_2 \cdots i_m} \bar{y}_{i_1 i_2 \cdots i_m}\right) \left(M_{i_1' i_2' \cdots i_m'} \bar{y}_{i_1' i_2' \cdots i_m'}\right)$$

$$\times \left(b\lambda^{-1}_{i_1 i_2 \cdots i_m, i_1' i_2' \cdots i_m'} - v^2 k^{-2}\right) .$$

7. RELATIVE EFFICIENCIES

The multidimensional cluster sampling described in this paper is easy to adopt. In the absence of more than two-dimensional cluster sampling in the literature, our estimates cannot be compared with any known estimators in those situations. Even different PBIB designs could be used in a given case and comparison of different designs is a tedious job as the parameters involved and the structure of designs will be different. However,,on a set of artificial data, we compared the efficiency of the estimator given by Hansen *et al.* (See Murthy [1967]) in deep stratification and our estimator given by GD design in the following example:

Numerical Example: Let us consider a GD design with parameters v=10, b=20, r=8, k=4, M=5, N=2, $\lambda_1=0$, $\lambda_2=3$ and let the values of the study variable in different strata be as follows:

7, 8, 11, 13, 12	1, 3, 4, 6, 7, 8
12, 13, 11, 18, 15, 13	3, 1, 4, 6
15, 14, 12	6, 8, 11, 7, 10
16, 15, 9	8, 10, 2, 3
12, 9, 15, 8, 16	4, 7, 10, 14

Here the population size is 45 and let us agree to have a sample of size 8.

The required GD design is given on page 141 of Raghavarao [1971]. In the example the variance of the estimator given by Hansen *et al.* is 3774.4, while our estimator of Section 6 has variance 1550.87, thereby showing the usefulness of our sampling procedure.

ACKNOWLEDGEMENT

The authors wish to thank the referee for his helpful comments on the first draft of this manuscript.

REFERENCES

[1] Chakrabarti, M. C. (1963). "On the use of incidence matrices of designs in sampling from finite populations." Journal of the Indian Statistical Association, 1, 78-85.

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

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

[4] Raghavarao, D. and Federer, W. T. (1973). "Applications of BIB designs as an alternative to the randomized response method in survey sampling." BU-490-M of the Biometrics Unit, Cornell University.

[5] Raghavarao, D. and Singh, R. (1972). "Applications of PBIB designs in cluster sampling." Paper presented at the International Symposium on Combinatorial Mathematics and its Applications at New Delhi, India.

[6] Raj, D. (1968). Sampling Theory. McGraw-Hill.

Revised 6/74