J. Ind. Soc. Agril. Statist.
62(2), 2008 : 126-131

Application of Experimental Designs in Survey Sampling

J.N.K. Rao and K. Vijayan[1]
Carleton University, Ottawa, Canada

SUMMARY

Some early uses of experimental designs and more recent applications are reviewed. Role of experimental designs in controlled sampling is appraised. Some new work on constructing balanced bootstrap replicates from stratified random samples is also reported.

Key words: Linear programming, Controlled sample selection, Balanced sampling plans.

1 . INTRODUCTION

Experimental designs have long been used in survey sampling. More recent applications include controlled sampling, handling of sensitive questions and dentiality of survey data and the construction of balanced subsamples for variance estimation. Section 2 gives a brief account of early uses followed by more recent applications in Section 3. We also report some new work on constructing balanced bootstrap replicates from stratified samples in Section 4.

2. SOME EARLY USES

Use of experimental designs in survey sampling dates back to Frankel and Stock (1942). They used Latin square designs to increase the effective depth of stratification in the selection of primary sampling units (clusters) when the number of sample clusters is small. The primary units in the population are divided into L^2 cells and one cell is selected from each row and column using a Latin square. One cluster is then selected at random from each of the L selected cells, leading to a sample of n = L clusters. Homeyer and Black (1946) used the principle of Latin square in sampling rectangular fields of oats. Patterson (1954) studied two- and three-dimensional lattice sampling.

Mahalanobis (1946) advocated the use of interpenetrating subsamples to measure interviewer variability and to estimate total variance (response variance plus sampling variance). The sample is divided into k subsamples and k interviewers (treatments) are assigned at random to the k subsamples (plots). This is an example of a completely randomized design. Fellegi (1964) used cross-over designs to measure the components of response variance.

Chakrabarti (1963) did pioneering work in the use of balanced incomplete block designs (BIBD) for drawing samples with the same first- and second-order inclusion probabilities, π_i and π_{ij}, as simple random sampling (SRS), i.e., $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/[N(N-1)]$, where n and N denote the sample and population sizes, respectively. This approach ensures variance equivalence with SRS for the sample mean and yet leads to support size (number of samples s with probability of selection $p(s) > 0$) smaller than the support size $\binom{N}{n}$ for SRS. In the BIBD, number of treatments v = N, plot size k = n and number of blocks, b, is the support size. For example, if a symmetric BIBD exists for the desired (N, n), then b = N which is much smaller than $\binom{N}{n}$. Practical implications of Chakrabarti's result are discussed in Section 3.

3. MORE RECENT APPLICATIONS

Raghavarao and Singh (1975) extended Chakrabarti's (1963) work to more complex sampling. They applied two associate class partially balanced incomplete block designs (PBIB) to cluster sampling. Singh *et al*. (1976) extended this work to multidimensional cluster sampling by using higher

[1] *University of Western Australia, Crawley, Australia*

associate class PBIB designs. Singh and Raghavarao (1975) applied linked block designs to sampling on two occasions.

Raghavarao and Federer (1979) used BIBD and spring balance weighting designs as an alternative to the randomized response method for eliciting reliable responses from individuals on sensitive questions. A sample respondent is required to give only the total of responses to k questions, sensitive or not. Raghavarao *et al.* (1971) used spring balance weighing designs to handle measurement errors in surveys. In this application, each respondent gives the value of the study variable for each of the other (n - 1) units in the sample. For example, a sample farmer might provide more correct information about the produce of other farmers than his own. Raghavarao and Chang (1992) used BIBD and contaminated block totals to protect confidentiality of data. Lakatos and Raghavarao (1987) used undiminished residual effects designs in ordering sensitive questions. Using these designs, they estimated the residual effects of questions and then ordered the questions according to increasing size of residual effects. This approach can improve the quality of responses and increase the response rates.

Experimental designs have also been used to obtain inclusion probabilities proportional to size (IPPS) sampling designs (Gupta *et al*. 1982 and Nigam *et al.* 1984). IPPS sampling leads to efficient estimation of a finite population total Y.

It often happens in practice that certain samples, s, are known to be non-preferred (for example, the units in s may be too widespread, thus increasing the travel cost). It is desirable to minimize the probability of selecting a non-preferred sample and at the same time ensure variance equivalence to SRS or to a more general design. Controlled sampling aims to achieve this objective. Most of the literature on controlled sampling used various incomplete block designs to construct designs with minimum support size (i.e., minimum number of distinct blocks) and then identify maximum number of distinct blocks with the non-preferred samples. One of the b blocks is then selected at random and the units in it form the sample. Avadhani and Sukhatme (1973) applied BIBD to controlled sampling, but the application readily follows from Chakrabarti's (1963) results. Unfortunately, the class of BIBD's with distinct blocks do not exist for many v = N and k = n. For example, no BIB with distinct blocks exists for v = 8, k = 3 and b < $\binom{8}{3}$ = 56 blocks. To handle such cases, Wynn (1977) and Foody and Hedayat (1977) proposed BIBD's with repeated blocks and variance equivalent to SRS. Hedayat and Majumdar (1995) used the technique of trade-off. in experimental design to generate desirable sampling plans. The focus of the above papers and others on controlled sampling is on reducing the support size rather than minimizing the probability of obtaining a non-preferred sample. To implement the latter objective, even approximately, from a given incomplete block design could often involve considerable trial and error and computations. Rao and Nigam (1990) used the linear programming approach to obtain optimal controlled sampling designs. If S_1 and S denote the set of non-preferred samples and the set of all $\binom{N}{n}$ possible samples, the optimal controlled design, $p_c(s)$, is obtained by minimizing

$$\phi = \sum_{s \in S_1} p(s)$$

subject to

$$\sum_{s \ni i,\, j} p(s) = n(n-1)/\{N(N-1)\},\ 1 \le i < j < N \quad (3.1)$$

and $p(s) \ge 0$ for all $s \in S$. The condition (3.1) ensures variance equivalence to SRS. Rao and Nigam (1990) gave an example with N = 8, n = 3 for which $\phi_{min} = 0.1607$ compared to $\phi = 32/56 = 0.5714$ obtained by Foody and Hedayat (1977). Through trial and error the latter ϕ could be reduced to 24/56 = 0.4286 by interchanging the units 3 and 5 in the Foody–Hedayat plan. Foody and Hedayat (1977) alluded to mathematical programming in the context of SRS, but did not pursue that approach to construct optimal controlled plans.

The linear programming approach readily extends to IPPS sampling designs and other general sampling designs, as demonstrated by Rao and Nigam (1990, 1992). However, a difficulty with this approach is that the dimensionality of the problem increases very rapidly with increase in N and n. Lahiri and Mukerjee (2000) attempted to address the dimensionality problem. They showed how from consideration of symmetry it is possible to achieve a drastic reduction in the dimensionality of the problem. Tiwari and Nigam (1998)

applied the linear programming approach of Rao and Nigam (1990) to two-dimensional optimal controlled sampling. Their method achieves the goal of "controls beyond stratification" (Goodman and Kish 1950) as well as minimize the probability of selecting a non-preferred sample.

Hedayat *et al*. (1988) studied balanced sampling plans excluding contiguous units (BSEC) for situations when the units in the population are ordered in time or space and the contiguous units provide similar observations. For a BSEC, pairs of contiguous units do not appear together in a sample whereas all other pairs appear equally often in the samples. The first and second order probabilities are given by $\pi_i = n/N$; i = 1, ... , N and $\pi_{ij} = n(n-1)/[N(N-3)]$; $i \neq j = 1, ..., N$ and $|i-j| > 1$ and $\pi_{ij} = 0$ if $|i-j| = 1$. Hedayat *et al*. (1988) showed that the variance of the sample mean under BSEC is smaller than the variance under SRS when the units are regarded as arranged round a circle with i and (i + 1)mod N as contiguous units and the first order circular serial correlation ρ_1 is greater than $-1/(N-1)$. Stufken (1993) generalized BSEC by excluding all those pairs where distance is less than or equal to m (≥ 1). He named such sampling plans as balanced sampling plans excluding adjacent units (BSAC(m)); note that BSA(1) = BSEC. Stufken *et al*. (1999) studied a generalization of BIBD, called polygonal designs (PD), that ensures variance equivalence with BSA(m). Existence and construction of PD have been studied in the literature (see Mandal *et al*. 2008a) but very few PD are available for m > 1. Hedayat *et al*. (1988) first introduced PD for the case m = 1. Mandal *et al*. (2008a) applied the linear programming approach to obtain balanced sampling plans excluding adjacent units (BSA(m)). Mandal *et al.* (2008b) extended this work to obtain IPPS sampling designs excluding adjacent units. The optimal solution gives $\phi = \sum_{s \in S_1} p(s) = 0$, where S_1 is the set of non-preferred samples.

Experimental designs have also been used for estimating the variance of linear and nonlinear statistics computed from stratified multistage samples. For the important special case of $n_h = 2$ clusters sampled with replacement from each primary stratum h (= 1, ... ,L), McCarthy (1969) proposed the method of balanced half-samples (BHS) based on a number of half-samples formed by deleting one cluster from the sample in each stratum. The set of R balanced half-samples used may be defined by an R × L design matrix $\left(\delta_h^r\right), 1 \leq r \leq R$ and $1 \leq h \leq L$ where $\delta_h^r = +1$ or -1 depending on whether the first or the second sample cluster in the h-th stratum is in the r-th half sample and

$$\sum_r \delta_h^r \delta_{h'}^r = 0 \text{ for all } h \neq h' \qquad (3.2)$$

The estimator of a parameter, θ, is computed from each half-sample r, leading to $\hat{\theta}^{(r)}$, $r = 1, ...R$. The BHS variance estimator of the estimator $\hat{\theta}$ from the full sample is then given by

$$v_{BHS}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \hat{\theta})^2$$

The property of balance, given by (3.2), ensures that v_{BHS} agrees with the customary variance estimator in the linear case $\hat{\theta} = \hat{Y}$, where $\hat{Y}$ is the unbiased estimator of the total Y. For nonlinear statistics $\hat{\theta}$ of the form $\hat{\theta} = g(\hat{Y})$ where $g(\cdot)$ is a smooth function, v_{BHS} is approximately equal to the Taylor linearization variance estimator, but the latter requires the derivation of a separate formula for each statistic $\hat{\theta}$ unlike v_{BHS}.

A minimal set of R balanced half-samples ($L + 1 \leq R \leq L + 4$) can be constructed by using Hadamard matrices of order R = 4m. Wolter (1985) listed such matrices for values of R up to 100, and gave rules for generating these matrices.

Gurney and Jewett (1975) extended BHS variance estimation to the case $n_h = q$ (a prime) for all h, using orthogonal arrays of strength two. Gupta and Nigam (1987) extended the BHS method to the case of unequal n_h by using mixed orthogonal arrays of strength two, but a disadvantage of the extensions is that they require a much larger number of replicates, R, than in the case of $n_h = 2$ for all h. Wu (1991) developed a variance estimator for this method that agrees with the customary variance estimator in the linear case $\hat{\theta} = \hat{Y}$ and approximately equal to the Taylor linearization variance estimator for nonlinear statistics $\hat{\theta} = g(\hat{Y})$. Sitter (1993)

extended the orthogonal array method to orthogonal multi-arrays which provides greater of flexibility in construction and thus can handle unequal n_h's with fewer replicates, R. But the method cannot handle odd prime values of n_h unless one uses artificially deined units to make all n_h even.

4. BALANCED BOOTSTRAP

Efron (1979) proposed bootstrap resampling for the case of simple random sampling with replacement (i.i.d.). Rao and Wu (1988) extended the i.i.d. bootstrap to stratfied multistage sampling. However, these methods are subject to simulation error in the sense that the bootstrap variance estimator does not agree with the theoretical bootstrap variance estimator in the linear case. Graham *et al*. (1990) and Nigam and Rao (1996) used experimental designs to construct balanced bootstrap replicates yielding second-order balance, *i.e*., the variance estimator agrees with the theoretical variance estimator in the linear case.

In this section we present second-order balanced bootstrap designs for both the i.i.d. case and stratified multistage sampling with arbitrary n_h. For the i.i.d. case, we give designs with the smallest possible number of balanced bootstrap replicates, B. Technical details of the results in this section will be reported in a separate paper.

4.1 Simple Random Sampling

Suppose $\{y_1, \ldots, y_n\}$ denotes a simple random sample with mean $\bar{y}$. The theoretical bootstrap variance estimator of $\bar{y}$ reduces to $\{(n-1)/n\}\,(s_y^2/n)$ where $s_y^2 = (n-1)^{-1}\Sigma(y_i-\bar{y})^2$. Suppose we want B bootstrap replicates and let f_{bi} be the frequency count of y_i in the b-th bootstrap sample. Then the conditions for second order balance are given by

$$\Sigma_b m_{bi} = 0; B^{-1}\Sigma_b m_{bi} m_{bj} = \Delta(i-j) - \frac{1}{n} \qquad (4.1)$$

where $\Delta(k) = 0$ or 1 according as $k \neq 0$ or $k = 0$, and $m_{bi} = f_{bi} - 1$. In matrix notation, (4.1) may be written as

$$M^T M = B(I - \frac{1}{n}J) \qquad (4.2)$$

where M is a B × n matrix with elements m_{bi} and J is a matrix with all elements equal to 1.

n Even

Smallest B = 2n when n is even. In this case, a Hadamard matrix H of order 2n for almost all even values of n exists. In its standard form H has all the entries in its first row and first column equal to 1 and by rearranging the rows of H we can write

$$H = \begin{bmatrix} 1 & 1 & H_1 \\ 1 & -1 & H_2 \end{bmatrix}$$

where 1 is a n-vector of all 1's. The balanced bootstrap design M^T is given by

$$M^T = (0 \quad 0 \quad H_2)$$

which satisfies (4.2).

n Odd

We consider two possible cases: (i) n = 4m – 1; (ii) n = 4m + 1, where m is a positive integer. In case (i) we consider a skew Hadamard matrix H of order 4m with the property $H + H^T = 2I$. Skew Hadamard matrices also exist for most of the orders 4m. We choose H such that the first row of H has all entries equal to 1, and M^T is obtained by removing the first row and column of H – I as

$$H - I = \begin{bmatrix} 0 & 1^T \\ -1 & M^T \end{bmatrix}$$

We have $M^T M = (4m-1)(I - \frac{1}{n}J)$

and $B = 4m - 1$

In case (ii), we consider a conference matrix of order 4m + 2 with entries ± 1 or 0 and satisfying $H^T H = (4m + 1)I$. When such a matrix H exists, we can assume without loss of generality that the diagonal entries of H are all zero and all the entries of the first row and the first column equal to 1, except for the first entry. The matrix obtained by deleting the first row and .rst column can be used as M^T

$$H = \begin{bmatrix} 0 & 1^T \\ -1 & M^T \end{bmatrix}$$

We have

$$M^T M = (4m + 1)(I - \frac{1}{n}J)$$

and B = 4m + 1. Raghavarao (1971) has shown that conference matrices do not exist for m = 5, 8, 14, 17, 19, 23.

4.2 Stratified Multistage Sampling

Let f_{bhi} be the frequency count of the (hi)-th sample cluster in the b-th balanced bootstrap sample, b = 1, ..., B. Then the conditions for second order balance are given by Rao and Nigam (1996) as

$$\Sigma_b m_{bhi} = 0,\ B^{-1}\Sigma_b m_{bhi} m_{bhj} = \Delta(i-j) - \frac{1}{n_h} \qquad (4.3)$$

$$B^{-1}\Sigma_b m_{bhi} m_{bkj} = 0,\ h \neq k = 1, ..., L \qquad (4.4)$$

We construct an $n_h \times t_h$ matrix M_h^T for the h-th stratum as in Section 4.1 (assuming existence), where $t_h = 2n_h$ if n_h is odd and $t_h = n_h$ if n_h is even. From M_h^T we construct an $n_h \times t$ matrix N_h by adding copies of M_h^T as

$$N_h = (M_h^T ... M_h^T)$$

where t is the least common multiple of $t_1, ..., t_L$. To construct M^T, we use a Hadamard matrix H of order 4m and replace 1's in the h-th row by N_h and –1's by $–N_h$. The matrix M^T would have 4mt columns and it satisfies the second order balance conditions (4.3) and (4.4). The matrix M is B × n with B = 4mt and $n = \Sigma n_h$.

When n_h is even it is quite often possible to write M_h^T as $(L_h - L_h)$. Suppose such a decomposition is possible for all even n_h's. Then we can halve the number of samples needed for second order balance. To do this, when constructing N_h use only half the number of copies of M_h^T if n_h is odd and use L_h instead of M_h if n_h is even. The only precaution that needs to be taken is to use a standard Hadamard matrix H and arrange n_h's such that n_1 is odd. If all n_h's are even, omit the first row of H.

Example 1. Suppose L = 3 and $n_h = 3$ for all h. Then t = 3, m = 1 and B = 12.

Example 2. Suppose L = 3, $n_1 = 3$ and $n_2 = n_3 = 4$. In this case t = 24 but we use only t/2 = 12 copies and use L_h when n_h is even. We have B = 48.

ACKNOWLEDGEMENT

This paper is a slightly revised version of our paper with the same title that appeared in the book "Recent Advances in Experimental Designs and Related Topics" published by Nova Science Publisher, Inc., Huntington, New York. We wish to thank the Publishers and the Editors of the book, Stan Altman and Jagbir Singh, for granting permission to publish the paper in this Special Volume of the Journal of the Indian Society of Agricultural Statistics in honour of Dr. K. Kishen. We thank Dr. V.K. Gupta for useful suggestions.

REFERENCES

Avadhani, M.S. and Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *Internal. Statist. Rev.*, **41**, 175–182.

Chakrabarti, M.C. (1963). On the use of incidence matrices in sampling from finite populations. *J. Ind. Statist. Assoc.,* **1**, 78–85.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7,** 1–26.

Fellegi, I.P. (1964). Response variance and its estimation. *J. Amer. Statist. Assoc.,* **59**, 1016–1041.

Foody, W. and Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *J. Amer. Statist. Assoc.,* **5,** 933–945.

Frankel, L.R. and Stock, J.S. (1942). On the sample survey of unemployment. *J. Amer. Statist. Assoc.,* **37**, 77–80.

Goodman, R. and Kish, L. (1950). Controlled selection - A technique of probability sampling. *J. Amer. Statist. Assoc.,* **45**, 350–372.

Graham, R.L., Hinkley, D.V., John, P.W.M. and Shi, S. (1990). Balanced design of bootstrap simulations. *J. Roy. Statist. Soc.,* **B52,** 185–202.

Gupta, V.K., Nigam, A.K. and Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, **69,** 191–196.

Gupta, V.K. and Nigam, A.K. (1987). Mixed orthogonal arrays for variance estimation when number of primary selections is unequal in each stratum. *Biometrika*, **74,** 735–742.

Gurney, M. and Jewett, R.S. (1975). Constructing orthogonal replications for variance estimation. *J. Amer. Statist. Assoc.,* **70,** 819–821.

Hedayat, A.S., Rao, C.R. and Stufken, J. (1988). Sampling plans excluding contiguous units. *J. Statist. Plann. Inf.,* **19,** 159–170.

Hedayat, A.S. and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *J. Statist. Plann. Inf.,* **44**, 237–247.

Homeyer, P.G. and Black, C.A. (1946). Sampling replicated field experiments on oats for yield determinations. *Proceedings of the Soil Society of America*, **11,** 341–344.

Lahiri, P. and Mukerjee, R. (2000). On a simplification of the linear programming approach to controlled sampling. *Statistica Sinica*, **10,** 1171–1178.

Lakatos, E. and Raghavarao, D. (1987). Undiminished residual effect designs and their applications in ordering sensitive questions in a questionnaire. *Comm. Statist. – Theory Methods*, **16,** 1345–1359.

Mahalanobis, P.C. (1944). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Statist. Soc.,* **109,** 325– 370.

Mandal, B.N., Prasad, P. and Gupta, V.K. (2008a). Computer-aided construction of balanced sampling plans excluding contiguous units. *J. Statist. Appl.,* **3(1)**, 59-85.

Mandal, B.N., Prasad, P. and Gupta, V.K. (2008b). IPPS sampling plans excluding adjacent units. *Comm. Statist. – Theory Methods*, **37(16)**, 2532-2550.

McCarthy, P.J. (1969). Pseudo–replication: Half samples. *Internal. Statist. Rev.*, **33,** 239–264.

Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *J. Roy. Statist. Soc.,* **B46,** 564–571.

Nigam, A.K. and Rao, J.N.K. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica,* **6,** 199–214.

Patterson, H.D. (1954). The errors in lattice sampling. *J. Roy. Statist. Soc.,* **B16,** 140–149.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments*. Wiley, New York.

Raghavarao, D., Sodhi, J.S. and Singh, R. (1971). A new assumption in spring balance weighing designs leading to a sampling application. *Cal. Stat. Assoc. Bull*, **20,** 83–88.

Raghavarao, D. and Singh, R. (1975). Applications of PBIB designs in cluster sampling. *Proceedings of the Indian National Science Academy*, **41A,** 281–288.

Raghavarao, D. and Federer, W.T. (1979). Application of BIB designs as an alternative to the randomized response technique. *J. Roy. Statist. Soc.,* **B41,** 40–45.

Raghavarao, D. and Chang, C.K. (1992). Contaminated block total method enhancing access to microdata protecting confidentiality. Preprint (courtesy of the authors).

Rao, J.N.K. and Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, **77,** 807–814.

Rao, J.N.K. and Nigam, A.K. (1992). Optimal controlled sampling: A unified approach. *Internal. Statist. Rev.*, **60,** 89–98.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.,* **83,** 231–241.

Singh, R. and Raghavarao, D. (1975). Application of linked block designs in successive sampling. In : *Applied Statistics* (R.P. Gupta Ed.), 301–309. North–Holland, Amsterdam.

Singh, R., Raghavarao, D. and Federer,W.T. (1976). Applications of higher order associate class PBIB designs to multidimensional cluster sampling. *Estadistica*, **30,** 202–209.

Stufken, J. (1993). Combinatorial and statistical aspects of sampling plans to avoid the selection of adjacent units. *J. Comb. Info. Sys. Sci.*, **18,** 81–92.

Stufken, J., Song, S.Y., See, K. and Driessel, K.R. (1999). Polygonal designs: Some existence and non-existence results. *J. Statist. Plann. Inf.,* **77,** 155–166.

Tiwari, N. and Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *J. Statist. Plann. Inf.,* **69,** 89–100.