*Jour. Ind. Soc. Ag. Statistics*
*49 (Golden Jub. No.), 1996-97 : 267–276*

# Variance Estimation Using Proportional Frequency Plans

A. Dhandapani, V. K. Gupta and A. K. Nigam[1]
*Indian Agricultural Statistics Research Institute, New Delhi, India*

## SUMMARY

For estimating the variance of nonlinear statistics in large scale complex surveys, the method of balanced repeated replication has received special attention. Gupta and Nigam [4] have shown that mixed orthogonal arrays of strength two are balanced subsamples needed for variance estimation. Wu [17] has shown that the proposed method of Gupta and Nigam provides inconsistent estimator for the nonlinear statistics and has advocated the use of nearly orthogonal arrays of strength two. Wu discouraged the use of proportional frequency plans or orthogonal main effect plans for asymmetrical factorial experiments. This paper attempts to show that certain proportional frequency plans can be used effectively in variance estimation. The case of linear and non-linear statistics has been dealt separately.

*Key words :* Balanced repeated replications, Proportional frequency plans, Bias.

## 1. *Introduction*

For estimating the variance of non-linear statistics like regression coefficient, correlation coefficient, etc. in stratified, multi stage designs, the method of *Balanced Repeated Replications (BRR)* has become popular, although other procedures like *Jackknife, Linearization and Bootstrap* are also available in the literature (for details see e.g. Wolter [16]). For *two* primary selections per stratum, the *BRR* method was introduced by McCarthy ([11], [12]). This method was subsequently extended to p primary selections per stratum by Gurney and Jewett [6], where p is a prime number or power of a prime number. Gupta and Nigam [4] extended the *BRR* method to arbitrary number of primary selections per stratum. Gupta and Nigam have shown that in a sampling design with arbitrary number of primary selections per stratum, a set of *BRR* is provided by a mixed orthogonal array of strength two (Rao [14]). The mixed orthogonal arrays of strength two do not always exist for all combinations that arise from sample surveys. This puts a severe restriction on the practical utility of the method. To overcome this drawback, several alternatives like the *grouped balanced repeated replication* of Kish and Frankel [7], the use of *partial*

[1] Institute of Applied Statistics and Development Studies, Lucknow

*balanced subsamples* (Lee [9]), the use of *proportional frequency plans* (Gupta and Nigam [4]), the use of *near orthogonal arrays* (Wu [17]), etc. have been proposed in the literature. Wu [17] using a conditional analysis has shown that the grouped method of Kish and Frankel [7] is quite inefficient though computationally easy. Wu also discouraged the use of proportional frequency plans suggested by Gupta and Nigam [4] due to their large bias and inconsistency in estimating the variance of a non-linear statistic. Wu advocated the use of near orthogonal arrays of strength two as a set of subsamples in the *BRR* method whenever the mixed oithogonal arrays of strength two do not exist (see also Wang and Wu [15]). This paper attempts to verify the validity of Wu's statement regarding the use of near orthogonal arrays of strength two as well as the use of proportional frequency plans. In *Section* 2 we show that certain proportional frequency plans with little modification in the variance estimator can be used to estimate the variance of a linear statistic with no efficiency loss, i.e., balancing can be achieved using the proportional frequency plans. In *Section* 3, we also propose a variance estimator based on proportional frequency plans which is valid for variance estimation for linear as well as non-linear statistics that can be expressed as smooth functions of estimated means. The asymptotic consistency of the proposed method is also established. The corrections needed for various sampling designs are given in *Section* 4. In *Section* 5, the relative advantages of the proportional frequency plans and the near orthogonal arrays are discussed with other issues.

## 2. *Variance Estimation Using Proportional Frequency Plans : Linear Case*

To simplify exposition, firstly we restrict ourselves to the estimation of the variance of an unbiased estimator of the population mean from a stratified random sample. We suppose that the sample design consists of a simple random sample with replacement of size $n_h \geq 2$ selected from a stratum h with population size $N_h$, for $h = 1, \ldots L$. The measurement on the ith member of stratum h will be denoted by $y_{hi}$, for $i = 1, \ldots, n_h$, so that an unbiased estimator of the population mean is $\bar{y}_{st} = \sum W_h \, y_{hi} / n_h$, where $W_h = \sum N_n / N$; $N = N_1 + \ldots + N_L$. An unbiased estimator of the variance of $\bar{y}_{st}$ is given by

$$v\left(\bar{y}_{st}\right) = \sum_h P_h^2 \, n_h^{-1} \sum_i \left(y_{hi} - \bar{y}_h\right)^2, \text{ where } P_h^2 = W_h^2 / (n_h - 1)$$

The *BRR* method based on mixed orthogonal arrays of strength two provides an alternative way to obtain the same expression in a different manner. Let A be the corresponding mixed orthogonal array of strength two denoted by $(R, L, n_1 x n_2 x \ldots x n_L)$ (for definition and notations, see e.g. Gupta and Nigam [4], Rao [14]). Let i(j,h) be one of $1, \ldots, n_h$ given by (j,h)th entry of the array A and $y_{i(j,h)} = y_{hi}$ with $i = i(j, h)$. Thus, the jth replicate consists of $y_{i(j,h)}$, for $h = 1, \ldots, L$. For estimating the variance of $\bar{y}_{st}$, Gupta and Nigam proposed the estimator,

$$\tilde{v}(\bar{y}_{st}) = R^{-1} \sum_{j=1}^{R} \left(\tilde{y}_{(j)} - \tilde{y}_{(\cdot)}\right)^2, \text{ where}$$

$\tilde{y}_{(j)} = \sum_h P_h y_{i(j,h)}$ and $\tilde{y}_{(\cdot)} = R^{-1} \sum_j \tilde{y}_{(j)}$ and showed that it equals to $v(\bar{y}_{st})$.

Wu [17] showed that the above variance estimator is asymptotically inconsistent for estimating variance of a non-linear statistic which can be expressible as a function of population means. Wu proposed rescaling of $\tilde{y}_{(j)}$ by

$\bar{y}_j = \bar{y}_{st} + \sum_h P_h (\tilde{y}_{i(j,h)} - \bar{y}_h)$ and $\bar{y}_{(\cdot)} = R^{-1} \sum_j \bar{y}_j$ and the balanced repeated replication variance estimator is given by

$$v_B(\bar{y}_{st}) = R^{-1} \sum_j (\bar{y}_j - \bar{y}_{(\cdot)})^2$$

For estimating the variance of a non-linear statistic $\hat{\theta}$ that can be written as $\theta = g(\bar{y}_{st})$, where $g(\cdot)$ is a smooth function, the BRR variance estimator is given by

$$v_B(\hat{\theta}) = R^{-1} \sum_j (\hat{\theta}_j - \hat{\theta}_{(\cdot)})^2, \text{ where}$$

$$\hat{\theta}_j = g(\bar{y}_j) \text{ and } \hat{\theta}_{(\cdot)} = R^{-1} \sum_j \hat{\theta}_j$$

Consider a proportional frequency plan $A(R, L, n_1 x \ldots x n_L)$. Here A is an $R \times L$ array with entries of the jth column being from a set $\sum_j = \{1, \ldots, n_h\}$ and $n_1 x \ldots x n_L$ denote the totality of the rows from which R rows are selected. Let $f_{hi}$ denotes the number of times the ith symbol from the hth column appears in the proportional frequency plan and $f_{hh', ii'}$ denotes the number of times the pair (i,i′), from the columns (h,h′) appears together in the plan. For a proportional frequency plan, it is true that (Dey [3])

$$f_{hh',\, ii'} = \frac{f_{hi}\, f_{h'i'}}{R},\ i \in \sum\nolimits_h \text{ and } i' \in \sum\nolimits_{h'},\ h \neq h'$$

It is well known that directly using the proportional frequency plan in variance estimation leads to efficiency loss. We now make a little modification in the technique of *BRR* so as to make use of proportional frequency plans. Consider a proportional frequency plan **A** = (t : **B**), where **B** forms a mixed orthogonal array $(R, L-1, n_2 \text{ x} \ldots \text{x } n_L)$ of strength two and t is the first column of **A**. Such proportional frequency plans are usually obtained when one collapses the levels of a factor using the technique of *collapsing* of symbols given by Addelman [1]. Let

$$\delta_{hi} = \sqrt{\frac{\text{LCM}(f_{hi})}{f_{hi}}},\ i = 1, \ldots, n_h \text{ and } h = 1, \ldots, L$$

where LCM($f_{hi}$) denotes the least common multiple of $f_{hi}$, $i = 1, \ldots, n_h$, $h = 1, \ldots, L$, $f_{hi}$ being the number of times ith unit appears in the hth column of the proportional frequency plan considered. For the jth replication, define

$$\bar{y}_j = \sum_{h=1}^{L} P_h\, C_h\, \delta_{i(j,h)}\, y_{i(j,h)}$$

and

$$\bar{y}_j^* = \sum_{h=1}^{L} P_h\, C_h\, \delta_{i(j,\,h)}\, \bar{y}_h$$

where $P_h = \frac{W_h}{\sqrt{n_h - 1}}$, i(j, h) is the (j, h)th entry in the proportional frequency plan and $\delta_{i(j,h)} = \delta_{hi}$ whenever i(j, h) = i, $C_h$ 's are constants so chosen that the variance estimator based on proportional frequency plans,

$$v_p(\bar{y}_{st}) = \frac{1}{R} \sum_{j=1}^{R} (\bar{y}_j - \bar{y}_j^*)^2$$

reduces to $v(\bar{y}_{st})$.

It is straight forward to show that the variance estimator $v_p(\bar{y}_{st})$ reduces to the usual variance estimator $v(\bar{y}_{st})$ for the proportional frequency plans that can be written as **A** = (t : **B**), where **B** forms a mixed orthogonal array

$(R, L-1, n_2 \text{ x} \ldots \text{x } n_L)$ of strength two if and only if the constants $C_h$'s are chosen as

$$C_h = \sqrt{\frac{R}{n_h \, \text{LCM}\,(f_{hi})}}\,, \; h = 1, \ldots, L$$

Further, one can easily show that for any general proportional frequency plan,

$$v(\bar{y}_{st}) - v_p\,(\bar{y}_{st}) = \frac{1}{R} \sum_{h \neq k} P_h\, P_k\, C_h\, C_k \sum_{i} \sum_{i'} \delta_{hi}\, \delta_{ki'}\, (y_{hi} - \bar{y}_h)\, (y_{ki'} - \bar{y}_k)\, \frac{f_{hi}\, f_{ki'}}{R}$$

Define $$\Delta_{hi} = \frac{\delta_{hi}\, f_{hi}}{R}\,, \;\; i = 1, \ldots, n_h, \, h = 1, \ldots, L$$

If $\Delta_{hi}$ is constant for all i, within each h, then $v_p\,(\bar{y}_{st})$ reduces to $v\,(\bar{y}_{st})$. Hence, it is clear that with little modification in the *BRR* method one can use proportional frequency plans as a set of replications. There will be no efficiency loss in estimating the variance of a linear statistic, if the proportional frequency plan can be written as $\mathbf{A} = (t : \mathbf{B})$, where **B** forms a mixed orthogonal array $(R, L-1, n_2 \text{ x} \ldots \text{x } n_L)$ of strength two. The bias will be negligible if most of the $\Delta_{hi}$ 's are constant for all i. The bias is coming from the cross-product terms, which is precisely what happens in the case of near orthogonal arrays suggested by Wu [17]. Thus, there seems to be not much advantage in using near orthogonal arrays and one can instead use the proportional frequency plans in the BRR method.

## 3. *Variance Estimation Using Proportional Frequency Plans : Non-linear Case*

The variance estimator defined in the previous section based on proportional frequency plans needs rescaling of each replicate so as to get an asymptotically consistent variance estimator for a non-linear statistic that can be written as a function of sample means. Without loss of generality, let $\theta = g(\bar{Y})$, where $\bar{Y}$ is the population mean. The usual estimator of θ can be written as (Krewski and Rao, [8] )

$$\hat{\theta} = g(\bar{y}_{st})$$

Note that the bias in $\bar{y}_j$ and $\bar{y}_j^*$ is

$$\text{Bias}(\bar{y}_j) = \text{Bias}(\bar{y}_j^*) = \left[\sum \left(\frac{1}{\sqrt{(n_h - 1)}}\, C_h\, \delta_{i(j,h)} - 1\right) W_h\, \bar{r}_h\right]$$

Following on similar lines of Wu [17], for the jth replicate, define

$$\tilde{y}_j = \bar{y}_{st} + \sum_h P_h C_h \delta_{i(j,h)} (y_{i(j,h)} - \bar{y}_h)$$

$$\tilde{y}_j^* = \bar{y}_{st} = \sum_h W_h \bar{y}_h \text{ and}$$

$$\tilde{\theta}_j = g(\tilde{y}_j) \text{ and } \tilde{\theta}^*_{(\cdot)} = g(\bar{y}_{st}) \text{ and the } BRR \text{ variance estimator as}$$

$$v_p(\hat{\theta}) = \frac{1}{R} \sum_{j=1}^{R} (\tilde{\theta}_j - \tilde{\theta}^*_{(\cdot)})^2$$

Assume that

(i) $\max\left(\frac{nW_h}{n_h}\right)$ is bounded as $n = n_1 + \ldots + n_L \to \infty$ for $h = 1, \ldots, L$

(ii) $\sum W_h S_h^2$ is bounded, where $S_h^2 = \frac{1}{N_h} \sum (Y_{hi} - \bar{Y}_h)^2$, $i = 1, \ldots, N_h$ and $h = 1, \ldots L$

(iii) $\max \frac{R}{n_h f_{hi}}$ is bounded, as $n \to \infty$

Under conditions (i) - (iii), it is easy to show that

$$\text{Var}(\tilde{y}_j - \bar{y}_{st}) = O(n^{-1})$$

where Var (·) denotes the variance operator.

Thus one can expand $\hat{\theta}_j$ around $\bar{y}_{st}$ using Taylor's series expansion (see Prakasa Rao [13], Example 1.15.10, pp. 134 ) as

$$\tilde{\theta}_j = g(\bar{y}_{st}) + (\tilde{y}_j - \bar{y}_{st})^2 [g'(\bar{y}_{st})]^2 + o_p(n^{-0.5})$$

where $g'(\cdot)$ denotes the first order derivatives of the function $g(\cdot)$. Thus, it can be shown that

$$v_p(\hat{\theta}) = v_p(\bar{y}_{st}) \left\{g'(\bar{y}_{st})\right\}^2 + o_p(n^{-1})$$

Here, $v_p(\bar{y}_{st})$ is the estimate of variance of stratified mean obtained using the proportional frequency plan. However, in *Section* 2 it has been already shown that

$$v_p\,(\bar{y}_{st}) = v(\bar{y}_{st})$$

if the proportional frequency plans can be written as $\mathbf{A} = (t : \mathbf{B})$, where $\mathbf{B}$ forms a mixed orthogonal array $(R, L-1, n_2 x \ldots x n_L)$ of strength two.

$$\text{Thus,} \qquad v_p\,(\hat{\theta}) = v(\bar{y}_{st})\,[g'\,(\bar{y}_{st})]^2 + o_p\,(n^{-1})$$

$$v_p(\hat{\theta}) \;=\; v_L\,(\bar{y}_{st}) + o_p\,(n^{-1})$$

where $v_L\,(\bar{y}_{st})$ is nothing but the linearization variance estimator. Thus, $v_p(\hat{\theta})$ is asymptotically consistent since the linearization variance estimator is asymptotically consistent. Thus, for the special type of proportional frequency plans considered, one can use the proportional frequency plans in estimating the variance of non-linear statistics. For a general proportional frequency plan, the bias and inconsistency in $v_p\,(\hat{\theta})$ is due to the bias and inconsistency in $v_p\,(\bar{y}_{st})$ which is due to unequal $\Delta_{hi}$'s.

The previous results for the proposed method hold for $\theta = g(\bar{Y}_1, \ldots, \bar{Y}_k)$, where g(.) is a smooth function of a vector of population means. This class includes several parameters of interest such as the ratio, correlation and regression coefficients.

## 4. *Other Sampling Designs*

It has been assumed in the previous sections that the sampling is carried out by simple random sampling with replacement within each stratum. Consider the case in which the sampling is carried out with simple random sampling without replacement within each stratum. An unbiased estimator of variance of the stratified sample mean is given by

$$v_{srswor}\,(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{(1-f_h)\,P_h^2}{n_h} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2\,,\; P_h^2 = \frac{W_h^2}{n_h - 1}\,,\; f_h = \frac{n_h}{N_h}$$

One can obtain the above expression alternatively using proportional frequency plans. Consider a proportional frequency plan $\mathbf{A} = (t : \mathbf{B})$, where $\mathbf{B}$ forms a mixed orthogonal array $(R, L-1, n_2 \, x \ldots x \; n_L)$ of strength two. The only modification needed in the variance estimator given in *Section* 2 using proportional frequency plan for which there is no efficiency loss is that redefine $C_h$'s as

$$C_h \;=\; \sqrt{\frac{R(1-f_h)}{n_h\,\text{LCM}\,(f_{hi})}}$$

It is easy to show that the variance estimator with new $C_h$ 's reduces to the usual variance estimator. For estimating the variance of a non-linear statistic, the condition (iii) is given by,

(iii) $\dfrac{R(1-f_h)}{n_h f_{hi}}$ is bounded as $n = n_1 + \ldots + n_L \rightarrow \infty$

Consider a multi-stage stratified sampling design in which $n_h$ primary sampling units are selected with probabilities $p_{hi}$ and with replacement within each stratum independently. Subsampling is carried out independently each time a primary sampling unit is selected. An unbiased estimator of population total is given by

$$\hat{Y}_{pps} = \sum_{h=1}^{L} \hat{Y}_{pps,h} = \sum_{h=1}^{L} \left( \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{P_{hi}} \frac{1}{n_h} \right) = \sum_{h} \bar{r}_h$$

where $\bar{r}_h = \dfrac{1}{n_h} \sum\limits_{i=1}^{n_h} r_{hi}$, $r_{hi} = \dfrac{\hat{Y}_{hi}}{P_{hi}}$ and $\hat{Y}_{hi}$ is an unbiased estimator of the $i^{th}$ primary sampling unit total in the $h^{th}$ stratum. The usual estimator of variance of the unbiased estimator of population total is given by

$$v(\hat{Y}_{pps}) = \sum_{h=1}^{L} \frac{1}{n_h (n_h - 1)} \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2$$

Define $$\tilde{y}_j = \sum_{h=1}^{L} r_{i(j,h)} \delta_{i(j,h)} C_h$$

and $$\tilde{y}_j^* = \sum_{h=1}^{L} \bar{r}_h \delta_{i(j,h)} C_h$$

where $\delta_{i(j,h)}$'s are defined as in the previous sections, and $i(j,h)$ is the $(j,h)^{th}$ entry in the proportional frequency plan, $r_{i(j,h)} = r_{hi}$ and $\delta_{i(j,h)} = \delta_{hi}$ when $i = i(j,h)$.

Then the variance estimator $v_p(\hat{Y}_{pps}) = \dfrac{1}{R} \sum\limits_{j=1}^{R} (\tilde{y}_j - \tilde{y}_j^*)^2$ reduces to $v(\hat{Y}_{pps})$ if the constants $C_h$ 's are chosen as

$$C_h = \sqrt{\frac{R}{n_h (n_h - 1) \, \mathrm{LCM}\, (f_{hi})}}$$

For estimating the variance of a non-linear statistic, modify $\tilde{y}_j$ and $\tilde{y}_j^*$ as

$$\tilde{y}_j = \hat{Y}_{pps} + \sum_h (r_{i(j,h)} - \bar{r}_h) \, \delta_{i(j,h)} \, C_h$$

$$\tilde{y}_j^* = \hat{Y}_{pps}$$

and
$$v_p(\hat{\theta}) = \frac{1}{R} \sum_{j=1}^{R} \left( \tilde{\theta}_j - \tilde{\theta}_{(.)}^* \right)^2$$

where
$$\tilde{\theta}_j = g(\tilde{y}_j), \; \tilde{\theta}_{(.)}^* = g(\tilde{y}_j^*)$$

The condition (iii) in *Section* 3 is given by $\dfrac{R}{(n_h - 1) \, n_h \, f_{hi}}$ and is bounded as $n = n_1 + \ldots + n_L \to \infty$

## 5. *Discussion*

The proportional frequency plans can be effectively used to estimate the variance of a non-linear statistic. For certain type of proportional frequency plans, there is no efficiency loss in estimating the variance of a linear statistic. The near orthogonal arrays can never estimate the variance of a linear statistic with no efficiency loss. Unlike the near orthogonal arrays, the construction of proportional frequency plans are well developed in the literature of Experimental Designs. Large collection of proportional frequency plans are available [see e.g., Addelman and Kempthorne [2], Dey [3] and Gupta *et al.* [5]. Thus there seems to be not much advantage in using the near orthogonal arrays in variance estimation. Hence, the *BRR* method based on proportional frequency plans may be preferred whenever the mixed orthogonal arrays of strength two do not exist.

## ACKNOWLEDGEMENTS

The authors wish to thank the referee for his useful comments on an earlier version of the paper.

## REFERENCES

[1] Addelman, S., 1962. Orthogonal main effect plans for asymmetrical factorial experiments. *Technometrics*, **4**, 21-46.

[2] Addelman, S. and Kempthorne, O., 1961. *Orthogonal main-effects plans.* Report No. 79, Aerospace Research Laboratory, Wright - Patterson Air Force Base.

[3] Dey, A., 1985. *Orthogonal Fractional Factorial Designs.* Wiley Eastern (Halsted Press), New Delhi.

[4] Gupta, V.K. and Nigam, A.K., 1987. Mixed orthhogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, **74**, 735-742.

[5] Gupta, V.K., Dey, A. and Nigam, A.K., 1988. *Main-effect orthogonal plans-construction and tabulation.* Technical Report, Indian Agricultural Statistics Research Institute, New Delhi.

[6] Gurney, M. and Jewett, R.S., 1975. Constructing orthogonal replications for variance estimation. *J. Amer. Statist. Assoc.*, **70**, 819-821.

[7] Kish, L. and Frankel, R.M., 1970. Balanced repeated replications for standard errors. *J. Amer. Statist. Assoc.*, **65**, 1071-1094.

[8] Krewski, D. and Rao, J.N.K., 1981. Inference from stratified samples : Properties of the Linearization, Jackknife and Balanced Repeated Replication methods. *Ann. Statist.*, **9**, 1010-1019.

[9] Lee, K.H., 1972. Partially balanced designs for half-sample replication method of variance estimation. *J. Amer. Statist. Assoc.*, **67**, 324-334.

[10] Lee, K.H., 1973. Using partially balanced designs for the half-sample replication method of variance estimation. *J. Amer. Statist. Assoc.*, **68**, 612-614.

[11] McCarthy, P.J., 1966. *Replication : An approach to the analysis of data from complex surveys.* Vital and Health Statistics, Series 2, No. 14, Washington, D.C. : US Department of Health, Education and Welfare, National Centre for Health Statistics.

[12] McCarthy, P.J., 1969. Pseudoreplication half samples. *Internat. Statist. Rev.*, **37**, 239-264.

[13] Prakasa Rao, B.L.S., 1987. *Asymptotic Theory of Statistical Inference.* John Wiley and Sons, New York.

[14] Rao, C.R., 1973. Some combinatorial problems of arrays and applications to Design of Experiments. In : *A Survey of Combinatorial Theory* (J.N. Srivastava *et al.*, eds.), North-Holland, Amsterdam.

[15] Wang, J. C. and Wu, C.F.J., 1992. Nearly orthogonal arrays with mixed levels and small runs. *Technometrics*, **34**, 409-422.

[16] Wolter, K.M., 1985. *Introduction to Variance Estimation.* Springer-Verlag, New York.

[17] Wu, C.F.J., 1991. Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, **78**, 181-188.