VARIANCE ESTIMATION FROM COMPLEX SURVEYS USING BALANCED REPEATED REPLICATION

Rajender Parsad and V.K.Gupta
I.A.S.R.I., Library Avenue, New Delhi – 110 012
rajender@iasri.res.in

For estimating the variance of nonlinear statistics like regression and correlation coefficients in stratified sampling designs, the Balanced Repeated Replication (BRR) method has received special attention, although other procedures like linearization (Taylor's series expansion method), Jackknife repeated replications and the Bootstrap method are also available in the literature. BRR method involves in forming replication by choosing one of the units selected from each stratum to form a replication. Each of the replications provides an estimate of the non-linear statistic. The procedure is repeated many times to get more stable estimator.

For two primary selections per stratum, sampling with equal or unequal probabilities and with replacement, McCarthy (1966, 1969) proposed the balanced repeated replications method that involves forming half-samples by randomly selecting one primary sampling unit from the two units in each stratum and showed that using the columns of Plackett and Burman (1946) plans in two symbols (0 and 1) for re-sampling from each of the stratum, there is no loss in efficiency for linear statistics. This is illustrated with the help of an example, in the sequel:

Consider that the sample design consist of a simple random sample with replacement of size $n_h = 2$ selected from a stratum h with population size N_h, for $h = 1, \ldots, L\ (=4)$. Further let $N_1 = 5; N_2 = 6; N_3 = 8$ and $N_4 = 6$: N= $N_1 + N_2 + N_3 + N_4 = 25$. The values of the characteristic under study for the units selected from the 4 strata are respectively:

Stratum	Selected Observations	Stratum mean = $\bar{y}_h$	$W_h = n_h / N_h$	$W_h \bar{y}_h$
1	45, 66	55.5	0.20	11.10
2	33, 44	38.5	0.24	9.24
3	35, 42	38.5	0.32	12.32
4	73, 82	77.5	0.24	18.60
$\bar{y}_{st} = \sum_{h=1}^{4} W_h \bar{y}_h$ (an unbiased estimator of population mean)				51.26

It is well known that an unbiased estimator of the variance of $\bar{y}_{st}$ is

$$\hat{v}(\bar{y}_{st}) = \sum_h P_n^2 n_h^{-1} \sum_i (y_{hi} - \bar{y}_h)^2,$$

where $P_h^2 = W_h^2 / (n_h - 1)$ and $\bar{y}_h = \sum_i y_{hi} / n_h$.

The computations involved in the estimation of an unbiased estimator of the variance of $\bar{y}_{st}$ are given in the sequel.

Stratum	Selected Obs.	Strata mean= $\bar{y}_h$	$\sum_i (y_{hi} - \bar{y}_h)^2$	W_h	P_h^2 / n_h	(4)*(6)
(1)	(2)	(3)	(4)	(5)	(6)	
1	45, 66	55.5	220.5	0.20	0.0200	4.41
2	33, 44	38.5	60.5	0.24	0.0288	1.7424
3	35, 42	38.5	24.5	0.32	0.0512	1.2544
4	73, 82	77.5	40.5	0.24	0.0288	1.1664
$\hat{v}(\bar{y}_{st}) = \sum_h P_n^2 n_h^{-1} \sum_i (y_{hi} - \bar{y}_h)^2$						8.5732

Now consider a Plackett and Burman Plan for 2^4 in 8 runs:

0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

Using the above, get 8 repeated samples, one corresponding to each run or row of the array. If the symbol in the column *h* is 0, then select unit 1 from stratum *h* and unit 2 for 1 in stratum *h.* Thus the 8 repeated samples are given as

45	45	45	45	66	66	66	66
33	33	44	44	33	33	44	44
35	42	35	42	35	42	35	42
73	82	82	73	82	73	73	82

Each of the sub-sample means are obtained as $\bar{y}_j = \sum_h P_h y_{i_{jh}}$. Single observation in stratum *h* for sample *j* will be denoted by where $y_{i_{jh}}$, where i_{jh} is one of $1, \cdots, n_h$. Here $P_h = W_h$ as $n_h = 2$. The various computations involved are shown below:

	←Repeated Samples→							
	R1	R2	R3	R4	R5	R6	R7	R8
$P_1 y_{i_{j1}}$	9	9	9	9	13.2	13.2	13.2	13.2
$P_2 y_{i_{j2}}$	7.92	7.92	10.56	10.56	7.92	7.92	10.56	10.56
$P_3 y_{i_{j3}}$	11.2	13.44	11.2	13.44	11.2	13.44	11.2	13.44

$P_4\, y_{i_{j4}}$	17.52	19.68	19.68	17.52	19.68	17.52	17.52	19.68
$\bar{y}_j$	45.64	50.04	50.44	50.52	52	52.08	52.48	56.88

Now $\bar{y} = \frac{1}{8}\sum_{j=1}^{8} \bar{y}_j$ =51.26 and Var($\bar{y}_j$)=8.5732.

Therefore, the results of repeated replications give unbiased estimator for population mean with same variance, i.e., without any loss in efficiency. The set of replications that achieve the full precision is called a *Balanced Set* and the method, therefore, is termed as *Balanced Repeated Replications.* Since the method consists of selecting one of the two units, the number of units selected for each replication is exactly one half the total sample size. Hence the nomenclature Balanced Half sample method is commonly used to describe the McCarthy's case of BRR. The plackett and Burman plans are nothing but Hadamard matrices or orthogonal arrays of strength two in two symbols. A brief description of orthogonal arrays is given in the Appendix.

Gurney and Jewett (1975) used orthogonal arrays to extend this to a prime number *p* of primary selections from each stratum, by forming a set of balanced sub samples. Gupta and Nigam (1987, *Biometrika, 74(4), 735-742)* extended the method of balanced repeated replications to arbitrary number of primary selections per stratum designs. It is shown that mixed orthogonal arrays of strength two are balanced subsamples needed for variance estimation. The algebra of the use of mixed orthogonal arrays in variance estimation is given in the sequel.

To simplify exposition attention will be confined to the estimation of a population mean from a stratified random sample. Extension to more complicated situations can be handled by using linear approximations. We suppose the sample design to consist 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, ... , L*. The measurement on the j^{th} member of stratum *h* will be denoted by y_{hi}, for $i = 1, \cdots, n_h$, so that an unbiased estimator of the population mean is

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h y_{hj/n_h} \text{ , where } W_h = N_h / N,\ N = \left(N_1 + \cdots + N_L\right).$$

An unbiased estimator of the variance of $\bar{y}_{st}$ is given by

$$\hat{v}\left(\bar{y}_{st}\right) = \sum_h P_n^2 n_h^{-1} \sum_i \left(y_{hi} - \bar{y}_h\right)^2,$$

where $P_n^2 = W_h^2 / \left(n_h - 1\right)$ and $\bar{y}_h = \sum_i y_{hi/n_h}$.

The same estimator of $\text{var}\left(\bar{y}_{st}\right)$ may be calculated by the use of balanced subsamples with observation per stratum constructed by the use of mixed orthogonal arrays of strength two. In effect, these mixed orthogonal arrays of strength two define balanced subsamples, so that the

existence of such an array implies the existence of balanced subsamples with the requisite properties.

To see this define a set of R subsamples with one observation per stratum. The single observation in stratum h for sample j will be denoted by $y_{i_{jh}}$ where i_{jh} is one of $1, \cdots, n_h$. We can equally write $y_{i_{jh}} = \sum_i \delta(i, i_{jh}) y_{hi}$, where $\delta(k, l) = 1$ if $k = l$ and $\delta(k, l) = 0$ other wise. For the j^{th} subsample let $\bar{y}_j = \sum_h P_h y_{i_{jh}}$. Then the average of these terms may be written as

$$\bar{y} = R^{-1} \sum_{h=1}^{L} P_h \sum_{i=1}^{n_h} y_{hi} \sum_{j=1}^{R} \delta(i, i_{jh}),$$

where $\sum_j \delta(i, i_{jh})$ represents the number of times that unit i from stratum h occurs over all R subsamples. If this is constant for all i within h, $\sum_j \delta(i, i_{jh}) = \mu_h$, then, as $\sum_j \sum_i \delta(i, i_{jh}) = R$, we have that $R = \mu_h n_h$ or $\mu_h = R n_h^{-1}$, whence $\bar{y} = \sum_h P_h \bar{y}_h$.

Consider now $$\sum_{j=1}^{R} (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{R} \bar{y}_j^2 - R_y^{-2} = S + T - R_y^{-2}$$

with $$S = \sum_{j=1}^{R} \sum_{h=1}^{L} P_h^2 \left\{ \sum_{i=1}^{n_h} \delta(i, i_{jh}) y_{hi} \right\}^2, \quad T = \sum_{j=1}^{R} \sum_{h \neq h'}^{L} P_h P_{h'} \sum_{i=1}^{n_h} \sum_{i'=1}^{n_{h'}} y_{hi} y_{h'i'} \delta(i, i_{jh}) \delta(i', i_{jh'}).$$

As only one unit is selected from each stratum in each subsample, it is readily deducted that $S = R \sum_h P_h^2 \sum_i y_{hi}^2 n_h^{-1}$. In summation T, the term $\sum_j \delta(i, i_{jh}) \delta(i', i_{jh'})$ represents the number of times that unit i from stratum h appear in same subsample as unit i' from stratum h'. If this is constant, say $\mu_{hh'}$ for all pairs (i, i') from strata (h, h'), then, as

$$\sum_{j=1}^{R} \sum_{i=1}^{n_h} \sum_{i'=1}^{n_{h'}} \delta(i, i_{jh}) \delta(i', i_{jh'}) = R = n_h n_{h'} \mu_{hh'},$$

we have $\mu_{hh'} = R / (n_h n_{h'})$. This allows us to write

$$T = R \sum_{h \neq h'} P_h P_{h'} \bar{y}_h \bar{y}_{h'},$$

whence $$S + T = R \sum_{h=1}^{L} P_h^2 n_h^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 + R_y^{-2}$$

Thus $R^{-1} \sum_{j=1}^{R} (\bar{y}_j - \bar{y})^2 = \sum_h P_n^2 n_h^{-1} \sum_i (y_{hi} - \bar{y}_h)^2 = \hat{v}(\bar{y}_{st})$ as required.

Thus, a clear relationship exists between the mixed orthogonal arrays of strength two and the set of balanced repeated replications. The method of Balanced Repeated Replication is applicable

whenever the mixed orthogonal array of strength two exists. However, the mixed orthogonal arrays of strength two do not always exist for all the combinations of symbols. This puts a severe restriction on the application of BRR method in real sample survey situations. Kish and Frankel (1970, 1974) introduced a method called grouping method to overcome the obstacle of non-availability of a balanced set in the case of arbitrary number of primary selections. The method is to divide the sample sizes within each stratum at random into two equal subsamples, and then to form balanced repeated replications treating each subsample as if it is one member of a sample of size 2. This method is commonly called as grouped balanced half sample method or grouped balanced repeated replication. In fact, the repeated replications based on this method are partially balanced. Wu (1991) have shown that the grouping method is very inefficient to estimate the variance of a non-linear statistic. Gupta and Nigam (1987) advocated the use of orthogonal main effect plans with unequal frequencies in general and with proportional frequencies in particular. Wu (1991) discouraged the use of proportional frequency plans in the BRR as these result in efficiency loss in estimating the variance of a non-linear statistic. Wu dvocated the use of near orthogonal arrays in which most of the columns are orthogonal. Dhandapani (1996) investigated the validity of Wu's statement regarding the use of proportional frequency plans. He has shown that certain proportional frequency plans can be used with no loss in efficiency for estimating the variance of a linear statistic. Further, they have also shown that the variance estimate using proportional frequency plans is asymptotically consistent for a non-linear statistic that is expressible as a general smooth function of population means.

References

Dhandapani, A. (1996). *Variance estimation from complex survey data using proportional frequency plans*. Unpublished Ph.D. Thesis, I.A.R.I., New Delhi.

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

Gurney, M. and Jewett, R.S. (1975). Constructing orthogonal replications for variance estimation. *J. Am. Statist. Assoc.*, **71**, 819-821.

Kish, L. and Frankel, R.M. (1970). Balanced repeated replications for standard errors. *J. Am. Statist. Assoc.*, **65,** 1071-1094.

Kish, L. and Frankel, R.M. (1974). Inference from complex samples (with discussion).. *J. Royal. Statist. Soc.*, **B36,** 1-37..

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.

McCarthy, P.J. (1969). Psuedo replication: half samples. *Rev. Int. Statist. Inst*., **37**, 239-264.

Plackett, R.L. and Burman, J.P. (1946). The design of optimum multi-factor experiments. *Biometrika*, **33**, 328-332.

Wu, C.F.. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, **78**, 181-188.

Appendix

Orthogonal arrays

Let S be the set of s symbols (or levels). These s levels are denoted by 0, 1, . . ., s-1.

Definition 1: An N × k Array **A** with entries from S is said to be an OA with s levels, strength t and index λ ($0 \leq t \leq k$) if every N × t subarray of **A** contains each t-tuple based on S exactly λ times as row.

The orthogonal arrays are denoted by OA (N, k, s, t) or OA (N, s^k, t) with N, k, s and t as the parameters, where N is the size of the array, or the number of runs, or the number of treatment combinations, k is the number of constraints or the number of factors, s is the number of symbols or levels and t is the strength of the orthogonal array.

Example 1: Let us consider an array

0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

In this array if we see any three columns, we find that all possible combinations 000, 001, 010, 011, 100, 101, 110, & 111 of symbols 0 & 1 appears same number of times, i.e. one. This property of the above array makes it to be called orthogonal arrays (OA) of strength 3 and index unity.

When all the factors in an orthogonal array are not at same level, orthogonal arrays are denoted by OA (N, $s_1 s_2 \ldots s_k$, t), where first factor is at the level s_1, second factor is at the level s_2 and so on.

Definition 2: A mixed orthogonal array $OA(N, s^{k_1} s^{k_2} \ldots s^{k_v}, t)$ is an array of size N× k, where k = $k_1+k_2+ \ldots +k_v$ is the total number of factors, in which the first k_1 columns have symbols from {0, 1, . . ., s_1-1}, the next k_2 columns have symbols from {0, 1, . . . , s_2-1}, and so on, with the property that in any N × t subarray every possible t-tuple occurs an equal number of times as arrow.

Example 2: An array OA (8, 4.2^4, 2). In this array the first factor is at four levels and the factors 2 to 5 are at two levels each.

0	0	0	0	1
0	1	1	1	0
1	0	0	1	0
1	1	1	0	1
2	0	1	0	0
2	1	0	1	1
3	0	1	1	1
3	1	0	0	0