THE ESTIMATION OF VARIANCE COMPONENTS USING BALANCED REPEATED REPLICATION

Robert J. Casady, National Center for Health Statistics

1. Introduction

The balanced repeated replication (BRR) technique as first proposed by McCarthy (1966) and later more fully developed by Kish and Frankel (1970) has been widely adopted for estimating variances for statistics from complex sample surveys. Unfortunately, this technique has not permitted the estimation of the components of variation inherent in a multistage sampling plan. Motivated by the analytical study of a linear first order estimator from a multistage design a BRR technique is developed to estimate these components of variation. An emperical study and a discussion of the proposed technique in the case of both linear and non-linear estimators is also included.

2. The Design and First Order Estimator

Suppose we have a two stage sampling design where x_{ijk} represents the measurement of interest on the k^{th} element from the j^{th} first stage unit in the i^{th} stratum where i=1, 2, ..., L, j=1, 2, ..., N_i and k=1, 2, ..., M_{ij}. Further, let f_i and f_{ij} be the first and second stage sampling fraction respectively and let n_i and m_{ij} be the first and second stage sample sizes respectively. X' is the usual inflation estimator of the population aggrigate X.

Assuming simple random sampling without replacement at both levels it can be shown that

$$\mathrm{Var}(X') = \sum_{i=1}^{L} N_i^2 (1-f_i)\, \sigma_i^2 / n_i + \sum_{i=1}^{L} \sum_{j=1}^{N_i} N_i\, M_{ij}^2\, (1-f_{ij})\, \sigma_{ij}^2 / n_i\, m_{ij} \qquad (1)$$

where $\sigma_{ij}^2 = \sum_{k=1}^{M_{ij}} (X_{ijk} - \bar{X}_{ij.})^2/(M_{ij}-1)$

$$\sigma_i^2 = \sum_{j=1}^{N} M_{ij}^2\, (\bar{X}_{ij.} - \bar{X}_{i..})^2/N_i-1)$$

$$\bar{X}_{ij.} = \sum_{k=1}^{M_{ij}} X_{ijk}/M_{ij}$$

and $\bar{X}_{i..} = \left(\sum_{j=1}^{N_i} \bar{X}_{ij.} \right)/N_i$.

If the sampling is done with replacement then the terms f_i and f_{ij} are set equal to zero in equation (1).

3. The BRR Second Order Estimators

3.1 The Usual BRR Estimator of Variance

Let us assume the stratified two stage design described in the previous section and in addition assume that n_i and m_{ij} are even for all i=1, 2, ..., L and j=1, 2, ..., N_i so as to avoid certain tedious but unconsequential algebraic complications. One procedure suggested by Kish and Frankel (1970) for BRR variance estimation was to partition the sample population into two groups by randomly assigning each of the first stage units in each stratum to one of two equal sized groups. Pseudo-replicates are formed by selecting one of the two groups from each of the strata and then each of these pseudo-replicates is used to calculate a numerical value for the first order estimator. Using the above procedure it can be shown that the expectation of the BRR estimator S_{hs}^2 of the variance of the simple inflation estimator X' is given by

$$E(S_{hs}^2) = \sum_{i=1}^{L} N_i^2\, \sigma_i^2/n_i + \sum_{i=1}^{L} \sum_{j=1}^{N_i} N_i M_{ij}^2\, (1-f_{ij})\, \sigma_{ij}^2/n_i m_{ij} \qquad (2)$$

when the sampling is without replacement. When the sampling is with replacement the f_{ij} are set equal to zero in equation (2).

It should be noted that when the half-sample estimates are calculated as described above the variation among them is a reflection of both the first and second stage sources of variation in Var (X'), however, they are not in the correct proportion when the sampling is with replacement. Even if one makes simplifying assumptions as in Frankel (1971) it is not possible at this stage to modify the BRR estimator to give a unbiased estimate of Var (X'). Thus if one is interested in such areas as optimum allocation, correcting the usual BRR estimator for bias, etc. it is necessary that an estimator of each of the components of variation be available. In what follows the basic ideas of the BRR method are modified to provide estimates of the components of variation.

3.2 BRR Estimates of Variance Components

First, each of the sampled first stage units is to be considered a pseudo-stratum thus resulting in $\sum_{i=1}^{L} n_i$ pseudo-strata. Secondly, within each of the pseudo-strata the second stage units are randomly placed in one of two equal sized groups. By selecting one of the two groups from each of the pseudo-strata the sample elements are

partitioned for use in a half-sample estimate. Observe that differences between half-sample estimates reflect only second stage sampling. Denoting the BRR variance estimator formed by using a balanced set of these half samples as

$S^2_{hs,2}$ it can be shown that

$$E\,(S^2_{hs,2}) = \sum_{i=1}^{L} \sum_{j=1}^{N_i} N_i M^2_{ij}\, \sigma^2_{ij}/n_i m_{ij} \qquad (3)$$

Thus, in the case of sampling with replacement this BRR method can be used to provide an estimate of the component of variation due to second stage sampling and in the case of linear first order estimator the estimator will be unbiased. Further, the estimator

$S^2_{hs,1} = S^2_{hs} - S^2_{hs,2}$ provides an estimator for the first-stage component of variance which is also unbiased for linear first order estimators.

When sampling without replacement it is necessary to make simplifying assumptions to get unbiased BRR variance estimators even when dealing with linear first order estimators. Extending Frankel's (1971) assumptions for a one stage design to the two stage design we assume that $f_i = f_1$ and $f_{ij} = f_2$ for all i and j. It then follows from equations (1), (2) and (3) that for linear estimators

$(1-f_2)S^2_{hs,2}$, $S^2_{hs,1} = (1-f_1)(S^2_{hs}-(1-f_2)S^2_{hs,2})$,

and $S^2_{hs,1} + (1-f_2)\, S^2_{hs,2}$ are unbiased estimators of the second stage component, the first stage component and total variance respectively.

4. A Numerical Example

The following example shows how the method outlined in the previous section can be used to calculate the variance components and the total variance in the case of a none linear estimator.

Suppose we have data from a survey where there are 3 strata with 2 first stage sample units per stratum and 2 second stage units per first stage unit. The sample units are denoted by I_{ijk} which represents the k^{th} second stage unit within the j^{th} first stage unit within the i^{th} stratum. Assume that

$f_i = .5$ for i-1, 2, 3

and $f_{ij} = .1$ for i=1, 2, 3 and j=1, 2.

Now suppose that two characteristics are measured on each sample unit, namely characteristic X and characteristic Y. We will let x_{ijk} and y_{ijk} represent the measurements of X and Y respectively on I_{ijk}. The combined ratio estimator

$$\hat{R} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2} x_{ijk} \Big/ \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2} y_{ijk}$$

will be used to estimate R= E(X)/E(Y). The (hypothetical) data is summarized below

$x_{111} = 10$	$y_{111} = 30$
$x_{112} = 12$	$y_{112} = 38$
$x_{121} = 21$	$y_{121} = 50$
$x_{122} = 17$	$y_{122} = 52$
$x_{211} = 8$	$y_{211} = 16$
$x_{212} = 11$	$y_{212} = 20$
$x_{221} = 9$	$y_{221} = 30$
$x_{222} = 12$	$y_{222} = 48$
$x_{311} = 14$	$y_{311} = 50$
$x_{312} = 20$	$y_{312} = 72$
$x_{321} = 20$	$y_{321} = 42$
$x_{322} = 24$	$y_{322} = 49$

It can be verified that $\hat{R} = 2.7921$ and to estimate Var $(\hat{R})$ we must calculate S^2_{hs} and $S^2_{hs,2}$.

First to calculate S^2_{hs} the sample population is grouped as indicated below

Strata	Group I	Group II
1	$(I_{111}\ I_{112})$	$(I_{121}\ I_{122})$
2	$(I_{211}\ I_{212})$	$(I_{221}\ I_{222})$
3	$(I_{311}\ I_{312})$	$(I_{321}\ I_{322})$

The selection matrix

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

is then used to form the four half samples which produce the $\hat{R}_1 = 3.0133$, $\hat{R}_2 = 2.2673$, $\hat{R}_3 = 2.7241$ and $R_4 = 3.2473$. Using these half sample estimates we find that

$$S^2_{hs} = \sum_{i=1}^{4} (\hat{R}_i - \hat{R})^2/4 = .1340.$$

Now to calculate $S^2_{hs,2}$ the data is grouped as indicated below

Pseudo-Strata	Group I	Group II
1	I_{111}	I_{112}
2	I_{121}	I_{122}
3	I_{211}	I_{212}
4	I_{221}	I_{222}
5	I_{311}	I_{312}
6	I_{321}	I_{321}

and then using the selection matrix

$$\begin{pmatrix} 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

we form eight half samples which produce the estimates

$\hat{R}_1 = 2.6585$, $\hat{R}_2 = 2.8824$, $\hat{R}_3 = 2.8556$,
$\hat{R}_4 = 2.8602$, $\hat{R}_5 = 2.7449$, $\hat{R}_6 = 2.8276$,
$\hat{R}_7 = 2.9070$, $\hat{R}_8 = 2.6044$

Using these half-sample estimates we find that

$$S^2_{hs,2} = \sum_{i=1}^{8} (\hat{R}_i - \hat{R})^2/8 = .0108.$$

Thus the estimate for the component of variance due to second stage sampling would be

$$(1-f_2)\ S^2_{hs,2} = .9(.0108) = .0097).$$

The estimate for the first stage component of variance is

$$(1-f_1)\ (S^2_{hs} - (1-f_2)\ S^2_{hs,2}) = .0622$$

Then adding the estimates for the two components together we have

$\widehat{Var}\ (\hat{R}) = .0622 + .0097 - .0719.$

It is of interest to note that prior to this paper S^2_{hs}, which is 86% larger than $\widehat{Var}\ (\hat{R})$, would have been used to estimate Var $(\hat{R})$.

5. The Monte Carlo Study

5.1 The Populations, Sample Design and First Order Statistics

For the Monte Carlo study of the BRR variance component estimators three populations were generated on the computer. Each of the three populations consisted of 10,000 individuals who were grouped into 10 equal sized strata. Each stratum was in turn partitioned into 10 groups of 100 individuals each. These groups were the first stage sampling units and the individuals within the groups were the second stage sampling units.

Measurements of two characteristics were available for each individual in the population. The numeric values for the two measurements were denoted by x_{ijk} and y_{ijk} respectively for the k^{th} individual in the j^{th} group in the i^{th} stratum. For this study x_{ijk} and y_{ijk} were realizations of the random variables.

$$X_{ijk} = \mu_{ijk} + (C_{ijk} - 5.0)/\sqrt{2.5}$$

and $$Y_{ijk} = R\ X_{ijk} + \varepsilon_{ijk}$$

where $C_{ijk} \sim \chi^2\ (5)$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

and C_{ijk} and ε_{ijk} were independent for all i, j and k. The values of the μ_{ij}'s may be found in Table 1. The parameter R was equal to 2 in all of the populations and for Population I, $\sigma = 2$; for Population II, $\sigma = 100$ and for Population III, $\sigma = 17.3205$. These values were chosen so the correlation between x_{ijk} and y_{ijk} would be .99, .1 and .5 respectively.

The sample design was a stratified two stage design using simple random sampling without replacement at both levels. Two first stage units were drawn from each stratum and four second stage units were selected from each first stage unit. From the sample three first order statistics were calculated. They were the usual inflation estimator of the aggregate of the x-measurements

$$x' = \frac{10{,}000}{80} \sum_{i=1}^{10} \sum_{j=1}^{2} \sum_{k=1}^{4} x_{ijk},$$

the usual inflation estimator of the aggregate of the y-measurements

$$y' = \frac{10{,}000}{80} \sum_{i=1}^{10} \sum_{j=1}^{2} \sum_{k=1}^{4} y_{ijk}$$

and the combined ratio estimator of the aggregates

$$\hat{R} = y' / x'.$$

5.2 Second Order Estimators

The variance of each of the first order estimators and the covariance of x' and y' were calculated by the following methods:

Table 1 - The values of the parameter μ_{ij}

	STRATA									
1st STAGE UNITS	1	2	3	4	5	6	7	8	9	10
1.	35.5	40.5	45.5	50.5	55.5	60.5	65.5	70.5	75.5	80.5
2.	36.5	41.5	46.5	51.5	56.5	61.5	66.5	71.5	76.5	81.5
3.	37.5	42.5	47.5	52.5	57.5	62.5	67.5	72.5	77.5	82.5
4.	38.5	43.5	48.5	53.5	58.5	63.5	68.5	73.5	78.5	83.5
5.	39.5	44.5	49.5	54.5	59.5	64.5	69.5	74.5	79.5	84.5
6.	40.5	45.5	50.5	55.5	60.5	65.5	70.5	75.5	80.5	85.5
7.	41.5	46.5	51.5	56.5	61.5	66.5	71.5	76.5	81.5	86.5
8.	42.5	47.5	52.5	57.5	62.5	67.5	72.5	77.5	82.5	87.5
9.	43.5	48.5	53.5	58.5	63.5	68.5	73.5	78.5	83.5	88.5
10.	44.5	49.5	54.5	59.5	64.5	69.5	74.5	79.5	84.5	89.5

Table 2 - A comparison of the variance estimators for Population I when the design is stratified two stage with random sampling without replacement at both stages

	Emperical Expectations				Emperical Variances	
Parameter	True Value	Standard Estimator	Uncorrected Half-sample Estimator	1st and 2nd Stage Corrected Half-sample Estimator	Standard Estimator	1st and 2nd Stage Corrected Half-sample Estimator
Variance (X')	$.66031 \times 10^{8}$	$.64864 \times 10^{8}$	$.73961 \times 10^{8}$	$.64862 \times 10^{8}$	$.56946 \times 10^{15}$	$.56925 \times 10^{15}$
Between Component	$.37472 \times 10^{8}$	$.36392 \times 10^{8}$	-	$.36399 \times 10^{8}$	$.56722 \times 10^{15}$	$.61161 \times 10^{15}$
Within Component	$.28559 \times 10^{8}$	$.28472 \times 10^{8}$	-	$.28463 \times 10^{8}$	$.04346 \times 10^{15}$	$.09788 \times 10^{15}$
Variance (Y')	$.26758 \times 10^{9}$	$.26287 \times 10^{9}$	$.29901 \times 10^{9}$	$.26286 \times 10^{9}$	$.92198 \times 10^{16}$	$.92177 \times 10^{16}$
Between Component	$.14911 \times 10^{9}$	$.14457 \times 10^{9}$	-	$.14461 \times 10^{9}$	$.92326 \times 10^{16}$	$.99753 \times 10^{16}$
Within Component	$.11847 \times 10^{9}$	$.11830 \times 10^{9}$	-	$.11825 \times 10^{9}$	$.07172 \times 10^{16}$	$.16350 \times 10^{16}$
Variance $(\hat{R})$	$.12726 \times 10^{-4}$	$.12969 \times 10^{-4}$	$.13134 \times 10^{-4}$	$.12965 \times 10^{-4}$	$.21430 \times 10^{-10}$	$.21904 \times 10^{-10}$
Between Component	$.00423 \times 10^{-4}$	$.00609 \times 10^{-4}$	-	$.00675 \times 10^{-4}$	$.23547 \times 10^{-10}$	$.31291 \times 10^{-10}$
Within Component	$.12303 \times 10^{-4}$	$.12361 \times 10^{-4}$	-	$.12290 \times 10^{-4}$	$.04631 \times 10^{-10}$	$.16357 \times 10^{-10}$
Covariance (X',Y')	$.13168 \times 10^{9}$	$.12932 \times 10^{9}$	$.14743 \times 10^{9}$	$.12931 \times 10^{9}$	$.22700 \times 10^{16}$	$.22689 \times 10^{16}$
Between Component	$.07471 \times 10^{9}$	$.07247 \times 10^{9}$	-	$.07249 \times 10^{9}$	$.22652 \times 10^{16}$	$.24437 \times 10^{16}$
Within Component	$.05698 \times 10^{9}$	$.05684 \times 10^{9}$	-	$.05683 \times 10^{9}$	$.01743 \times 10^{16}$	$.03919 \times 10^{16}$

Table 3 - A comparison of the variance estimators for Population II when the design is stratified two stage with random sampling without replacement at both stages

	Emperical Expectations				Emperical Variances	
Parameter	True Value	Standard Estimator	Uncorrected Half-sample Estimator	1st and 2nd Stage Corrected Half-sample Estimator	Standard Estimator	1st and 2nd Stage Corrected Half-sample Estimator
Variance (X')	$.67707 \times 10^{8}$	$.67733 \times 10^{8}$	$.77526 \times 10^{8}$	$.67748 \times 10^{8}$	$.60799 \times 10^{15}$	$.61002 \times 10^{15}$
Between Component	$.39094 \times 10^{8}$	$.39175 \times 10^{8}$	-	$.39113 \times 10^{8}$	$.62000 \times 10^{15}$	$.66765 \times 10^{15}$
Within Component	$.28613 \times 10^{8}$	$.28558 \times 10^{8}$	-	$.28635 \times 10^{8}$	$.04036 \times 10^{15}$	$.11004 \times 10^{15}$
Variance (Y')	$.13176 \times 10^{11}$	$.13323 \times 10^{11}$	$.13441 \times 10^{11}$	$.13329 \times 10^{11}$	$.23914 \times 10^{20}$	$.24312 \times 10^{20}$
Between Component	$.00451 \times 10^{11}$	$.00472 \times 10^{11}$	-	$.00448 \times 10^{11}$	$.28146 \times 10^{20}$	$.38104 \times 10^{20}$
Within Component	$.12726 \times 10^{11}$	$.12850 \times 10^{11}$	-	$.12881 \times 10^{11}$	$.05351 \times 10^{20}$	$.19794 \times 10^{20}$
Variance $(\hat{R})$	$.33143 \times 10^{-1}$	$.33456 \times 10^{-1}$	$.33684 \times 10^{-1}$	$.33500 \times 10^{-1}$	$.15205 \times 10^{-3}$	$.15466 \times 10^{-3}$
Between Component	$.00849 \times 10^{-1}$	$.00801 \times 10^{-1}$	-	$.00736 \times 10^{-1}$	$.17848 \times 10^{-3}$	$.24482 \times 10^{-3}$
Within Component	$.32294 \times 10^{-1}$	$.32655 \times 10^{-1}$	-	$.32764 \times 10^{-1}$	$.03530 \times 10^{-3}$	$.13004 \times 10^{-3}$
Covariance (X',Y')	$.12639 \times 10^{9}$	$.13492 \times 10^{9}$	$.15505 \times 10^{9}$	$.13446 \times 10^{9}$	$.73779 \times 10^{17}$	$.74074 \times 10^{17}$
Between Component	$.06896 \times 10^{9}$	$.08051 \times 10^{9}$	-	$.08236 \times 10^{9}$	$.79886 \times 10^{17}$	$.90431 \times 10^{17}$
Within Component	$.05743 \times 10^{9}$	$.05441 \times 10^{9}$	-	$.05211 \times 10^{9}$	$.06102 \times 10^{17}$	$.20750 \times 10^{17}$

Table 4 - A comparison of the variance estimators for Population III when the design is stratified two stage with random sampling without replacement at both stages

	Emperical Expectations				Emperical Variances	
Parameter	True Value	Standard Estimator	Uncorrected Half-sample Estimator	1st and 2nd Stage Corrected Half-sample Estimator	Standard Estimator	1st and 2nd Stage Corrected Half-sample Estimator
Variance (X')	$.66963 \times 10^{8}$	$.66751 \times 10^{8}$	$.75928 \times 10^{8}$	$.66715 \times 10^{8}$	$.66280 \times 10^{15}$	$.66331 \times 10^{15}$
Between Component	$.36933 \times 10^{8}$	$.36705 \times 10^{8}$	-	$.36852 \times 10^{8}$	$.65507 \times 10^{15}$	$.71514 \times 10^{15}$
Within Component	$.30030 \times 10^{8}$	$.30046 \times 10^{8}$	-	$.29862 \times 10^{8}$	$.05168 \times 10^{15}$	$.12930 \times 10^{15}$
Variance (Y')	$.62834 \times 10^{9}$	$.62329 \times 10^{9}$	$.66076 \times 10^{9}$	$.62278 \times 10^{9}$	$.58486 \times 10^{17}$	$.59209 \times 10^{17}$
Between Component	$.15591 \times 10^{9}$	$.14987 \times 10^{9}$	-	$.15190 \times 10^{9}$	$.61057 \times 10^{17}$	$.71967 \times 10^{17}$
Within Component	$.47243 \times 10^{9}$	$.47342 \times 10^{9}$	-	$.47088 \times 10^{9}$	$.08036 \times 10^{17}$	$.25288 \times 10^{17}$
Variance $(\hat{R})$	$.93790 \times 10^{-3}$	$.93019 \times 10^{-3}$	$.93389 \times 10^{-3}$	$.93068 \times 10^{-3}$	$.11931 \times 10^{-6}$	$.12370 \times 10^{-6}$
Between Component	$.02878 \times 10^{-3}$	$.01379 \times 10^{-3}$	-	$.01282 \times 10^{-3}$	$.13963 \times 10^{-6}$	$.17550 \times 10^{-6}$
Within Component	$.90912 \times 10^{-3}$	$.91640 \times 10^{-3}$	-	$.91785 \times 10^{-3}$	$.02866 \times 10^{-6}$	$.09529 \times 10^{-6}$
Covariance (X',Y')	$.13251 \times 10^{9}$	$.13183 \times 10^{9}$	$.15004 \times 10^{9}$	$.13165 \times 10^{9}$	$.45349 \times 10^{16}$	$.45417 \times 10^{16}$
Between Component	$.07310 \times 10^{9}$	$.07282 \times 10^{9}$	-	$.07357 \times 10^{9}$	$.45642 \times 10^{16}$	$.51028 \times 10^{16}$
Within Component	$.05941 \times 10^{9}$	$.05900 \times 10^{9}$	-	$.05808 \times 10^{9}$	$.03784 \times 10^{16}$	$.10857 \times 10^{16}$

(a) The standard textbook formulas which give unbiased estimates for Var (x'), Var (Y') and Cov (x', y'). These are nearly unbiased for Var ($\hat{R}$);

(b) The balanced half-sample method without correcting for either the first or second stage sampling fraction; (See Casady (1975) for a discussion of BRR estimates of covariance)

(c) The balanced half-sample method corrected for both levels of sampling.

In addition the "within" and "between" components of variance for Var(x'), Var(y') and Var ($\hat{R}$) and the "within" and "between" components of covariance are calculated by both the standard textbook formulas and by the balanced half-sample method.

5.3 Expectations and Variances of Second Order Estimators

The expectations and variances of all the second order estimators for each of the three populations were estimated by drawing 1000 samples from each of the three populations. These results are given Tables 2, 3, and 4. This study was primarily exploratory and a far more extensive study is necessary before any definitive conclusions can be drawn, however, it would appear that the BRR estimates of the variance components will be satisfactory. Also, the extensions of the BRR method to the estimation of covariance seems to be satisfactory. The most encouraging result at this time is the fact that the variance of the BRR variance estimators are of the same magnitude as the standard estimators.

6. References

Casady, Robert J. (1975) "A BRR Estimator of Covariance". National Center for Health Statistics (Internal Memorandum).

Frankel, Martin R. (1971) Inference to Survey Samples An Emperical Investigation. University of Michigan, Ann Arbor, Michigan.

Kish, Leslie and Frankel, Martin R. (1971) "Balanced Repeated Replication for Standard Errors". Journal of The American Statistical Association, 65, 1071-94.

McCarthy, Philip J. (1966) "Replication: An Approach to the Analysis of Data from Complex Surveys". Vital and Health Statistics. PHS Publication No. 1000, Series 2, No. 14. Government Printing Office, Washington, D.C.