COMPONENTS OF VARIANCE BY REPLICATED BRR

Eric Schindler and Stanley Kulpinski, Bureau of Labor Statistics

SUMMARY

This paper derives components of variance from several variance estimates made by balanced repeated replication. Two-stage sample designs with one within- and one between-p.s.u. component of variance, and two-way sample designs with one within- and two between-sampling unit components of variance are discussed. Variances are recalculated using only a portion(s) of the second stage (or within cell) sample from all primary sampling units. The increase(s) in the variance estimate(s) now made by balanced repeated replication is (are) used to isolate the components of the classical variance model as the solution of two (three) simultaneous equations in two (three) unknowns.

INTRODUCTION

In the past decade many statistical surveys have turned to balanced repeated replication (BRR) to estimate variances as outlined by McCarthy(1966) and others. The difficulty of developing and/or using exact formulae, and general satisfaction with the results of BRR when applied to complex estimators have resulted in this change. Although Casady(1975), Bean and Schnack(1977), and perhaps others have done some work isolating components of variance when using BRR, very little has been published on this topic.

The next section of this paper gives a quick review of BRR. The main section replicates the variance calculation to obtain estimates for the within- and between-p.s.u. components of variance for a two-stage sample design. The remainder of the paper discusses two-way sampling, compares replicated BRR with the method suggested by Casady, and provides a simple example.

This technique was developed for the Bureau of Labor Statistics' Producer Price Index Revision. Components of variance from this expanded program will not be available for several years.

Our thanks are extended to Drs. John C. Koop and Babubhai Shah of the Research Triangle Institute, Dr. Robert Casady of the National Center for Health Statistics and Wes Schaible and Ronny Schaul of BLS.

BALANCED REPEATED REPLICATIONS

Assume a frame with N primary sampling units divided into K strata each with N(k) units. We select two independent samples of size n(k) in each stratum with the selected p.s.u. indexed by (k,i) or (k,i') for $k=1,\ldots,K$; $i=1,\ldots,n(k)$; $i'=1,\ldots,n(k)$. Sampling weights W(k,i) and W(k,i') are developed as appropriate for the application. We note that the same p.s.u. can be selected in both of the independent samples.

Assume that a second stage sample of size n(k,i) or n(k,i') is chosen in each p.s.u. with n(,) at least 2 for all k, i and i'. For each selected second stage unit an observation x(k,i,j) or x(k,i',j) is made with k,i and i' as above and $j=1,\ldots,n(k,i)$ or $j=1,\ldots,n(k,i')$. We can safely assume that the second stage sampling procedure is self-weighting and that two independent second stage samples can be selected from p.s.u.'s selected in both first stage samples.

We can use $\sum_{j=1}^{n(k,i)} x(k,i,j)/n(k,i)$ as an unbiased estimator for $\bar{X}(k,i)$ and $\sum_{j=1} x(k,i',j)/n(k,i')$ as an unbiased estimator for $\bar{X}(k,i')$.

Assuming that the sampling weights are normalized, the population mean $\bar{X}$ can be estimated by:

$$\hat{\bar{X}} = \sum_{k=1}^{K} \sum_{i,i'=1}^{n(k)} W(k,i)\hat{\bar{X}}(k,i)+W(k,i')\hat{\bar{X}}(k,i')$$

Following McCarthy we can construct an appropriate matrix with perpendicular columns and then define M replicates and complements. If k is less than 112, we can control M such that $K < M < K + 5$. A matrix for 4, 5, 6 or 7 strata is:

Stratum	Replicate Number 1	2	3	4	5	6	7	8
1	A	B	B	A	A	B	B	A
2	A	A	B	B	A	A	B	B
3	A	B	A	B	A	B	A	B
4	A	A	A	A	B	B	B	B
5	A	B	B	A	B	A	A	B
6	A	A	B	B	B	B	A	A
7	A	B	A	B	B	A	B	A

where an "A" means that unprimed p.s.u.'s are used from a particular stratum in a particular replicate, and a "B" means that primed p.s.u.'s are used. For example, using the above matrix, the 4th replicate would contain data from the unprimed p.s.u.'s from strata 1, 4 and 5 and data from the primed p.s.u.'s from strata 2, 3, 6 and 7.

We now define

$$\hat{\bar{X}}(m)= \sum_{k=1}^{K} \sum_{\substack{i=1\\ i'=1}}^{n(k)} [2\,M(k,m)\,W(k,i)\,\hat{\bar{X}}(k,i)] + [2\{1-M(k,m)\}W(k,i')\hat{\bar{X}}(k,i')]$$

and

$$\hat{\bar{X}}(m')= \sum_{k=1}^{K} \sum_{\substack{i=1\\ i'=1}}^{n(k)} [2\{1-M(k,m)\}W(k,i)\hat{\bar{X}}(k,i)]+ [2\,M(k,m)\,W(k,i')\hat{\bar{X}}(k,i')]$$

to be the estimated values for the mth replicate and for the mth complement respectively.

Following McCarthy and others we use:

$$V^2(\hat{\bar{X}}) = \sum_{m=1}^{M} [\hat{\bar{X}}(m) - \hat{\bar{X}}(m')]^2 \Big/ 4M$$

or

$$V^2(\hat{\bar{X}}) = \sum_{m=1}^{M} [\hat{\bar{X}}(m) - \hat{\bar{X}}]^2 \Big/ M$$

as an estimate for the variance of $\hat{\bar{X}}$.

For linear estimators, the estimate of the variance is algebraically equivalent to the classical estimates. It is known to be biased for ratio estimators and for most other complex non-linear estimators. However, for many complex estimators this bias seems sufficiently small to make BRR an attractive alternative to the myriad complexities of developing and using exact formulae.

REPLICATED BRR FOR TWO-STAGE DESIGNS

Traditional statistical theory for a stratified multi-stage sample tells us that:

$$V^2(\bar{X}) = S_B^2/P + S_W^2/PQ$$

where S_B^2 and S_W^2 are the components of variance between- and within-primary sampling units, P is the first stage sample size, and Q is the average second stage sample size.

If the variance was estimated by BRR as indicated above, then

$$P = \sum_{k=1}^{K} 2\, n(k)$$

and

$$P\,Q = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i'=1}}^{n(k)} [n(k,i,j) + n(k,i',j)]$$

If we now randomly take a fraction, such as one-half or two-thirds of the second stage sample, and recalculate the variance using BRR for this subset of the data, the expected value of the revised variance estimate will be larger. If half the second stage data are used, we will have $P' = P$ and $Q' = Q/2$. We now have two estimates of the variance and we can define two simultaneous equations with the between- and within-p.s.u. components of variance as the unknowns:

$$V^2(\hat{\bar{X}}) = S_B^2/P + S_W^2/PQ$$

$$V'^2(\hat{\bar{X}}) = S_B^2/P' + S_W^2/P'Q'$$

These equations have the solution:

$$S_B^2 = \frac{P\,Q\,V^2(\hat{\bar{X}}) - P'\,Q'\,V'^2(\hat{\bar{X}})}{Q - Q'}$$ and

$$S_W^2 = \frac{(P'\,V'^2(\hat{\bar{X}}) - P\,V^2(\hat{\bar{X}}))\,Q\,Q'}{Q - Q'}$$

If $P' = P$ and $Q' = Q/2$ then

$$S_B^2 = P(2V^2 - V'^2) \text{ and } S_W^2 = PQ(V'^2 - V^2).$$

REPLICATED BRR FOR TWO-WAY DESIGNS

If we have a two-way sample design, such as that used in the Consumer Price Index, the variance estimate is given by:

$$V^2(\bar{X}) = S_{BU}^2/P + S_{BI}^2/Q + S_W^2/PQR$$

where S_{BU}^2, S_{BI}^2, and S_W^2 are the two between- and the one within-sampling unit components of variance respectively; P is the number of p.s.u.'s selected in the first sample design; Q is the number of p.s.u.'s selected in the second sample design; R is the average number of selections made in the PQ selected cells.

If we generate two random subsets of different sizes of the PQR observations, we will have three simultaneous equations with the variance components as the three unknowns.

$$V^2(\hat{\bar{X}}) = S_{BU}^2/P + S_{BI}^2/Q + S_W^2/PQR$$

$$V'^2(\hat{\bar{X}}) = S_{BU}^2/P' + S_{BI}^2/Q' + S_W^2/P'Q'R'$$

$$V''^2(\hat{\bar{X}}) = S_{BU}^2/P'' + S_{BI}^2/Q'' + S_W^2/P''Q''R''$$

COMPARISION WITH CASADY'S METHOD

The method proposed by Casady in 1975 differs substantially from our technique of replicating the basic BRR procedure. Casady, in his 1975 paper, treats each p.s.u. as a separate stratum in order to recalculate variances. The second stage observations are then split into two pseudo-p.s.u.'s. In thus eliminating the between p.s.u. component of variance, Casady at least doubles the number of strata and the number of replications required to estimate the components of variance. Replicated BRR, however, uses the same variance estimation techniques to isolate the components of variance as was originally used for the variance. This simplification and the flexibility of being able to use different subsamples of different sizes, could make our method preferable for many applications.

EXAMPLE

We consider the following example with three strata, two sampled p.s.u.'s per stratum, and two sampled second-stage units per p.s.u..

		A half-sample		B half-sample	
		odd data	even data	odd data	even data
STRATUM	1	4	5	6	6
	2	8	7	5	5
	3	6	4	7	9

The sample mean is 6.00.

We will use the following matrix with perpendicular columns to determine the replicates.

		REPLICATE			
		1	2	3	4
STRATUM	1	A	A	B	B
	2	A	B	A	B
	3	A	B	B	A

We can now estimate four replicates and four complements using, for example, the A-data from Stratum 1 and the B-data from Strata 2 and 3 in Replicate 2.

$\hat{\bar{X}}(1) = 5.67$ $\hat{\bar{X}}(1') = 6.33$

$\hat{\bar{X}}(2) = 5.83$ $\hat{\bar{X}}(2') = 6.17$

$\hat{\bar{X}}(3) = 7.17$ $\hat{\bar{X}}(3') = 4.83$

$\hat{\bar{X}}(4) = 5.33$ $\hat{\bar{X}}(4') = 6.67$

The variance is now calculated using the formula:

$$\hat{V}^2(\bar{X}) = \sum_{m=1}^{M} [\hat{\bar{X}}(m) - \hat{\bar{X}}(m')]^2 \Big/ 4M$$

The estimate is:

$$\bar{V}^2(\bar{X}) = \frac{(2/3)^2 + (1/3)^2 + (7/3)^2 + (4/3)^2}{16} = 0.486$$

Four separate sets of four replicates are aggregated: one using only odd data from both half-samples; one using only even data from both half-samples; and two using odd data from one half-sample and even data from the other. Variances for all four sets of estimates are calculated by the same formula. The full results for the odd/odd data and the variances for all four sets of data are as follows:

$\hat{\bar{X}}(1) = 6.00$ $\hat{\bar{X}}(1') = 6.00$

$\hat{\bar{X}}(2) = 5.33$ $\hat{\bar{X}}(2') = 6.67$

$\hat{\bar{X}}(3) = 7.00$ $\hat{\bar{X}}(3') = 5.00$

$\hat{\bar{X}}(4) = 5.67$ $\hat{\bar{X}}(4') = 6.33$

The variance results are:

$$\bar{V}'^2_{odd\ odd}(\bar{X}) = \frac{0^2 + (4/3)^2 + 2^2 + (2/3)^2}{4 * 4} = 0.389$$

Similarly, $\hat{V}'^2_{even\ odd}(\bar{X}) = 0.389$,

$\hat{V}'^2_{odd\ even}(\bar{X}) = 0.611$ and

$\hat{V}'^2_{even\ even}(\bar{X}) = 0.833$

Taking the average we obtain:

$\hat{V}'^2(\bar{X}) = 0.556$

Using the one variance estimate based on all available data and the average of the four estimates for the subsamples of the data, we generate and solve the two simultaneous equations in two unknowns indicated above. We obtain:

$$S_B^2 = \frac{6*2*.486 - 6*1*.556}{2 - 1} = 2.5000$$

$$S_W^2 = \frac{(6*.556 - 6*.486)*2*1}{2 - 1} = 0.8333$$

Using Casady's method we would need 6 strata and 8 replicates to isolate the within p.s.u. component of variance for the same data set. Identical results would be obtained. Classical methods also provide the same results for this simple estimator.

CONCLUSION

The equivalence with classical methods of calculating components of variance can be shown analytically for simple linear estimators provided that all possible combinations of subsets of data are used when recomputing the increased variances. In the case that second-stage samples are split in two, this involves estimating increased variances four times at what we suspect would be a reasonable cost when compared to the total processing costs. Moreover, the replication method proposed here is independent of the technique used in estimating the variances. Even if the variance was estimated by some technique other than balanced repeated replication, this method can be applied to replicate whatever method was used for the original variance estimates in order to calculate the increased variances needed to isolate the components of variance.

In an attempted analysis of a small and less than perfect data set from the just completed BLS Hospital and Nursing Home Construction Labor and Material Requirements Survey, results by Casady's method seemed more appropriate than those using the replicated BRR method proposed in this paper. We suspect that the type and quality of the data being processed may have been the primary source of the many negative estimates of between p.s.u. variances using both methods, but final determination has not yet been made. Producer Price Index data should provide

more enlightening results in several years because PPI sampling methodology has been redesigned with this application in mind. Also, the data generated by the much larger PPI will more readily lend itself to this type of analysis than the construction data.

Our next tasks include the further study of the theoretical possibilities for this method to isolate components of variance for multi-stage sample designs, within- and between-interviewer effects (using poststratification), etc.. We must identify techniques for insuring the independence of the two estimates of the variance and for determining the error of the estimated components of variance. We must also identify more appropriate data bases for additional empirical analysis of this replicated BRR variance technique for the isolation of variance components.

REFERENCES

McCarthy, Philip J., (1966),
"Replication: An Approach to the Analysis of Data From Complex Surveys,"
Vital and Health Statistics, Public Health Service Publication Number 1000, Series 2, Number 14.

Casady, Robert J., (1975),
"The Estimation of Variance Components Using Balanced Repeated Replication,"
Proceedings of the Social Statistics Section of the American Statistical Association, pp. 352-357.

Bean, Judy A. and Schnack, George A., (1977),
"An Application of Balanced Repeated Replication to the Estimation of Variance Components,"
Proceedings of the Social Statistics Section of the American Statistical Association, pp. 938-942.