Replicate Variance Estimates--Reducing Bias by Using Overlapping Replicates

Susan Hinkins and H. Lock Oh, Internal Revenue Service, and Fritz Scheuren, Ernst and Young

A replicate variance estimator can be useful when the form of the estimator is complex or when the sampling distribution is complex. In the example that motivated this work, the estimators are simple totals or means, but the sampling distribution is unwieldy as it involves the probability of being in different strata over time.

The original problem of interest grew out of the use of a permanent random number (PRN) for sample selection in the Statistics of Income (SOI) samples, in particular, the annual sample of corporate tax returns. This is a stratified probability sample designed, in part, to provide cross-sectional estimates of income and tax items for a particular year. Since it is an annual sample, estimates of year-to-year change are also of interest, and, for users within the Treasury Department, the primary interest is in modeling economic and tax dynamics over time using the microdata. By using a permanent random number in the sample selection, the year-to-year sample overlap is increased while maintaining the simplicity and validity of the cross-sectional estimation.

Because of the sample overlap, the precision of estimates of year-to-year change may be greatly improved. For variables with high year-to-year correlation, the standard error may be reduced by as much as one-half, compared to independent samples. Calculating estimates of variance is more difficult, however, since the probability of a unit being in both samples depends on its sampling stratum each year, and this can change.

Hinkins, Moriarity, and Scheuren (1996) describe the SOI corporate sample, the selection using a PRN, and the resulting year-to-year overlap. A method is given for defining replicates using a PRN, so that a unit stays in the same replicate over time. In this way, replicate variance estimation can be used for estimating the variance of estimates of year-to-year change.

One difficulty with the replicate procedure, in general, is that it does not account for the finite population correction (fpc). In the simple random sample case, the replicate variance estimate is an unbiased estimate of the variance of the mean or total only if the fpc can be ignored. The SOI corporate sample has sampling rates as large as .5, in which case, the fpc cannot be dismissed.

In this paper, we describe a general modification to the usual replicate variance estimator to adjust for the finite population correction. The second section gives a brief description of the replicate variance estimator in general and describes the proposed adjustment to the replicate methodology. The third section discusses the case where one wants a variance estimate for the original estimator rather than the replicate estimator. An example based on the SOI corporate sample is given. The fourth section briefly summarizes the results and describes future work.

■ Overlapping Replicate Variance Estimators

Replicate variance estimators are useful in many cases where the variance calculation is complex, as they only require calculation of the point estimate (mean, ratio, total, etc.). Suppose the sample is of size n where n=m*G. The dependent random group's variance estimator (e.g., Wolter 1985) is calculated by using a random mechanism to divide the sample into G groups, each of size m. The estimator of interest, say, $\hat{X}$, is calculated in each group, $\hat{X}_g$. The replicate estimator and variance estimator are

$$\hat{X}_\bullet = \frac{1}{G}\sum_{g=1}^{G} \hat{X}_g$$

$$V_1 = \hat{Var}(\hat{X}_\bullet) = \frac{1}{G(G-1)} \sum_{1}^{G} (\hat{X}_g - \hat{X}_\bullet)^2 \; .$$

For estimators of means and totals, the replicate estimator is equal to the original estimator, and, with simple random sampling with finite population correction (fpc), we have

$$\hat{X}_{\bullet} = \hat{X} \quad and\ E(V_1) = \frac{Var(\hat{X})}{(fpc)}.$$

The replicate variance estimate is a conservative estimate, overestimating the variance, and is approximately unbiased only when the fpc is close to one.

For a stratified sample, if the fpc's were equal for every stratum, one could simply correct the replicate variance estimator. Having nonconstant fpc's across strata would be typical in highly skewed populations and is true in the SOI sample; hence, a simple adjustment is unavailable to us.

Overlapping Replicates

For most situations, there is a reasonably straightforward way to adjust the definition of the replicates in order to get an approximately unbiased estimate of variance. Note that in the case n=m*G, the expected value of the variance estimator, V_1, can be written (e.g., Wolter 1985) as

$$E(V_1) = Var(\hat{X}_{\bullet}) - \frac{1}{G(G-1)}\sum_{a \neq b} Cov(\hat{X}_a, \hat{X}_b).$$

If there is no intervention, because of the fpc's, the covariance terms between the estimators from different random groups are all equal and negative, and V_1 will, as a result, be positively biased. What if one could alter the covariance between replicate estimates, so that the total covariance term was approximately zero? Then, the replicate variance estimator, V_1, would be nearly unbiased.

Assume that the original sample has been divided into G dependent random groups, each of size m. And assume that the groups are randomly ordered and that the units within groups are randomly ordered. Then, we can denote the sample and the random groups as n units, 1, 2,..., n, where the first group consists of units 1 through m, and the second group consists of units m+1, m+2, ..., 2m, etc. Figure 1 shows an example with n = 20 and G = 5; each group then has m=4 members.

In general, we want to form G new groups by randomly selecting k units in the original group 1 to overlap with the next consecutive t groups. Then, we select k units from the original group 2 to overlap with the next consecutive t groups. Only two values of t need to be considered: t=1 and t=3.

Figure 1 shows an example for both t=1 and t=3, using k=1. In each case, there are still 5 replicates, but

Figure 1.--Overlapping Replicates, K=1

A. With t=1

Group	Units																			
1	x	x	x	x																x
2	x				x	x	x	x												
3						x			x	x	x	x								
4									x				x	x	x	x				
5															x		x	x	x	x

B. With t=3

Group	Units																			
1	x	x	x	x					x						x					x
2	x				x	x	x	x							x					x
3	x					x			x	x	x	x								x
4	x					x			x				x	x	x	x				
5						x			x						x		x	x	x	x

now, each replicate contains m + t*k units.

With overlapping replicates (k>0), the replicate estimate of the total, $\hat{X}_*$, no longer equals the original sample estimate of the total, $\hat{X}$. But conditional on the sample achieved,

$$E(\hat{X}_* \mid sample) = \hat{X} .$$

However, the replicate variance estimate, V_1, is now an estimator of the replicate estimate, $\hat{X}_*$, rather than the original estimate, $\hat{X}$.

In the case of t=3, the restriction $G \geq 5$ and $m \geq 3$ is needed, which is not an unreasonable requirement for using replicate estimates in general. Then, for t=1 or t=3,

$$E(V_1) = Var(\hat{X}_*) - h_t(k) * N * S^2$$

where

$$h_t(k) = \frac{t(t+1)kN}{(G-1)(m+tk)^2} - 1 .$$

By solving for the value of k, which makes $h_t(k) = 0$, an unbiased estimate, V_1, can be constructed.

In the case t=1, if the sampling rate and the number of replicates, G, satisfy

$$\frac{n}{N} < \frac{1}{2} + \frac{1}{2(G-1)}$$

then, there is a solution

$$k_1 = \frac{N-(n-m)-\sqrt{N(N-2(n-m))}}{G-1}$$

which satisfies $0 < k \leq m$. Therefore, for sampling rates no larger than .5, there is a solution for any value of G.

Most sampling designs probably fall into this category, i.e., with sampling rates all less than or equal to 0.5. If there are strata that are selected with probability 1.0, then, the usual solution is to include the entire certainty stratum in each replicate, as discussed later. For cases with sampling rates between .5 and 1.0, we can use t=3, and the solution

$$k_1 = \frac{2N-(n-m)-2\sqrt{N(N-(n-m))}}{3(G-1)}$$

satisfies $0 < k < m/3$ for all sampling rates.

Choosing t=1 vs. t=3

Since using t=3 gives a solution for all sampling rates, why bother with the case t=1? One reason is that the case t=1 is easier to construct. The second reason is that the case t=1 is more likely to result in a reduction in bias for smaller sampling rates.

At the exact solution, k, we would have an unbiased estimate. However, we get only an approximately unbiased estimate, V_1, because k must be rounded to an integer value. In order to assure a conservative estimate of variance, one should always round down. That is, in both cases where t=1 and t=3, one can show that rounding down will result in a negative value of $h_t(k)$, whereas rounding up will result in a positive value. So, rounding down will result in a (hopefully small) overestimate of the variance.

Therefore, if the exact solution, k, is less than 1, we round down to 0, and we do not reduce the bias. Conditions under which there will be a useful solution can be described in terms of the initial sampling rate, f, the population size, N, and the number of replicates, G. Namely, if

$$f \geq G * \left(\sqrt{\frac{t*(t+1)}{N*(G-1)}} - \frac{t}{N} \right)$$

then, the solution, k, will be greater than or equal to 1. One would hope that the value of the right-hand side of the inequality would be relatively small. Holding N and G fixed, the value of the right-hand side is smaller for t=1 than for t=3. Therefore, for sampling rates of .5 or less, the method of overlapping replicates using t=1 will reduce the bias of the variance estimate for smaller sampling rates compared to overlapping with t=3.

This is not such an important consideration for large populations. For example, with N=100,000 and G=25, one can get a bias reduction using t=3 for any design with sampling rate greater than .055. Using t=1, one can get a bias reduction for designs with sampling rates down to .023. But, at such small sampling rates, the bias of the usual replicate variance estimate is very small anyway. However, with smaller populations, the difference can be noticeable. Take, for example, N=10,000 and G=25. Using t=3, one gets a reduction in the bias only for sampling rates larger than .17. Using t=1, one can reduce the bias for sampling rates as low as .07. If the sampling fraction is .1, using the configuration with t=3 will not result in a bias reduction, but using t=1 will.

In general, the cases where this method does not reduce the bias appear to coincide with examples where the replicate estimate may not be useful in general, namely, small sample sizes. When the population size is small, one cannot have both a small sampling rate and a large number of replicates. This does not seem unreasonable; one cannot expect to use the replicate method if the sample is very small.

■ Replicate vs. Original Estimator

For estimation of means or totals, the usual random group's replicate estimator, with no overlap, is the same as the original estimator. In this case, V_1 is an estimator of the variance of the original estimate, $\hat{X}$. When overlapping replicates are used, the replicate estimator, $\hat{X}_*$, is no longer equal to the original estimator, $\hat{X}$. And the variance of the replicate estimator will be larger than the variance of the original estimator.

This is most immediately noticeable with certainty strata. The variance of the original estimate is zero. By including the entire certainty stratum in each replicate, this property is preserved, and we have

$$\hat{X}_* = \hat{X}\ ,\ \ Var(\hat{X}_*) = 0 \text{ and } V_1 = 0\ .$$

Note that one could also divide the certainty strata into G random groups and use the general solution with t=3 to find a value of k < m/3 that results in an unbiased estimator, $\hat{X}_*$, and an approximately unbiased variance estimator, V_1, for $\hat{X}_*$. But this is not the best solution for certainty strata, in the sense that

$$\hat{X}_* \neq \hat{X} \text{ and } Var(\hat{X}_*) > Var(\hat{X}) = 0.$$

In other cases as well, one may want a replicate variance estimate that is an unbiased estimate of the variance of the original sample estimate. This can be done using the fact that

$$Var(\hat{X}_*) = Var(\hat{X}) + E\left(Var(\hat{X}_*|sample)\right).$$

In the case t=1, we find, for totals,

$$E(V_1) = Var(\hat{X}) - N\left(h_1(k) - \frac{k(m-k)N}{n(m+k)^2}\right) S^2$$

where $h_1(k)$ was defined in Section 2. By solving for the value of k, say, k_2, that makes the coefficient on S^2 equal to zero, we have an unbiased estimate. In order for $0 < k_2 \leq m$, the same condition as before is required, and the solution in terms of the proportion of overlap is

$$\frac{k_2}{m} = \frac{1}{1-f}\left[f - \frac{(G+1)}{2(G-1)}\left(1 - \sqrt{1 - \frac{8f(G-1)}{(G+1)^2}}\right)\right]$$

which lies between 0 and 1.

There are several choices here. For a given value of t, either t=1 or t=3, there are three replicate estimators of interest, namely, those associated with k=0 (no overlap) or $k=k_1$ or $k=k_2$. With each replicate estimator, there is an associated replicate variance estimator, $V_1(k)$.

Using k=0, the replicate estimator is the same as the original estimator. But the associated variance estimate, $V_1(0)$, can be exceedingly conservative when the sampling rates are not small.

Using $k=k_1$, V_1 is an approximately unbiased estimate of the variance of the **replicate** estimator. It is a conservative estimate of the variance of the original estimator. That is, using exact values of k, we would have

$$E(V_1(k_1)) = Var(\hat{X}_*(k_1)) \geq Var(\hat{X})\ .$$

Using $k=k_2$, V_1 is an approximately unbiased estimate of the variance of the **original** estimator. But V_1 will underestimate the variance of its associated replicate estimator:

$$E(V_1(k_2)) = Var(\hat{X}) \leq Var(\hat{X}_*(k_2))\ .$$

For best results, one needs to decide on the estimator of interest before determining the amount of overlap in the replicates, or else provide more than one definition of replicates. For a general purpose data base, a reasonable compromise might be to use the construction with an overlap of k_1 units. Then V_1 is an unbiased estimate of the replicate estimate of the total. And, as we will see in the next example, even though it is a biased (but conservative) estimate of the variance of the original estimator, it can be much better than the usual replicate variance estimator.

An Example from SOI

Take as an example a simplified version of some of the non-certainty SOI strata for the regular corporations, as shown in Table 1. The second and third columns give the population and sample sizes, respectively. Using G=25 replicates, the fourth column shows the resulting original group size, m.

Since the largest sampling rate is .5, we can use the configuration, t=1, for all strata. Column 5 gives the value of k_1, the number to overlap in order to get an approximately unbiased variance estimate of the replicate estimator. Column 6 shows the value of k_2, the overlap needed in order to use V_1 as an (approximately) unbiased estimate of the original stratified estimate of the total (or mean).

The last two columns show the relative increase in variance, by strata, if we use the replicate estimate of the total, compared to the original weighted stratum estimate. For example, in stratum 3, using overlapping replicates with k=6, the variance of the replicate estimate, $\hat{X}_*$, is approximately 5 percent larger than the variance of the usual estimate, $\hat{X}$.

For t=1, the maximum increase in variance occurs at k= m/3. So, if $k_1<k_2<m/3$, using k_2 results in a larger variance of $\hat{X}_*$ than using k_1. But if, as in stratum 5, $m/3 < k_1 < k_2$, then the variance of the replicate estimator using k_2 is smaller than the variance of the replicate estimator using k_1.

Suppose we are interested in using V_1 as an estimator for the variance of the **original** stratified estimate of the total. We can calculate the relative bias, B, of the estimator, $V_1(k)$, for each value of k:

$$E(V_1(k)) = Var(\hat{X}) * (1+B(k))$$

where B depends on population and sample sizes, the number of replicates, G, as well as the size of the over-

Table 1.--Example of Overlapping Replicates for Stratified Design, G=25, t=1

						Relative increase in Var	
Stratum	N_h	n_h	m_h	k_1	k_2	Using k_1	With k_2
1	140,000	7,050	282	7	13	.024	.042
2	50,000	2,950	118	3	6	.025	.046
3	28,000	2,950	118	6	12	.049	.084
4	20,000	5,950	238	49	92	.160	.176
5	10,000	5,000	200	133	184	.161	.040

lap, k. Table 2 shows the relative bias by stratum. Note that even though k_1 is not optimal, it gives considerably better estimates of the variance than the usual replicate variance estimate (k=0), especially when the fpc is not close to one. And it should give approximately unbiased estimates of the replicate estimate of the total.

Table 2.--Relative Bias in V_1, for Estimating $\hat{X}$

Stratum	k=0	k=k_1	k=k_2
1	.053	.025	.003
2	.063	.034	.007
3	.120	.059	.007
4	.420	.164	.003
5	1.000	.161	.0001

The relative bias using $k=k_2$ should be zero. It is only approximately zero because k is rounded to an integer. The bias in stratum 5 is so much smaller than the others because, in this case, the exact value of k is 184.03, compared to stratum 3, where the exact value of k is 12.8.

Suppose we decide to define the replicates using k_2. In each stratum, we randomly divide the sample units into 25 groups and randomly order the 25 groups. In stratum 1, in each group of 282, we randomly select 13 units and include these in the "next" group as well, etc. Therefore, each replicate in stratum 1 will have 295 units; each replicate in stratum 2 will have 124 units, etc. In this way, 25 replicates, each of size 2,021, are formed.

■ Conclusions and Future Work

The results shown here imply that when the fpc factor cannot be ignored, we could improve considerably over the usual dependent group estimates of variance by using overlapping replicates. And this technique is programmable. The results shown here are exact for the relatively unrealistic case where n=m*G. In practice, we will have some slight variation in the size of replicates (m vs. m+1), and, for overlapping units, it would be more convenient to use a rate of overlap, k/m, so that there might not be exactly k units selected each time. We are in the process of doing simulation studies to evaluate the reduction in bias using this technique in more realistic conditions, and in the original problem of estimation of year-to-year change.

■ References

Cochran, W. (1977), *Sampling Techniques*, John Wiley and Sons, Inc., New York.

Hinkins, S.; Moriarity, C.; and Scheuren, F. (1996), "Replicate Variance Estimation in Stratified Sampling with Permanent Random Numbers," *Proceedings, Section on Survey Research Methods, American Statistical Association.*

Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag.