REPLICATE VARIANCE ESTIMATION IN STRATIFIED SAMPLING WITH PERMANENT RANDOM NUMBERS

Susan Hinkins, IRS, Chris Moriarity, NCHS, and Fritz Scheuren, George Washington University
Susan Hinkins, 1122 South 5th Ave., Bozeman MT 59715

KEY WORDS: Time Series, Repeated Sampling

1. INTRODUCTION

A permanent random number (PRN) is a convenient way of sampling from administrative lists, especially if they are computerized. The idea goes back a long way, even before the advent of computers -- at least to the days when statistical work was still being done on tab machines that mechanically sorted punch cards.

Ben Mandel, for example, instituted the use of PRN's for sampling in the 1930's and 1940's at the Social Security Administration (e.g., Perlman and Mandel, 1944 and Perlman, 1988).. At that time he began the Continuous Work History Sample. This sample has now tracked U.S. social security accountholders for some 60 years.

In large automated record systems, like Social Security, each record typically has an identifying number (supposedly unique) assigned in some systematic, usually nonrandom fashion (for example, an ID number used internally by an agency). Administratively, this number is not supposed to change over time. For use in sampling, depending on how it was assigned, the number can be employed directly or transformed by a conventional pseudo-random number generator. In any case, as we will see, permanent random numbers afford many advantages, from cost savings to variance reduction, especially when estimates are required over time.

In the present paper, we will examine a PRN application at the Internal Revenue Service (IRS). Virtually all major administrative statistical samples at Internal Revenue employ the PRN technology; however, our interest here will center primarily on just one of these -- the highly stratified Statistics of Income (SOI) annual sample of corporate tax returns (e.g., Hughes et al 1996).

Organizationally, the paper is divided up into 5 parts. This introduction is Section 1. In Section 2 the current SOI sampling process using permanent random numbers is described and alternative variance estimators are briefly reviewed. Section 3 illustrates, via a simulation, how and under what circumstances we might improve the current variance estimation for year-to-year change. Then Section 4 describes what happens when the finite population correction factor cannot be ignored. Areas of future study are outlined in the last section -- Section 5.

2. SAMPLE SELECTION USING THE PRN

The use of permanent random numbers is a fairly standard procedure in the sample designs at the Statistics of Income Division of the Internal Revenue Service (e.g., Harte, 1986). The SOI sample discussed in this paper is from the population of corporate tax returns (filed on Forms 1120). For each tax year, a stratified probability sample is selected, with sample rates ranging from .0025 to 1. The strata are defined by size of the corporation and the type of corporation, in terms of the tax form used. The certainty strata (generally large corporations) constitute the bulk of the sample.

2.1 Background and Problem Statement. -- For all SOI samples, the main advantage of using a PRN in sample selection is that the year-to-year sample overlap can be increased; this, in turn reduces the variance of estimates of year-to-year change. There is a disadvantage, though -- namely that the variance estimation can in some cases become more complicated. Because of this complexity, currently in SOI, sometimes the variance estimator may be very conservative.

To get better estimates of precision for year-to-year change, we investigated the use of replication. This approach was first set forth by Mahalanobis (e.g., 1946) and popularized by Deming (e.g., 1960), among others. In order for replication to work, each identically distributed subsample or replicate must capture the properties of the overall sample design. For PRN sampling, this means that each subsample must include the mechanism that results in sample overlap. Therefore, the replicates must be defined using the permanent random numbers, themselves, so that a unit is always put in the same replicate over time.

The use of replicate methodology could also allow users to calculate variance estimates in the cross-sectional more easily too -- and without detailed knowledge of the sample design -- a clear benefit to users.

While the annual cross-section estimates of tax and revenue items have many applications, SOI's major user, the Office of Tax Analysis, is primarily interested in modeling economic/tax dynamics over time. In order to increase the year-to-year sample overlap, without sacrificing cross-section estimates, the current PRN design for corporations was first advocated in a 1975 WESTAT contract report. Some earlier theoretical papers which may have lead to the suggestions in the WESTAT report may be traceable to Brewer et al (1972). In any case, the use of the current PRN approach to sample selection, plus the large proportion of the sample drawn with certainty can result in a two-year sample overlap as large as 70%.

2.2 PRN Selection Method. -- The PRN sampling builds on an administrative process that assigns each corporation an employer identification number (EIN) which is unique to that corporation. A corporate return is selected into the sample as follows:

(1) The EIN is transformed to an 11 digit permanent random number, denoted as T, using a pseudo-random numerator generator of the form: T = (c * EIN) modulo n, where c and n are large, predetermined integers and $0 \leq T < n$.

(2) The key is to choose c and n so that values of the transforms, T, are roughly uniform and independent.

(3) Suppose that an individual falls into a stratum with probability of selection equal to p.

(4) If the number consisting of the last four digits of T is less than 10000*p, this return is selected into the sample.

If the same values of c and n are used each year, then the probability of an individual being in the sample over k years is equal to the minimum of its sampling rates over those years. In particular, large or "growing" corporations are more likely to stay in the sample over time. One last comment, whatever is happening to the sample from year to year, at any given point, the PRN method always yields a representative cross-section of the population, including births.

Typically, there can be a high correlation in some variables from one year to the next, in which case this selection technique significantly reduces the sample variance for estimates of year-to-year change. Problems arise in variance estimation, however. The variance of the estimate of a simple difference, say , between the totals in years 1 and 2, is

$$V(\hat{X}_1-\hat{X}_2)=V(\hat{X}_1)+V(\hat{X}_2)-2Cov(\hat{X}_1,\hat{X}_2)\ .$$

The first two terms above are the cross-section variances in each year and can be estimated in a standard way. The estimation of the covariance term, though, cannot -- since the probability of a unit being in both samples depends on its sampling stratum each year and this can change.

2. 3 Current Variance Estimator. -- Historically, what has often been done is to estimate the year-to-year variance assuming no correlation, i.e.

$$\widehat{Var}(\hat{X}_1-\hat{X}_2)=\widehat{Var}(\hat{X}_1)+\widehat{Var}(\hat{X}_2)$$

where, using the standard notation (Cochran 1977), and treating N_h and n_h as the known population and sample sizes for the h^{th} stratum, the variance of the cross-sectional estimate is

$$V(\hat{X}) = \sum_h (1-f_h) N_h^2 \frac{s_h^2}{n_h}$$

where $f_h = n_h/N_h$ is the sampling fraction in stratum h.

For variables which are highly correlated from year to year, and with the SOI design resulting in year-to-year overlaps that can approach 70%, this estimate of the variance of the difference can be extremely conservative, even "punishingly" so.

2.4 Replicate Variance Estimation . -- For a population total, say D, a replicate variance estimator is available as

$$\hat{V}(\hat{D}_*) = \frac{1}{G(G-1)}\sum_{g=1}^{G}(\hat{D}_g-\hat{D}_*)^2$$

$$where\quad \hat{D}_*=\frac{1}{G}\sum_1^G \hat{D}_g$$

and the G random group replicate estimates, $\hat{D}_g$, are each an estimate of the population total D.

This variance estimator will be approximately unbiased, assuming the sampling rates are small in every stratum (e.g.,Wolter 1985). If the finite population correction factors cannot be ignored, Wolter provides an alternative approach that, while awkward, seems workable for cross-section or level estimates. In Section 4 we will provide another technique that works for both level and change estimates.

In Sunter (1986), there is some discussion of the variance properties of permanent random number samples; but mainly in the cross-section and for the special case of statistics based solely on a simple longitudinal overlap from year to year. In Hinkins, Jones and Scheuren (1988), some of Sunter's observations are made with the SOI corporate application emphasized. The focus in the 1988 Hinkins et al paper, however, is on how to better control the overlap as the population elements shift from stratum to stratum over time.

Cross-section variances can be approximated in the usual way as all authors point out. However, in none of these settings are detailed approaches given for handling the impact of the overlap when estimating the variance of measures of change from year to year.

Our proposal here, for both cross-section and time series variances, is to use the PRN itself (digits other than those used in the selection) to create permanent replicates that can be employed in calculating variances. Two digits of the 11 digit PRN should allow for adequate degrees of freedom in the estimates. Perhaps 25 replicates, with 24 degrees of freedom, might be enough. The "t" value for a 95% confidence interval in this case would be 2.06, only slightly larger than the "z" value of roughly 2 that would ordinarily be used.

While clearly arbitrary, 25 replicates should be workable for SOI. Too many replicates raises the possibility of strata without observations. The SOI samples are large enough, though, to safely handle 25 replicates without this problem arising.

3. A SIMULATION STUDY

A population was created to form the basis of the simulations to be discussed. Two variables were generated, X and Y, highly related to each other and connected over time as well. These were simulated from a mixture of Gamma and Normal random variables for a population of 10,000 units. A stratified design of five strata, with an expected sample size of 500 in each year, was superimposed on this structure. The strata boundaries were fixed in the first year and remained unchanged even though the population values for X (the strata variable) grew over time.

A form of optimum allocation was employed for the first year to select strata boundaries but in both years the selection rates (f_h) chosen were far from proportional to the individual strata sizes. Births and deaths were not included in the simulation. A take-all stratum generally would arise in such populations too; but this was a feature we thought could be added later.

A 100 identically designed stratified samples were drawn from the population and divided into 11 replicates. The variables of interest were X, Y, and the ratio Y/X. Naturally, our focus was on measures of change in these quantities from one year to the next. The random group replicates were defined within each sample using part of the PRN, so that units would stay in the same replicate over time. Crucial to the interpretation of any approximate variance calculations is the degree of year-to-year correlation. Scatterplots, again based on all the replicates for all 100 samples, show that for X there is clear evidence of a strong relation. For Y this is less and for the ratio Y/X there is almost no pattern.

A simple natural competitor for the replicate variance calculations is to use the cross-section estimates calculated for each year and to treat the yearly samples as independent, as described in Section 2. This is, in effect, what is done currently in SOI. It is referred to here as the "naive" variance estimator. For the simulations, we ignored the fpc's in both the replicate and naive approaches. For each variable and for each of the 100 samples, the estimated standard error of the change was calculated using both methods. To compare the two techniques, the ratio (see Table A) was then calculated of the standard error using the replicate method over the standard error using the naive variance estimator.

Table A. -- Simulation Results: Ratio of Replicate to "Naive" Standard Error in 100 Samples

	X	Y	Y/X
Mean	.73	.93	.99
Median	.73	.94	1.00
IQR	(.60,.83)	(.82,1.05)	(.89,1.09)
Range	(.37,1.16)	(.57,1.31)	(.59,1.29)

Using the naive variance estimator for X resulted in an overstatement in standard error that was "punishingly conservative." In 94 of the 100 samples the naive estimator of variance overestimated the standard error as compared to using the replicate variance estimator. The replicate estimate averaged just 73% of the naive, conservative approach.

For the variable Y, the overstatement in standard error using the naive approach, as table A shows, turned out to be only a minor annoyance. Still the replicate variance approach was marginally better. For the ratio of Y/X, the confidence interval was lengthened slightly by using the replicate estimates over what would have been obtained by combining standard textbook estimates taken from the separate cross-sections. This last result is the penalty that has to be paid for using "t" rather "z" as the reference distribution for inference.

In summary, for variables with strong year-to-year relationships, permanent replicate variance estimators make sense. For variables with weaker or almost no year-to-year relationship, there appears to be little to gain or lose. Since SOI samples typically have variables of all types -- from those that are weakly related to those that are strongly related over time -- then acquiring the ability to calculate replicate variance estimates seems wise.

4. FINITE POPULATION CORRECTION

The replicate variance estimate is approximately unbiased if the finite population correction (fpc) is close to 1, i.e. if the sampling rates are small. This problem was not addressed in the simulation study -- neither estimator included a finite population correction.

4.1 Problem Formulation. -- Consider estimates of means or totals from a random sample of size n = mG. The sample can then be divided into G random groups, each of size m. Within each random group, an estimate of the total is calculated as

$$\hat{X}_g = \frac{N}{m}\sum_{i=1}^{m} X_{(g)i}$$

and the replicate estimate of the total is equal to the original estimate

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

The replicate variance estimate is

$$\hat{V}_1 = \frac{1}{G(G-1)}\sum_{g=1}^{G}(\hat{X}_g-\hat{X}_*)^2$$

and

$$E(\hat{V}_1) = \frac{Var(\hat{X})}{(fpc)} = \frac{Var(\hat{X}_*)}{(fpc)}\ .$$

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

For a stratified sample, such as ours, 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 our case too; hence a simple adjustment is unavailable to us.

4.2 A Partial Solution. -- When the sampling rates are no larger than .5, there is a reasonably straightforward way to adjust the definition of the replicates in order to get an approximately unbiased estimate of variance. When sampling rates are between .5 and 1.0, there is also a solution, but the approach we have so far is rather cumbersome; and will not be discussed here.

To motivate our idea, notice that the expected value of the variance estimator $\hat{V}_1$ can be written (e.g., Wolter 1985) as

$$E(\hat{V}_1) = Var(\hat{X}_*) - Cov(\hat{X}_a,\hat{X}_b)\ ,\quad a\neq b$$

If there is no intervention, because of the fpc's, the covariance between the estimators from different random groups will be negative; and V_1 will, as a result, be positively biased. What if one could alter the covariance between replicate estimates, so that the covariance was approximately zero? Then, the replicate variance estimator $\hat{V}_1$ 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; also suppose that the units within groups are randomly ordered, so that we can denote the sample and the random groups as n units, 1, 2,..., n. The first group consists of units 1 through m, the second group consists of units m+1, m+2, ..., 2m, etc. For example, if n = 12 and G = 3, we can denote the groups as:

Group	Units		
1	x x x x		
2		x x x x	
3			x x x x

Now we want to form 3 new groups by randomly selecting k units in the original group 1 to overlap with group 2, k units from the original group 2 to overlap with group 3, and k units in the original group 3 to overlap in turn with the new group 1. In the example above, if k = 1, this might look like:

Group	Units		
1	x x x x		x
2	x	x x x x	
3		x	x x x x

Therefore there are still 3 groups, but now each group has m+k units, and each adjacent group shares k units. Unfortunately, the replicate estimate of the total, X_* no longer exactly equals the original sample estimate of the total, X. Still, one may interested in using $\hat{V}_1$ as an estimate of the variance of X_*. In this case we have

$$E(\hat{V}_1) = Var(\hat{X}_*) - N\left(\frac{2kN}{(G-1)\,(m+k)^2}-1\right)S^2$$

and by solving for the value of k which makes the coefficient on S^2 equal to zero, we can construct an unbiased estimate V_1. To be useful, the solution k must satisfy $0 \leq k \leq m$. If the sampling rate and the number of replicates, G, satisfy

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

then the solution is

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

Therefore, for sampling rates no larger than .5, this is a solution for any value of G. This result is only an approximately unbiased estimate because we must round the exact solution to get an integer value for k. In order to assure a conservative estimate of variance, one should always round down.

4.3 Further Considerations. -- If rather than picking a value of k, we instead want the subsampling rate for the overlap, k/m, we find that this is a function only of the original sampling rate and the number of groups:

$$p \triangleq \frac{k}{m} = r-1-\sqrt{r(r-2)}$$

$$\text{where} \quad r=\frac{1}{f}\frac{G}{(G-1)} = \frac{N}{n}\frac{G}{G-1}$$

In some instances, one may want a replicate variance estimate that is an unbiased estimate of the variance of the **original** sample estimate. In this case we find

$$E(\hat{V}_1)=V(\hat{X}) - N^2\left(\frac{2k}{(G-1)\,(m+k)^2}-\frac{1}{N}-\frac{k(m-k)}{n(m+k)^2}\right)S^2$$

and by solving for the value of k, say k_1, that makes the coefficient on S^2 equal to zero, we have an unbiased estimate. In order for $0\leq k_1 \leq m$, the same condition as before is required, and the solution in terms of the proportion overlap is

$$p_1 \triangleq \frac{k_1}{m} = \frac{f}{1-f} - \frac{(G+1)}{2(1-f)\,(G-1)}\left(1-\sqrt{1-\frac{8f(G-1)}{(G+1)^2}}\right)$$

where this solution lies between 0 and 1. Note that for both p and p_1, if the sampling rate, f, is very close to zero, then the proportion of overlap is zero. Now if the fpc is approximately 1, this method indicates that no overlap is required -- as we would have hoped from our earlier discussion.

Table B shows some examples of the proportion of overlap required to achieve an approximately unbiased replicate estimate of variance for three original sampling rates, f, which are all less than or equal to .5. As expected, the amount of overlap needed increases with the sampling rate. In the table, the proportion of overlap required is shown for estimating both the variance of the replicate estimator and the variance of the original estimate. One needs significantly more overlap in order to estimate the variance of the original estimate.

Table B.-- Proportion of Overlap

f	G	p	p_1
.10	20	.053	.100
	50	.054	.107
	100	.055	.109
.30	20	.208	.387
	50	.218	.412
	100	.225	.428
.50	20	.635	.900
	50	.752	.960
	100	.818	.980

To illustrate, suppose one has a stratified sample with three strata and sampling rate .10, .30, and .50 respectively, as shown above. Suppose further we wish to employ a replicate variance estimator with G=20 replicates. Then each stratum is randomly divided into 20 groups. In the first stratum, approximately 5% in each group is chosen to also be included in the next group. In the second stratum, again 20 groups are defined, but in this case the rate of overlap is 20%, etc. In this way, G replicates of equal size are defined.

When the sampling rates are no larger than .5, then, we can devise replicates that produce nearly unbiased estimates of variance. It can be seen from Table B why this is limited to sampling rates of no larger than .5. When the sampling rate is equal to .5, as G increases, the percent overlap approaches 100%. The maximum amount of overlap has been reached.

As noted already, it is possible to use an overlap technique when the sampling rates fall between .5 and 1.0; but the method becomes more complicated and is not discussed here.

5. FUTURE STUDY

This paper indicates several directions for further study. Two problems were discussed: (1) The use of "permanent" replicates for variance estimation -- in order to improve on the "naive" standard error estimates of year-to-year change. (2) A method to pursue for improving the replicate variance estimator for any stratified sample -- when the fpc's cannot be ignored.

5.1 Value of Replicate PRN Variances. -- For SOI samples, like that for corporations, the number of variables which have punishingly conservative confidence intervals for estimates of change is large. Clearly in such settings a look at the operational issues of providing for permanent replicate variance estimation may be warranted. For other SOI samples, such as the sample of individual returns, permanent replicates may or may not be cost effective. In any event, to determine the relative merits of this methodology for the SOI environment, more simulations need to be done, constructing variables with longer tails; and, perhaps, stronger year-to-year relationships than we used initially.

5.2 Handling fpc's. -- We already have a general approach that builds in enough dependency between replicates so that the conventional replicate approach works routinely, no matter what the fpc's differences are from strata to strata -- as long as all sampling rates are no larger than .5. The SOI problem lies within these bounds. The next step is to put this technique in place, using both the overlap and the "permanent" replicate definition; and see, perhaps, how well the two work together in the SOI corporate sample.

For applications where at least one sampling rate is between .5 and 1.0, further work is needed. Our starting point for this paper and for this continuing work is the notion that when doing replicate estimation, the certainty strata are usually included in each replicate. For these strata, in other words, the random groups are identical, having complete and identical overlap in every replicate. As the sampling rates get larger and larger, therefore, we want a method which converges to this form of replication .

5.3 Handling complex estimation. -- We have already calculated the replicates with the full sample weights and with weights conditioned on the actual replicate strata sample sizes. Only the full sample results are discussed here. The raking estimates used in the corporate SOI program have not been reflected (e.g., Oh and Scheuren 1987). The performance of our approach in a raking setting needs to be checked. We are unsure whether separate raking weights by replicate will be needed.

5.4 Last Words. -- The key, as in all applied work, is to give the customer all the good they can afford (to rephrase Jefferson whose original words were "To give the people all the good they can endure"). We hope to continue this effort with that dictum in mind.

REFERENCES

Brewer, K., Early, L., and Joyce, S. (1972), Selecting Several Samples from a Single Finite Population, *Australian Journal Of Statistics,* 231-239.

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

Deming, W.E. (1960), *Sample Design in Business Research,* Wiley.

Harte, J.M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 603-608.

Hinkins, S., Jones, H., and Scheuren, F. (1988), Design Modifications for the SOI Corporate Sample: Balancing Multiple Objectives, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 216-221.

Hughes, S., Collins, R. and Uberall, B. (1996), Section 3, *Statistics of Income -- 1993, Corporation Income Tax Returns*, Internal Revenue Service, Washington, DC.

Mahalanobis, P.C. (1946), Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *Journal Royal Statistical Society,* 109, 325-370.

Oh. H.L. and Scheuren, F. (1987), Modified Raking Ratio Estimation, *Survey Methodology*.

Perlman, J. (1988), The Continuous Work History Sample: The First 12 Years, *Social Security Bulletin*, April, 1988.

Perlman, J. and Mandel, B. (1944), The Continuous Work History Sample Under Old-Age and Survivors Insurance, *Social Security Bulletin*, February, 1944.

Sunter, A.B. (1986). Implicit Longitudinal Sampling from Administrative Files: A Useful Technique, *Journal o Official Statistics*, 2, Statistics Sweden: Stockholm.

WESTAT, INC., (1974), Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income, Working Paper. Bethesda

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