COMPOSITE ESTIMATION FOR SAMPLES WITH OVERLAPPING SAMPLING UNITS

Jiahe Qian
ETS, Rosedale Road, 02-T, Princeton, NJ 08541

KEY WORDS: Repeated Surveys, Overlapping Sampling Units, Composite Estimation

In repeated surveys, overlapping of sampling units in different samples often occurs especially when target population is relatively small or when sampling units are large. For example, overlapping of sampled schools is common in some of the National Assessment of Educational Progress (NAEP) data, either between two repeated state samples in different years or between a state sample and a national main sample. The average scale scores, which is the statistic of primary interest in NAEP, usually have relatively high correlation coefficients in the overlapping part of two samples. Two basic issues for data from overlapping samples are: 1) what extent of overlapping would occur in repeated sampling and, 2) how to improve estimation with data from overlapping sampling units. There is some similarity between samples with overlapping and rotation samples. Some attention has been given to the issues in repeated surveys (Fuller, 1990; Lent, 1996).

1. Overlapping in Repeated Surveys

This discussion of the extent of overlapping will focus on simple random sampling (SRS) and area sampling. Moreover, sample selection schemes can also be characterized by sampling with replacement (wr) and sampling without replacement (wtr).

1.1 Overlapping in Repeated Simple Random Samples

There are three combinations of sampling schemes for two repeated simple random samples: SRS/wtr and SRS/wtr, SRS/wtr and SRS/wr, SRS/wr and SRS/wr. Suppose that N is a population size; n_1 and n_2 are the sample sizes for first and second samples.

1.1.a. Repeated Samples: SRS/wtr + SRS/wr. The first sample is SRS without replacement and the second is SRS with replacement. There are n_1 distinct units in the first sample. Define f_1, sampling fraction, as n_1/N, and $f_2 = n_2/N$.

In the second sample, the number of sample units which were selected from the first sample, ξ, forms a binomial distribution:

$$P(\xi = k) = \left[\begin{matrix} n_2 \\ k \end{matrix}\right] f_1^{k}(1 - f_1)^{n_2-k}.$$

Let η be the exact number of distinct units among these k units (Feller 1957, 92):

$$P(\eta = m \mid \xi = k) = \frac{1}{n_2^{k}} \left[\begin{matrix} n_2 \\ m \end{matrix}\right] \sum_{i=0}^{m} (-1)^i \left[\begin{matrix} m \\ i \end{matrix}\right] (m-i)^k.$$

The conditional expectation is

$$E(\eta \mid \xi = k) = n_2 \left[1 - \left[\frac{n_2 - 1}{n_2} \right]^k \right];$$

then the expectation of the number of overlapping units is

$$E(\xi, \eta) = n_2 \left[1 - \left[1 - \frac{n_1}{n_2 N} \right]^{n_2} \right].$$

It is assumed here that the two sample sizes are equal. When N=100, if n_1=n_2=40, there is an average of 12.93 sample units which overlap; if n_1=n_2=80, 43.84 sample units on average overlap. When N=500, if n_1=n_2=40, there is an average of 3 sample overlapping units; if n_1=n_2=100, there is an average of 17.98 sample units which overlap.

1.1.b. Repeated Samples: SRS/wr + SRS/wr. Consider both samples to be SRS/wr. The exact number of distinct units among n_1 sampled units in first sample is random, and can be expressed as

$$P(\zeta = g) = \frac{1}{N^{n_1}} \left[\begin{matrix} N \\ g \end{matrix}\right] \sum_{i=0}^{g} (-1)^i \left[\begin{matrix} g \\ i \end{matrix}\right] (g - i)^{n_1}.$$

As in (1.1.a), the conditional expectation of the number of overlapping units equals

$$E(\xi, \eta \mid \zeta = g) = n_2 \left[1 - \left[1 - \frac{g}{n_2 N} \right]^{n_2} \right];$$

therefore, the average of the number of overlapping units is $\sum_{g=1}^{n_1} E(\xi, \eta \mid \zeta = g) p(\zeta = g)$.

1.1.c. Repeated Samples: SRS/wtr + SRS/wtr. In simple random samples without replacement, data can be expressed in a two-by-two table; see Table 1.1.C. The number of overlapping units for two repeated samples forms a hypergeometric distribution:

$$P(\eta = m) = \left[\begin{matrix} n_2 \\ m \end{matrix}\right] \left[\begin{matrix} N - n_2 \\ n_1 - m \end{matrix}\right] \Big/ \left[\begin{matrix} N \\ n_1 \end{matrix}\right];$$

This research was partially supported under the National Assessment of Educational Progress (Grant No. R999G50001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The author thanks John Tukey and Juliet Shaffer for many helpful comments.

Table 1.1.C. Data for Two Repeated Samples of SRS/wtr + SRS/wtr

	Second Sample		
S1	In S2	Not in S2	Total
In S1	m	n_1-m	n_1
Not in S1	n_2-m	N-n_1-n_2+m	N-n_1
Total	n_2	N-n_2	N

the expectation of η is $E(\eta) = n_1 n_2/N$.

When N=100, if n_1=n_2=40, there is an average of 16 overlapping sample units; if n_1=n_2=80, on average 64 sample units overlap. When N=500, if n_1=n_2=40, there is an average of 3.2 sample units which overlap; if n_1=n_2=100, 20 sample units on average overlap.

1.2 Overlapping in Repeated Unequal Probability Samples

Let p_i be the inclusion probability for unit i in unequal probability sampling, and it is proportional to the unit size (PPS). Area sampling is a special scheme of PPS sampling. There are also three combinations of repeated samples: PPS/wtr + PPS/wr, PPS/wtr + PPS/wtr, PPS/wr + PPS/wr.

For unequal probability sampling, the formulas of the number of overlapping units would be more complex though the general ideas are similar. First, $P(\eta = m \mid \xi = k)$ can be expressed as in SRS samples (see the appendix). Given a sampling scheme, then the conditional expectation, $E(\eta \mid \xi = k)$, can be formulated. Finally, the expectation of the number of overlapping units, $E(\xi, \eta)$, can also be formulated. Although the formulas can be expressed, the complexity of calculation is NP-complete. However, computer simulation can be used to estimate the average number of overlapping units.

In the simulation of this research, assume that the size of sampling unit i forms a linear relationship with g: $m_i = \alpha g + m_0$, where α and m are constant and g forms a Gamma distribution. First select two PPS samples, then count the number of overlapping units. The replication of this procedure is between 300-500. PPS/wtr is implemented by systematic PPS sampling.

The results of the simulation are given in the Tables 1.2.A-1.2.C. The results show that the percentage overlapping in PPS is similar to that in SRS when population sizes are small; but when population sizes become larger, say 200 or more, the percentage overlapping in PPS is less than that in SRS. These conclusions hold for either sampling with replacement or without replacement.

2. Estimation in Repeated Samples with Overlapping

2.1 Composite Estimator. To improve the accuracy of estimates from samples with overlapping, different types of composite estimators can be applied. A general composite estimator can be expressed as

$$\bar{y} = \alpha \bar{y}_U + (1 - \alpha)\bar{y}_R,$$

where $\bar{y}_R$ is a model-based estimator for $\bar{Y}_2$, and $\bar{y}_U$ is an unbiased estimator for $\bar{Y}_2$. $\bar{y}_U$ can be estimated by $\bar{y}_2$, a weighted mean estimate based on second sample. And $\bar{y}_R$ can be constructed by different models. When correlation coefficients of two samples in the overlapping part are high, $\bar{y}_R$ could be a ratio estimator or regression estimator:

$$\bar{y}_R = \bar{y}_1 + \beta(\bar{y}_2^{o} - \bar{y}_1^{o}),$$

where $\bar{y}_1^{o}$ and $\bar{y}_2^{o}$ are two weighted mean estimators based on data from two samples in the part where they overlap. Empirical data also suggest to use regression estimators. For example, in Table 2.1, most of the correlation coefficient of school scale scores between 1996 and 1992 NAEP state math samples range from 0.6 to 0.85. The variance of $\bar{y}_R$ is

$$V(\bar{y}_R) = V(\bar{y}_1) + 2\beta Cov(\bar{y}_1, \bar{y}_2^{o} - \bar{y}_1^{o}) + \beta^2 V(\bar{y}_2^{o} - \bar{y}_1^{o}).$$

2.2 Optimization of β in $\bar{y}_R$. Consider $V(\bar{y}_R)$ under different βs. Let n_1^{o} be the size of the overlapping part from the first sample, n_2^{o} is defined similarly. Clearly $n_1^{o} = n_2^{o}$. Assume both subsamples in the overlapping part are random representatives of two original samples.

First, let $\beta = 1$, $\bar{y}_R$ is then unbiased. Define

$$\delta = \frac{n_1 - n_1^{o}}{n_1 n_1^{o}} S_1^2 - 2\rho \frac{n_1 - n_1^{o}}{n_1} \sqrt{\frac{1}{n_1^{o}} S_1^2 \frac{1}{n_2^{o}} S_2^2} ,$$

so $V(\bar{y}_R) \approx S_2^2/n_2^{o} + \delta$. If $\rho \geq 0.5 S_1/S_2$, $\bar{y}_R$ would gain more efficiency than $\bar{y}_2^{o}$; if $\delta > \frac{n_2 - n_2^{o}}{n_2 n_2^{o}} S_2^2$, $\bar{y}_R$ is more efficient than $\bar{y}_2$.

Second, minimum variance β can be obtained by solving $\partial V(\bar{y}_R)/\partial\beta = 0$:

$$\beta_{min} = \frac{n_1^{o}}{n_1} \frac{Cov(\bar{y}_1^{o}, \bar{y}_1^{o} - \bar{y}_2^{o})}{V(\bar{y}_2^{o} - \bar{y}_1^{o})};$$

substitute β_{min} into the variance formula for $\bar{y}_R$, then

$$V_{min}(\bar{y}_R) = V(\bar{y}_2) - \beta_{min}Cov(\bar{y}_1^o, \bar{y}_1^o - \bar{y}_2^o).$$

Assume $n_1^o/n_1 = 0.5$ and $\rho = 0.5$, then $V_{min}(\bar{y}_R) \approx 0.75V(\bar{y}_2)$. But, $\bar{y}_R$ is biased under β_{min} and bias= $(1-\beta)\Delta$ where $\Delta = \bar{Y}_1 - \bar{Y}_2$.

In general, by $\partial MSE(\bar{y}_R)/\partial\beta = 0$, minimum mean square error β can be derived:

$$\beta_{opt} = \frac{n_1^o/n_1 Cov(\bar{y}_1^o, \bar{y}_1^o - \bar{y}_2^o) - \Delta^2}{V(\bar{y}_2^o - \bar{y}_1^o) - \Delta^2}.$$

2.3 Cohen-Spencer Optimal Shrinkage Estimator. To treat $\bar{y}_R$ as "model-based" estimator, the Cohen-Spencer approach can be used to obtain optimal composite estimator (Cohen & Spencer, 1991), $\bar{y}_{CS} = \alpha_{CS}\bar{y}_U + (1-\alpha_{CS})\bar{y}_R$ where

$$\alpha_{CS} \triangleq \begin{cases} A_m/(A_u + A_m) & \text{if } A_m \geq 0 \text{ and } A_u \geq 0, \\ 1 & \text{if } A_u < 0, \\ 0 & \text{otherwise,} \end{cases}$$

with $A_m \triangleq V(\bar{y}_R) + Bias^2(\bar{y}_R) - Cov(\bar{y}_R, \bar{y}_U)$ and $A_u \triangleq V(\bar{y}_U) - Cov(\bar{y}_R, \bar{y}_U)$. In calculation, α_{CS} can be estimated by sample moments for A_m and A_u.

This shrinkage estimator was applied to estimate the mean scale scores for NAEP state assessments. The data used were the 1996 and 1992 NAEP state math assessments. The results in Table 2.3 show that the composite mean estimates had smaller mean square errors than the simple mean estimates.

Among these, those with minimum variance β and minimum MSE β in $\bar{y}_R$, for 7 of 9 states, had particularly small mean square errors while the bias introduced was trivial, though there wasn't much difference between minimum variance β and minimum MSE β.

APPENDIX

The sizes of two repeated samples are equal. Let $\mathbf{S}^k = \{\boldsymbol{S}_1^k, \boldsymbol{S}_2^k, ..., \boldsymbol{S}_T^k\}$, where $T = \begin{pmatrix} n \\ k \end{pmatrix}$, and $\mathbf{S}_i^k = (s_{i_1}, s_{i_2}, ..., s_{i_n})$ where s_{i_j} equals 1 if unit with subscript i_j was selected in the first sample, otherwise 0, and S_i^k has k s_{i_j}s that equal 1. In second sample, the number of sample units which were also selected in the first sample has a probability:

$$P(\xi = k) = \sum_{S_i^k \in S^k} \prod_{t=1}^{n} p_{i_t}^{s_{i_t}}(1 - p_{i_t})^{1-s_{i_t}}.$$

Define $\mathbf{S}^* = \{\boldsymbol{S}_1^*, \boldsymbol{S}_2^*, ..., \boldsymbol{S}_{T^n}^*\}$, where $T^n = \begin{pmatrix} N \\ n \end{pmatrix}$ and $\mathbf{S}_i^* = (s_{i_1}, s_{i_2}, ..., s_{i_n})$. Let $P_{i^{(1)}}$ be the probability that the range of overlapping for all k second sample units and the first sample is all the first sampled units except one; $P_{i^{(2)}}$ is the probability that the range of overlapping for all k second sample units and the first sample are all the first sampled units except two; and so forth.

$$\text{Define } S_1 = \sum P_{i^{(1)}} = \sum \left[1 - \frac{p_{i_j} |_{s_{i_j}^* \in S_i^*}}{\sum_{s_{i_j}^* \in S_i^*} p_{i_j}} \right]^k,$$

$$S_2 = \sum P_{i^{(2)}} = \sum \left[1 - \frac{(p_{i_j} + p_{i_k}) |_{s_{i_j}^* \in S_i^* \wedge s_{i_k}^* \in S_i^*}}{\sum_{s_{i_j}^* \in S_i^*} p_{i_j}} \right]^k, ...$$

so the probability that all the first sample units are overlapped is

$$P_0(k, n) = 1 - S_1 + S_2 - ....$$

Define $\mathbf{S}_i^{**} = \{\boldsymbol{S}_{i1}^{**}, \boldsymbol{S}_{i2}^{**}, ..., \boldsymbol{S}_{iT^m}^{**}\}$, where $T^m = \begin{pmatrix} n \\ m \end{pmatrix}$ and $\mathbf{S}_{ij}^{**} = (s_{i_{j1}}, s_{i_{j2}}, ..., s_{i_{jm}})$. Let η be the exact number of distinct units among these k units:

$$P(\eta = m \mid \xi = k) = \sum_{s_{i_j}^{**} \in S_i^{**}} \left[\sum_{s_{i_{jk}}^{**} \in s_i^{**}} p_{i_j} / \sum_{s_{i_j}^* \in s_i^*} p_{i_j} \right]^k p_0(k, m).$$

REFERENCES

Cochran, W. G. 1977. *Sampling Techniques,* 3rd ed. John Wiley & Sons, New York.

Cohen, T. and B. Spencer. 1991. "Shrinkage Weights for Unequal Probability Samples." *Proceedings of the Section on Survey Research Methods*, 625-630.

Feller, W. 1957. *An Introduction to Probability Theory and Its Application*, John Wiley & Sons, New York.

Fuller, W. 1990. *Analysis of Repeated Surveys*, Survey Methodology, 16, 2, 167-180.

Lent, J., Miller, S., and P. Cantwell. 1996. "Effect of Composite Weights on Some Estimates From the Current Population Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Qian, J., and B. Spencer. 1993. "Optimally Weighted Means in Stratified Sampling," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 863-866.

Table 1.2.A. The Average Number of Overlapping Units in PPS/wtr + PPS/wr ($n_1=n_2$) (Replication=500 in Simulation)

Sample size	Population size			
	100	200	500	1000
20	3.27 (1.4)	0.91 (0.6)	0.15 (0.2)	0.04 (0.1)
40	14.07 (2.4)	3.72 (1.1)	0.66 (0.3)	0.16 (0.1)
80	43.88 (4.1)	13.36 (1.9)	2.27 (0.6)	0.63 (0.2)
100	62.78 (7.0)	19.58 (2.5)	3.70 (0.7)	0.95 (0.3)

Table 1.2.B. The Average Number of Overlapping Units in PPS/wr + PPS/wr ($n_1=n_2$) (Replication=300 in Simulation)

Sample size	Population size			
	100	200	500	1000
20	3.80 (2.1)	1.14 (0.9)	0.18 (0.3)	0.04 (0.1)
40	11.89 (4.7)	4.12 (1.9)	0.68 (0.6)	0.17 (0.2)
80	30.57 (8.9)	11.93 (3.3)	2.58 (1.3)	0.75 (0.5)
100	39.14 (7.6)	16.06 (4.4)	3.83 (1.5)	1.03 (0.6)

Table 1.2.C. The Average Number of Overlapping Units in PPS/wtr + PPS/wtr ($n_1=n_2$) (Replication=500 in Simulation)

Sample size	Population size			
	100	200	500	1000
20	5.07	1.29	0.16	0.04
40	18.95	4.84	0.83	0.20
80	71.51	19.37	3.09	0.78
100	100.00	30.20	4.78	1.19

Table 2.1 The overlapping of schools between 1996 and 1992 NAEP state math assessments

State	N	n_1	n_2	# of Overlapping	ρ of scale score	ρ of 1st subscale score
DC	69	35	45	31	0.86	0.86
DE	60	28	40	26	0.67	0.49
HI	69	51	51	51	0.73	0.73
ME	189	80	92	50	0.38	0.23
NE	425	75	113	52	0.61	0.52
NH	113	69	68	41	0.42	0.34
RI	77	49	64	36	0.82	0.81
VT	72	82	91	76	0.53	0.43
WY	95	51	72	45	0.76	0.69

* Five subscales in the math assessments are number and operation, measurement, geometry, data analysis, and algebra.

* The overlapping is based on the available variable of school district.

Table 2.3. The composite means with different βs for 1996 NAEP state math assessments

State	Mean*	Std Dev	β = 1		β:min var		β:optimal	
			Comp. mean	RMSE	Comp. mean	RMSE	Comp. mean	RMSE
DC	233.4	15.9	233.4	15.9	233.4	15.9	233.4	15.9
DE	262.9	8.7	262.9	8.7	262.9	8.7	262.9	8.7
HI	256.3	12.1	256.3	12.1	256.1	10.3	256.1	10.4
ME	278.6	10.9	277.8	8.8	277.6	8.5	277.8	8.8
NE	278.1	12.3	278.0	9.3	277.9	8.8	278.0	9.4
NH	278.4	9.2	278.6	8.1	278.7	7.6	278.7	7.6
RI	266.5	16.8	266.2	15.1	264.7	11.8	264.9	11.8
VT	274.7	8.6	274.7	7.9	274.6	7.1	274.6	7.2
WY	274.7	10.5	274.7	10.5	274.7	9.8	274.7	9.7

* The means in this column are calculated by 1996 state math data only.