



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium  
Series - Proceedings**

**Symposium 2003: Challenges  
in Survey Taking for the Next  
Decade**

2003



Statistics  
Canada    Statistique  
Canada

Canada

Proceedings of Statistics Canada Symposium 2003  
Challenges in Survey Taking for the Next Decade

## AN ALTERNATE METHODOLOGY FOR ESTIMATING VARIANCE COMPONENTS

G. Hussain Choudhry, Gary Shapiro, Varma Nadimpalli, and Joseph Croos<sup>1</sup>

### ABSTRACT

Variance components are useful in allocating the sample to the various stages of sampling in a multi-stage sample designs. This paper discusses a regression model to estimate the variance components for a two-stage sample design. We use a simulation study to compare the bias and root mean square error of the estimated variance components using the traditional and the proposed regression approaches and conclude that the traditional is actually superior. The traditional approach estimates the within first-stage variance component and then computes the between first-stage variance component by subtraction from the total variance, and hence can produce negative estimates of the between first-stage variance component. The regression approach can also produce negative estimates of the between first-stage variance component. In both cases, the negative values of the estimated variance components are set to zero. Alternatively, for the regression approach the model can be estimated with the constraint that the model parameters be non-negative.

KEYWORDS: Between PSU Variance; Cost Model; Optimum Number of PSUs; Two-Stage Sample Design.

### 1. INTRODUCTION

In multistage design surveys, there is a frequent need to obtain components of variance in addition to simply estimating the variance over all stages of sampling. The need most commonly occurs when making sample design decisions about how many first stage and later stage units to sample. The traditional way of estimating the first stage variance component has been to estimate the total variance and the within-first-stage unit variance independently, and then obtain the between-first-stage-unit variance by subtraction. This typically results in unstable estimates of the variance components, with negative variance estimates for the between-first-stage component not a very uncommon situation. We examine in this paper a linear regression modeling approach to obtain components of variance. We believed that this approach would produce more stable variance estimates. After the presentation of the paper at the 2003 Symposium, Wayne Fuller questioned whether this was so, since the approach is based on using subsets of the full data set and can not bring in larger data sets. Because of Dr. Fuller's questions, we compared the bias and the root mean square of the estimates of variance components based on a large-scale simulation. The simulation shows that the traditional method is superior to the proposed regression approach.

### 2. DESCRIPTION OF THE APPROACHES FOR ESTIMATION OF VARIANCE COMPONENTS

The estimate of the total for a stratified two-stage sample is given by

$$\hat{Y} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{hij}$$

---

<sup>1</sup> G. Hussain Choudhry, Gary Shapiro, Varma Nadimpalli, and Joseph Croos, Westat, 1650 Research Boulevard, Rockville, Maryland 20850, USA.

where  $Y_{hij}$  is the  $Y$ -value for the sampled second-stage unit  $j$  from first stage unit  $i$  from stratum  $h$  and  $W_{hij}$  is the corresponding sampling weight. The total variance of the estimate  $\hat{Y}$  can be written as:

$$V = (1 - f_1) \frac{S_1^2}{n} + (1 - f_2) \frac{S_2^2}{nm}$$

where

$$\begin{aligned} (1 - f_1) \frac{S_1^2}{n} &= \text{between PSU variance, and } (1 - f_2) \frac{S_2^2}{nm} = \text{within PSU variance;} \\ n &= \text{average number of sampled PSUs per stratum;} \\ m &= \text{average number of sampled households per PSU;} \\ f_1 &= \text{sampling fraction at the first stage; and} \\ f_2 &= \text{sampling fraction at the second stage.} \end{aligned}$$

In the traditional approach both the total variance and the within PSU variance component for a stratified two-stage sample design are estimated from the sample. The between PSU variance is then estimated from the estimated total variance by subtraction. The problem with estimating the between PSU variance by subtraction is that it is usually unstable and can produce negative estimates, especially when the between PSU variance component is small.

Baskin (1992, 1993) used hierarchical Bayes and Gibbs sampling to estimate certain components of the variance for the United States Consumer Price Index (CPI). Baskin and Johnson (1995) and Shoemaker (2001 and 2002) estimated components of variance of rent and commodities in the U.S. CPI by Restricted Maximum Likelihood (REML) and the usual ANOVA type estimators. It was seen that the REML methods produce nonnegative estimates of components of variance whereas the ANOVA estimates would produce negative estimates of some components. Leaver and Larson (in publication 2003) estimated components of variance for CPI using a weighted REML method and contrasted these with those obtained using two different random effects models. Choudhry, et al (1985) ensured nonnegative estimates of the components of variance for Canadian Labour Force survey that is multistage design by estimating these using census data.

In this paper we propose a linear regression model to estimate the components of variance for a stratified two-stage sample design. Although the proposed regression approach can produce negative estimates of the components of variance, it can be modified with non-linear programming so that the estimates are constrained to be nonnegative. The ratio  $S_1^2/S_2^2$  cannot be directly calculated. We used variance-component model to estimate the two components of variance and then estimate the ratio from the estimated variance components.

The variance-component model for the estimate of total is:

$$V(\hat{Y}) = (1 - f_1) \frac{S_1^2}{n} + (1 - f_2) \frac{S_2^2}{nm} + \epsilon,$$

and our interest is in estimating the parameters  $S_1^2$  and  $S_2^2$ . We use the full sample to create 55 mini samples as described in Section 3.3 to estimate the parameters  $S_1^2$  and  $S_2^2$  in the above regression model. We set  $X_1 = (1 - f_1)/n$  and  $X_2 = (1 - f_2)/nm$ , and regress  $V(\hat{Y})$  on  $X_1$  and  $X_2$  to estimate the parameters  $S_1^2$  and  $S_2^2$  in the regression model. The variance components can be constructed from the estimated values of the parameters  $S_1^2$  and  $S_2^2$ .

### 3. SIMULATION METHODOLOGY

We used four artificial finite populations to select samples repeatedly with a stratified two-stage sample design with simple random sampling at both stages of sampling. We estimate the components of variance from each sample using the traditional and regression approaches and compare the two approaches with respect their biases and root mean square errors.

#### 3.1 Finite Populations

We constructed four artificial populations from which to draw repeated samples. Each population consists of 10 strata, 40 PSUs in each stratum, and 100-second stage units (such as households) in each PSU, for a total of 40,000-second stage units. The means for each stratum in all four populations are as indicated in Table 1.

**Table 1. Means by stratum**

Stratum	1	2	3	4	5	6	7	8	9	10
Mean	820	860	900	940	980	1,020	1,060	1,100	1,140	1,180

For the first population, PSUs were selected from a normal distribution with standard deviation 8.28. Thus, for example, in stratum 1, 40 PSUs were selected from the distribution  $N(820, 68.56)$ , where  $68.56 = (8.28)^2$ . Since the set of 40 PSUs was randomly selected, the mean value for the PSUs from stratum 1 is not exactly 820 although the expected value for the mean was 820.

The four populations differ in that the standard deviations differ for the PSUs. Table 2 shows the standard deviations for the populations.

**Table 2. Standard deviations (planned) among PSUs by population**

Population	1	2	3	4
Standard deviation	8.28	13.10	18.57	23.89

Once the population of PSUs was determined for each of the 10 strata in each of the four populations, selection of the second stage units in the population was possible. One hundred second stage units were selected for each PSU from a normal distribution with mean equal to the PSU mean and standard deviation 100. (The exact same standard deviation of 100 was used for all PSUs in each population.)

The achieved populations had between PSU variances as a proportion of total variance as indicated in Table 3.

**Table 3. Between PSU variance as proportion of total variance**

Population	1	2	3	4
Proportion between variance	12.43%	19.73%	29.45%	39.02%

#### 3.2 Simulation Samples

We selected 2,000 simulated samples with stratified two-stage sample design with simple random sampling at each stage of sampling. Note that each simulated sample was selected from four similar but different finite populations. The four finite populations were drawn from four super-populations as described above. For the purpose of our simulation study, we treat these finite populations as fixed.

For each simulation run, we selected 4 PSUs per stratum, and 10 households per PSU with simple random sampling (SRS) at both stages of sampling. Thus, each simulation sample consisted of 400 households—10 households per PSU, 4 PSUs per stratum, and 10 strata. We selected 2,000 simulation samples independently.

### 3.3 Mini Samples for Regression Approach

We created 55 mini samples from each simulation sample and used these mini samples for estimating the within and between components of the variance with the regression model as described in Section 2. Let P1, P2, P3, and P4 be the generic notation for the 4 PSU's in any stratum from the main sample. For example, in a given simulation, the PSU P1 in a stratum may be "02", whereas in the same iteration the PSU P1 in another stratum may be "06." For each stratum, there are 11 possible subsets of PSUs containing all 4 sampled PSUs, 3 out of 4 sampled PSUs, and 2 out of 4 sampled PSUs. For the simulation, we created subsamples from the original sample such that there were the same numbers of PSUs in each stratum. For each of the 11 subsets of the sampled PSUs defined above, we created 5 subsets of households as follows:

1. All 10 sampled households in each PSU;
2. Random half of the 10 sampled households (5 households per PSU);
3. Other half of the 10 sampled households (5 households not selected in step 2 above);
4. Randomly selected seven households out of the 10 sampled households per PSU; and
5. Remaining three households not selected in step 4 above.

Thus, we generate 55 mini samples from each simulation sample—one of these mini samples is the actual simulation sample.

The models fitted very well for all four populations. For the first population, the adjusted R-squared values varied from 0.64 to 0.95, with an average of 0.87. The adjusted R-squared values varied from 0.67 to 0.96 for the second population, it varied from 0.69 to 0.96 for the third population and from 0.70 to 0.96 for the fourth population. The average R-squared values for the second and third populations were the same as for the first population (0.87), and the average was 0.88 for the fourth population.

## 4. SIMULATION RESULTS

Our focus here will be on the ratio of between PSU variance to total variance. We also examined the estimates of within PSU variance and of between PSU variance. All methods produced comparable and good estimates of within PSU variance. The pattern for between PSU variance was generally similar to that for the ratio of between PSU variance to total variance.

We examined three versions of the traditional approach. The simplest version is to directly use the sample estimate, including negative values of between PSU variance. Especially for population 1, where the true between PSU variance is relatively low, there were many instances of negative values—709 of the 2,000 samples. Table 4 gives the number of negative values for each population, for the unadjusted regression estimate as well as the unadjusted traditional approach. The second version of the traditional approach is to change the negative values to zero, i.e., the estimated ratio is 0 percent. The third approach is to make a bias correction in addition to changing negative values to zero. Cochran (1977, p. 278) and other authors have shown that  $s_1^2/n$ , the sample estimate of the variance among first stage units, is not an unbiased estimate of  $V(\hat{Y})$ . Thus, the third version makes an adjustment for this fact.<sup>2</sup> The first version has much higher root mean square errors than the other two versions. The third version has only slightly lower mean square errors than the second version. We make comparisons here to the second version.

---

<sup>2</sup> The authors plan to write a paper for the 2004 Joint Statistical Meetings discussing the bias of the ultimate cluster variance as an estimate of total variance.

**Table 4. Frequency of negative estimates of between PSU variance**

Population	1	2	3	4
Unadjusted traditional	709	471	240	90
Unadjusted regression	812	636	385	217

As we discussed in Section 2, we also had two versions of the regression modeling approach (one with negative values changed to zero and one with constraints that prevented negative values). The two versions produced nearly identical results. Table 5 shows that the root mean square error (RMSE) is substantially higher for the regression approach than for the traditional approach for all four populations. The standard error dominates the RMSE. Table 6 shows that the standard error is also consistently higher for the regression approach. However, the relative bias is slightly lower for the regression approach for the third population. Note that the bias is usually negative for the traditional approach even though we increase the estimated ratio to zero when the sample estimate is a negative value.

**Table 5. RMSE for traditional and regression approaches**

Population	1	2	3	4
Traditional	0.146	0.160	0.171	0.166
Regression	0.220	0.209	0.219	0.221

**Table 6. Bias and Standard error for traditional and regression approaches**

Population	Bias				Standard error			
	1	2	3	4	1	2	3	4
Traditional	.093	-.048	-.075	-.058	0.145	0.160	0.169	0.165
Regression	.376	.080	-.037	-.061	0.194	0.208	0.219	0.220

## 5. DISCUSSION OF RESULTS AND CONCLUSIONS

The traditional approach to estimating the between first-stage variance in a two-stage sample design is not very satisfactory, usually resulting in unstable estimates that can frequently be negative or much too large. We developed what we thought was a superior regression-based method, but found it to be subject to significantly larger standard errors and root mean square errors than the traditional approach. We believed that the use of a number of mini-samples (55 were used in the simulation) would provide more stable estimates of variance components. However, since the mini-samples are subsets of the original sample, they apparently do not add any information and do not lead to better estimates. We therefore recommend against use of the new methodology.

We have not evaluated the various methods of Bureau of Labor Statistics staff (Baskin, Shoemaker and others) that we briefly cited in Section 2 of the paper, and so offer no opinion as to whether these may yield better estimates of variance components than does the traditional approach.

## REFERENCES

- Baskin, R.M. (1993). Estimation of variance components for the U.S. Consumer Price Index via Gibbs sampling, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 808-813.
- Baskin, R.M. (1992). Hierarchical Bayes estimation of variance components for the U.S. Consumer Price Index, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 716-719.

- Baskin, R.M., and Johnson, W.H. (1995). Estimation of variance components for the U.S. Consumer Price Index, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 126-131.
- Choudhry, G.H., Lee, H., and Drew, J.D. (1985). Cost-variance optimization for the Canadian Labor Force Survey, *Survey Methodology*, 11, pp. 33-50.
- Cochran, W.G. (1977). *Sampling Techniques (3<sup>rd</sup> Ed.)*. New York: John Wiley and Sons.
- Leaver, S.G., and Larson, W.E. (in publication 2003). Estimating Components of Variance of Price Change from a Scanner-Based Sample, *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Shoemaker, O.J. (2002). Estimation and Analysis of Variance Components for the Revised CPS Housing Sample. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Shoemaker, O.J. (2001). Estimation of Variance Components for the U.S. Consumer Price Index: A Comparative Study, *Proceedings of the Survey Research Methods Section, American Statistical Association*.