VARIANCE ESTIMATION FOR THE 1992 PUERTO RICO CENSUS OF AGRICULTURE[1]

Richard Griffiths and Inez Chen, Bureau of the Census
Richard Griffiths, Bureau of the Census, Suitland, MD 20233

Key Words: One primary unit per stratum, Collapsed strata, Bias, Precision

For the 1992 Puerto Rico Census of Agriculture, a sample was used to represent the smallest farms and to supplement the list of larger farms. The sample design called for the selection of one primary unit from each "stratum." In order to estimate variances for this portion of the design, the common technique of collapsing strata was employed. We used data from the 1987 Puerto Rico Census of Agriculture — a complete enumeration — in an empirical study to investigate several different variance estimators for the collapsed strata design and to address issues related to the biases and precision of the different variance estimators. This paper reports the results of the empirical study and compares the variance estimators.

I. Introduction

For the purposes of statistical estimation for the Census of Agriculture, Puerto Rico is divided into five mutually exclusive regions and 77 municipios, each of which is wholly located in one of the five regions. The municipios are subdivided into enumeration districts (EDs).

The 1992 Puerto Rico Census of Agriculture sample comprised three separate parts: an attempted complete enumeration of certainty farms; an area sample of noncertainty farms in noncertainty EDs; and an area sample of certainty mail EDs.

Noncertainty EDs were those EDs identified as having at least four noncertainty farms in the 1987 Census of Agriculture; all other EDs were classified as certainty mail EDs. Certainty farms were those farms either identified as large from the 1987 Census or expected to be large based on information obtained from sources such as the Puerto Rico Department of Agriculture. Certainty farms could appear in either noncertainty or certainty mail EDs. All other farms in noncertainty EDs were classified as noncertainty farms. In the certainty mail EDs those farms identified from the 1987 Census but not large enough to be certainty farms were classified as noncertainty farms. Farms in certainty mail EDs not otherwise identified as certainty or noncertainty farms were classified as new farms.

In this paper we will be concerned with the method used to estimate variances for totals estimated from the noncertainty ED area sample of farms. In this area sample a single cluster was chosen from each municipio, and the potential farms in this cluster were enumerated.

The clusters in each municipio were groups of EDs. The clusters had been formed by studying the intracluster correlation coefficients of particular items for different ED groupings. These intracluster correlation coefficients were calculated from the data obtained in the 1987 Puerto Rico Census of Agriculture. ED groupings which resulted in small intracluster correlations were chosen for the formation of clusters for the 1992 Census.

The most direct method of estimating variances for this sample design would have been to use a formula for a cluster sample; unfortunately, due to the selection of only one cluster from each municipio, calculation of the necessary intracluster correlations based on the 1992 sample data was impossible.

Some other possibilities examined for the calculation of these variances included the use of a design effect and a collapsing method. The method chosen was the collapsing method.

When two strata are collapsed in order to estimate variances, a between-stratum variance is added to the variance estimator; this between-stratum variance contributes to an overestimation of the true variance. In order to help reduce this positive bias, strata are generally collapsed based on similarity of totals; thus, we used similarity of total values of agricultural products sold from the 1987 Census as the criterion for grouping municipios.

This grouping was done within region. In four of the regions there was an even number of municipios; thus, the municipios were paired. In the fifth region the number of clusters was odd; thus, there was one group of three municipios in this region. The remainder of this paper presents formulas only for the cases in which two municipios were paired.

For each pairing, then, there were two observations — two sampled clusters. These two observations were used to produce an estimate of the variance for each

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

municipio in the group.

In this paper we report the results of an empirical study designed to examine the collapsing method of variance estimation. Data from the 1987 Puerto Rico Census of Agriculture, a complete enumeration, were used to conduct this empirical study.

In the empirical study we investigate the accuracy of three variance estimators: the one derived for the 1992 Puerto Rico Census of Agriculture; a formula given by Hansen, Hurwitz, and Madow (1953); and one credited by Hartley, Rao, and Kiefer (1969) to Seth (1966). This study serves as an after-the-fact evaluation of the variance estimation method actually chosen for the 1992 Census. The formulas given by Hansen, Hurwitz, and Madow and Seth were not considered for use as variance estimators in the 1992 Census; however, this study allows us to examine not only the chosen variance estimator in an absolute sense by comparing it to the true variance, but also in a relative sense by comparing it to other variance estimators. Such information could be useful for the next Puerto Rico Census of Agriculture or another survey or census with a similar design.

II. The Variance Estimation Formulas

The three variance estimation formulas, along with their defining notation, are as follows:

The PR formula:

$$V(\hat{X}_i) = \left(1 - \frac{1}{C_{gi}}\right) C_{gi}^2 \cdot \sum_{h=1}^{2} \left(x_{gh} - \frac{x_{g1} + x_{g2}}{2}\right)^2$$

This formula is the one used to estimate municipio-level variances for the 1992 Puerto Rico Census of Agriculture.

Basically, this formula consists of a sum of squares — the sum of the squared differences of the cluster totals, x_{g1} and x_{g2}, and their average over the two collapsed municipios. This sum of squares is based on the unweighted sampled cluster totals.

The mean used in this formula is unweighted. By use of this unweighted mean, we are explicitly assuming the equality of the collapsed municipios' means.

The sum of the squares is multiplied by the square of the total number of clusters, C_{gi}^2, in the municipio for which we're estimating variances; this gives us the variance estimator for a municipio-level total.

Finally, the formula is multiplied by a finite population correction (fpc). This fpc should not have been included, since the design calls for the selection of only one unit per municipio; however, the fpc was erroneously included in calculations for the 1992 Puerto Rico Census of Agriculture and, thus, we will include it in this paper since we are concerned with evaluating the actual formula used.

HHM's formula:

$$V(\hat{X}_i) = \left(1 - \frac{1}{C_{gi}}\right) \sum_{h=1}^{2} C_{gh}^2 \left(x_{gh} - \frac{C_{g1}x_{g1} + C_{g2}x_{g2}}{C_{g1} + C_{g2}}\right)^2$$

Seth's formula:

$$V(\hat{X}_i) = \left(1 - \frac{1}{C_{gi}}\right) \sum_{h=1}^{2} C_{gh}^2 \left(x_{gh} - \frac{C_{g1}x_{g1} + C_{g2}x_{g2}}{2\,C_{gh}}\right)^2$$

Both of these formulas use a weighted mean in the sum of squares, although the weight is different for each.

HHM's formula was designed to make use of auxiliary information: an auxiliary variate highly correlated with municipio totals. In our study we used the number of clusters as the auxiliary variate. So, HHM's formula more closely resembles the other two formulas than it might otherwise.

For some of the items examined in the empirical study, the number of clusters was fairly highly correlated with the item totals. For others it wasn't. The implications of this will be discussed in more detail later in the paper.

Another difference between the PR formula and the other two is the use of a weighted mean in the sums of squares of HHM's and Seth's formulas.

Also, HHM's and Seth's formulas were both designed to provide variance estimates at the aggregate (region) level. The PR formula was specifically designed to yield estimates at the municipio level.

III. Empirical Study

The accuracy of the PR formula was evaluated through the use of an empirical study. This empirical study provided both relative and absolute comparisons of the PR formula. Comparisons of the PR formula to HHM's and Seth's formulas produced a relative evaluation. Comparison of the three formulas to the true variance provided an absolute evaluation.

The empirical study was based on the use of data from the 1987 Puerto Rico Census of Agriculture, a complete enumeration. For the purposes of this empirical study, we assumed that the 1987 Puerto Rico Census of Agriculture data were generated from the complete and accurate response of all farms in the universe. Any imputed data were assumed to be the data that would have been reported had we obtained complete and accurate response. This assumption was necessary for us to calculate true variances and expected values of the three variance estimators to be compared.

A. Methodology

Using the 1992 Census cluster designations, we calculated item totals for each cluster in each municipio using the 1987 Census data. The point estimators of municipio totals were based on a simple random sample of a single cluster per municipio in the 1992 Census. It was thus a simple step to replicate all possible samples and calculate the true variances of municipio-level total estimators using the 1987 data; thus, we were able to calculate true variances for estimated totals for each of the 24 items in each of the 77 municipios.

Furthermore, we were able to construct municipio-level variance estimates for all possible samples using each of the three formulas. Averaging over all possible samples for the 24 items in each of the 77 municipios gave us the expected values of the variance estimators. We then calculated biases and mean square errors of the three variance estimators.

Region-level true variances and biases and mean square errors of the variance estimators were also calculated for the 24 items for each of the five regions.

The following section gives some of the results of the empirical study.

B. Results

Table 1 of the attachment displays the results of a relative comparison of the three variance estimators for municipio-level variance estimates. This table provides the frequency of estimates for which each of the three variance estimators produced the minimum absolute bias. From this table we see that, in general, the PR formula tended to estimate variances with the smallest absolute bias, though, all three formulas did have the minimum absolute bias for a good number of items.

Table 2 provides the frequency of municipio-level variance estimates for which each of the three variance estimators produced the minimum mean square error. From this table we see that all three variance estimators are more even: each provided the variance estimator with the minimum mean square error for a large number of cases. The fact that HHM's formula closed the gap somewhat on the other two formulas, especially on the PR formula, is an indication that the variance of this variance estimator is probably smaller than the variances of the other two variance estimators, offsetting some of the advantage the other two had in terms of bias.

Table 3 provides another comparison of the variance estimators based on the estimation of municipio-level variances. This table gives the percentage of cases for which each of the formulas yielded variance estimates with absolute biases less than 50 percent. Again, the results appear similar for all three formulas: all have biases of less than 50 percent for between 30 and 40 percent of the items. Although, it does appear that the PR and Seth's formula have a slight advantage.

With Tables 4 and 5 we turn to the results of comparisons for the estimation of region-level variances. In Table 4 we have the frequency of cases for which each of the three variance estimators produced the minimum absolute bias. It appears that the PR formula does not do as well as the other two in the estimation of region-level variances and that Seth's formula often produces variance estimates with smaller absolute biases than HHM's. This is to be expected, though, since the PR formula was specifically designed to estimate municipio-level variances, while the other two formulas were designed to estimate variances at the region level.

Table 5 gives more region-level results. The results here are the percent of items for which each of the variance estimators had an absolute bias of less than 50 percent. A clear winner is not obvious. Each formula estimates variances with less than 50 percent absolute bias for between 70 and 80 percent of the items. This indicates that the differences among the three formulas for region-level variances may not be as great as would seem from Table 4, though, HHM's and Seth's formulas do have a small advantage.

C. Limitations

Before offering our conclusions, there are two study limitations which should be mentioned. Both have been alluded to previously in this paper.

As mentioned earlier, HHM's formula was designed to make use of an auxiliary variate. The variable we use as the auxiliary variable, number of clusters, was fairly highly correlated with the municipio totals for some items. For other items, the correlations were very low. Perhaps, the use of a variable more highly correlated with municipio totals would have produced better results for HHM's formula.

Also, we need to study the effect of erroneously including the finite population correction in the PR formula. Exclusion of the fpc would not affect any of the relative comparisons, since all three of the formulas included the same fpc in this study. It would, however, affect the absolute comparisons.

We know that for most of the items, the variance estimators displayed a positive bias. It then follows that if we excluded the fpc, in general, the absolute biases of the variance estimators would increase. It also follows that any negative biases would become larger, with some even becoming positive biases. This, however, could be viewed in a positive light, since for variance estimation we generally prefer positive biases to negative ones.

IV. Conclusion

The general conclusion from the empirical study is somewhat unclear. It is hard to say which of the variance estimation methods would serve our purposes best. No one formula is dominant in any of the measures, i.e. bias, MSE, frequency of bias less than 50%.

It might be argued, though, that the PR formula, in comparison to the other two formulas, performs better when estimating municipio-level variances than it does when estimating region-level variances. This seems plausible since the PR formula was designed to estimate variances at the municipio level, while the other two formulas were designed to estimate variances at the region level. So, we could conclude that it would be better to use the PR formula to estimate municipio-level variances and one of the other two formulas to estimate region-level variances.

Another argument is that while the PR formula may appear to have a slight advantage at the municipio-level, the magnitude of this advantage is not large. So, it wouldn't seem completely inappropriate to suggest that we use either HHM's or Seth's formula to estimate both municipio- and region-level variances. In fact, if we were to find an auxiliary variable more highly correlated with stratum totals, HHM's formula might prove to be superior to the other two formulas.

It could also be argued that none of the formulas estimates variances particularly accurately, especially at the municipio level. With more than half of the municipio-level variance estimates having biases of at least 50%, we may question whether we want to use any of the variance estimators under study here. (Note that while we are discussing the poorness of the variance estimates here and, in general, the positive biases of these estimates, the estimated coefficient of variation for the overall farm count for Puerto Rico was 4.7%.) What could be done to improve the variance estimators?

One possibility is to design the sample so that we needn't use the collapsing method: we could select two clusters per municipio rather than one. This would eliminate the problem of the biased variance estimator. Of course, it might be argued that the point estimators based on a sample of two clusters, each of size n/2, would be less precise than those based on a sample of one cluster of size n. This loss in precision would probably be minimal, but do we want to compromise the precision of our point estimators — no matter how minimally — just to allow better variance estimators?

If we use a similar sample design for the 1997 Puerto Rico Census of Agriculture — that is, a design which calls for the selection of one cluster per municipio — we may want to look into other methods of estimating variances. Two possibilities are 1) A method given in Hartley, Rao, and Kiefer (1969) which doesn't require collapsing. Since this method requires use of a concomitant variable(s), we could also retest HHM's formula using this concomitant variable if it were more highly correlated with municipio totals than number of clusters; 2) A design effect. As we have seen in this paper, the 1987 variances can be calculated (almost) exactly. If we were to account for the change in design and if we knew something about the change in distributions of the characteristics over the years, we might be able to effectively implement a design effect approach.

V. Acknowledgement

The authors would like to thank Toni Wooten for running the programs which produced the empirical study results.

BIBLIOGRAPHY

Cochran, William G. (1977), *Sampling Techniques*, Third Edition, John Wiley and Sons.

Hansen, Morris, William Hurwitz, and William Madow (1953), *Sample Survey Methods and Theory, Volume I*, John Wiley and Sons, Inc.

Hartley, H.O., J.N.K. Rao, and Grace Kiefer (1969), "Variance Estimation with One Unit Per Stratum," Journal of the American Statistical Association, 64, pp. 841-851.

Kalton, Graham (1977), "Practical Methods for Estimating Survey Sampling Errors," Bulletin of the International Statistical Institute, 47(3), pp. 495-515.

Seth, G.R. (1966), "On Collapsing Strata," Journal of the Indian Society of Agricultural Statistics, 18, pp. 1-3.

Table 1 Municipio-level Variances
Minimum Absolute Bias
Number of Items*

HHM	Seth	PR
184	353	391

*115 cases not included in table for which Seth's and PR formula were tied and both had an absolute bias less than HHM's.

Table 2 Municipio-level Variances
Minimum Mean Square Error
Number of Items*

HHM	Seth	PR
283	358	339

*115 cases not included in table for which Seth's and PR formula were tied and both had a mean square error less than HHM's.

Table 3 Municipio-level Variances
Absolute Bias
Percent with Bias < 50%

HHM	Seth	PR
33.7	39.8	39.2

Table 4 Region-level Variances
Minimum Absolute Bias
Number of Items

HHM	Seth	PR
39	68	13

Table 5 Region-level Variances
Absolute Bias
Percent with Bias < 50%

HHM	Seth	PR
79.2	76.7	71.7