Estimating Upper and Lower Bounds of Census Nonsampling Variance

Paul P. Biemer, Bureau of the Census

1. INTRODUCTION

How good are the census data? Since 1950, the Census Bureau has attempted to provide data users with an answer to this question. Studies have been conducted to evaluate census operations such as interviewing, coding, editing, keying, microfilming, and tabulating. The methods most often used for evaluation require either some form of replication of the operation or interpenetration of the operator assignments. For large scale surveys, these methods are expensive to implement and difficult to control due to their complexity. Furthermore, since their purpose is usually evaluation of operations, these studies usually focus on only a few operations in the data collection and processing chain. As a consequence, estimation of total census error is not possible unless it is assumed that the operations not evaluated do not contribute significantly to total error.

The direction of the present research is toward methods for estimating total census variance which do not require expensive experimental designs or reinterview surveys. By taking advantage of the complete geographic coverage of the census, the estimation method developed in the paper produces a narrow range of error which, in expectation, contains the total nonsampling variance of the census total under an assumed model. These lower and upper bounds could serve as indicators of census data quality; or perhaps their utility could be extended, for example, (1) to determine which census items are most effected by processing error, (2) to compare, say, a decentralized data collection procedure with a centralized one, and (3) to form rough estimates of the total census variance.

The present paper describes the general approach of the method which was tested for feasibility in the 1980 Census. At the time of this writing, the computer results are not yet available; however, a subsequent paper will give the results of that study.

2. DESCRIPTION OF THE 1980 CENSUS

It is useful, first, to describe the data collection and processing operations of the 1980 Census. The 1980 Census data collection operation in mailout/mail back areas was handled by 375 District Offices (DO's). Eighty-seven were centralized offices headed by permanent Census Bureau employees. In these offices, which were responsible for major urban centers, the questionnaires were checked in, edited and followed up by clerical staff. Enumerators responsible for the personal follow-up were each assigned areas of about 350 housing units called enumeration districts (EDs). They used the office as their base of operations.

The remaining 288 offices in these areas used a decentralized office procedure and were headed by locally recruited people. Procedures in these offices involved turning over more responsibility for checking and editing questionnaires to the enumerators who worked out of their homes. Furthermore, there was no editing of nonmail returns and the follow-up of nonresponse was done more informally by the enumerators. As in centralized offices, the operations were performed by census personnel on an ED basis.

In other areas, the collection operation was handled by 24 offices using the conventional procedure and 12 offices using a combination of the decentralized and conventional procedures. This conventional system was used in more sparsely settled sections of the country. Households were mailed census questionnaires and were asked to hold them for pick up by census enumerators. Systematically canvassing their EDs, the enumerators picked up the questionnaires and followed up nonresponse at the same time.

The processing of the census data was done in three processing sites located in Jeffersonville, Indiana, New Orleans, Louisiana and Laguna Niguel, California. The principal operations performed included (1) receiving, sorting and storing the questionnaires, (2) clerical coding of written responses, (3) microfilming all questionnaires, (4) converting responses to machine-readable form and (5) clerically reviewing the collected information for each ED. Again, the operations were completed on an ED by state basis.

The remainder of the paper is concerned with the assessment of combinations of effects. That is, we attempt to estimate the joint impact on total variance of (1) all operations in a district office which are performed on whole EDs (such as enumeration) and (2) all other operations in a district office (check-in, editing, second phase follow-up, etc.). In the analysis of the experimental data, the effects, (1) and (2), will be estimated separately for centralized and decentralized district offices by the procedure described below. By the nature of the operations in each type of office, (1) is expected to be larger for centralized offices while (2) should be larger for decentralized offices. A test of these hypotheses will provide some evidence as to the validity of the proposed estimation procedure.

3. THE THEORETICAL BACKGROUND

3.1 The Basic Idea

The idea forming the basis of this methodology is similar to the collapsed strata concept of variance estimation ([2], p.138). Under additive model assumptions, the effects of the respondent, the enumerator, the joint effects of the district office operations and the joint effects of the processing center operations can be be roughly estimated by functions of the following contrasts:

(i) between EDs within the same district, (ii) between EDs within different districts but processed in the same processing centers, and (iii) between EDs within different districts and different processing centers.

In the present paper, the technique is developed for estimating (i) and (ii); (iii) has also been considered in another paper (see [1]). Furthermore, it will be shown that forming these contrasts only between contiguous EDs results in bounds having a smaller expected range.

The success of this technique depends to a great extent on the number of pairs of contiguous EDs handled by different processing centers and lying in different districts of the same type. Because no planning was done for this study prior to the census, there are very few contrasts of the form in (iii) for some types of district offices - in particular, for conventional district offices. However, if this method proves to be successful for centralized and decentralized district offices, future censuses could ensure the estimability of the nonsampling error components by purposefully creating more ED pairs of the form (ii) and (iii) in the census planning stage.

3.2 Notation and Assumptions

To simplify the exposition of the theory, consider a census operation in which all operations of data collection and processing are centralized in the district offices.[1/] As in the 1980 Census, assume that some operations, such as enumeration, are performed on an ED basis while other operations cut across EDs.

Our model assumes that the error that exists in the census arises from three general sources: (i) the respondent (or, generally, the elementary unit), (ii) the ED, and (iii) all operations in the district office which are not ED-based.

We, consider the case where the elementary unit is a person within a household whose response error may be correlated with other persons within the household. Let the subscript (i, j, k, ℓ) denote the ℓ-th person in the k-th household within the j-th ED within district office i. Denote by $y_{ijk\ell}$ the final tabulated value for unit (i, j, k, ℓ) and let $\mu_{ijk\ell}$ denote the true value for the unit. It is assumed that (i), (ii), and (iii) above contribute the following additive nonsampling errors to the true value of unit (i, j, k, ℓ):

$$\text{District office error} = \delta_i + \overset{*}{\delta}_{ijk\ell}$$

$$\text{ED error} = \alpha_{ij} + \overset{*}{\alpha}_{ijk\ell}$$

$$\text{Respondent error} = r_{ijk\ell}$$

where

δ_i = error variable arising from the i-th district office which is a combined effect of office operations other than those based upon EDs; may be interpreted as the average systematic error of these operations.

α_{ij} = error variable contributed by ED (i, j) common to all units in this ED,

$r_{ijk\ell}$ = error variable associated with respondent (i, j, k, ℓ) and,

$\overset{*}{\delta}_{ijk\ell}$ and $\overset{*}{\alpha}_{ijk\ell}$ = uncorrelated errors contributed by district office i and ED (i, j) for unit (i, j, k, ℓ).

Then, the full model is

$$y_{ijk} = \mu_{ijk\ell} + \delta_i + \alpha_{ij} + \varepsilon_{ijk\ell} \qquad (3.2.1)$$

where $\varepsilon_{ijk\ell} = \overset{*}{\delta}_{ijk\ell} + \overset{*}{\alpha}_{ijk\ell} + r_{ijk\ell}$.

The assumptions made for the model are:

(i) the errors δ_i and α_{ij} are random samples from infinite populations of errors with zero means and variances given by σ_δ^2 and σ_α^2, respectively.

(ii) the errors $\varepsilon_{ijk\ell}$ are random variables with mean 0, variance σ_ε^2 (i, j) and covariance $\mathrm{Cov}(e_{ijk\ell}, e_{i'j'k'\ell'}) = \rho_{ij}\sigma_\varepsilon^2$ (i, j) for $(i, j, k) = (i', j', k')$ and $= 0$, otherwise.

(iii) the errors δ_i, α_{ij}, and $\varepsilon_{ijk\ell}$ are uncorrelated with each other and with the true true values $\mu_{ijk\ell}$.

Now consider the variance of a census total, Y, under these assumptions. Let N denote the size of the total population of units and define

m_i = number of units in district office i,

m_{ij} = number of units in ED (i, j), and

m_{ijk} = number of units in household (i, j, k).

From (3.2.1), Y can be written as

$$Y = N\bar{\mu} + \sum_i m_i \delta_i + \sum_{ij} m_{ij}\alpha_{ij} + \sum\sum_{ijk\ell}\varepsilon_{ijk\ell}$$

where $\bar{\mu}$ is the true mean of the population. Hence, from the previous assumptions,

$$V(Y) = \sigma_\delta^2 \sum_i m_i^2 + \sigma_\alpha^2 \sum\sum_{ij} m_{ij}^2 + \sum\sum_{ij} m_{ij}\left[\sigma_\varepsilon^2(i,j)\,(1 + c_{ij}\rho_{ij})\right] \qquad (3.2.2)$$

where $c_{ij} = \sum_k m_{ijk}^2/m_{ij} - 1$.

3.3 Appropriateness of the Model

Perhaps the most striking feature of the proposed model is its simplistic representation of the multitudinous and complex nonsampling errors of a census. A natural question at this stage of the theoretical development is whether such a simple model is appropriate for the purposes of estimation or whether a more accurate representation of the census errors is needed.

For example, it may not be realistic to assume independence of the errors $\varepsilon_{ijk\ell}$ within a district office for some operations such as clerical editing where editors may make errors which are correlated. Or, the assumption that the δ_i are sampled from the one distribution may not hold and this will have implications for the interpretation of the estimates.

It should, therefore, be emphasized that our current research objectives are simply to determine whether the proposed estimation method, or a similar method, (a) yield measures which are, at least, an indication of magnitude of nonsampling error in census data and (b) is feasible for implementation in future censuses.

Upon satisfying these objectives, future work will be concerned with building models which better aid us in the formulation and interpretation of the estimates. The focus in this paper is, therefore on the estimation technique itself. The model is used only as a rough guide for interpreting the estimators and for combining the component estimators to create the lower and upper bounds.

4. ESTIMATION OF UPPER AND LOWER BOUND ON V(Y)

4.1 Derivation of V_L and V_U

Two EDs will be called adjacent if their boundary lines connect at some point. Two EDs may be adjacent and lie within the same district office boundary. These pairs of EDs will be referred to as EDs of type A. ED pairs of type B are defined as adjacent EDs which lie in two different census districts. Let A denote the set of all pairs of adjacent EDs of the type A and let B denote the set of all pairs of adjacent EDs of the type B.

To simplify the notation, the symbol $\underset{h\in S}{\text{Avg}}\, d_h$ will denote the simple expansion mean of a characteristic d for all elements in some specified set S.

Hence, the mean of the characteristic y for ED (i,j) will be denoted as $y_{ij} = \underset{h\in S}{\text{Avg}}\, y_h$ where S is the set of all units in the ED. Similarly, the true mean of the population for ED (i,j) will be denoted by $\mu_{ij} = \underset{h\in S}{\text{Avg}}\, \mu_h$.

Further, let the indexes 1 and 2 be assigned arbitrarily to each ED in a pair, h, of adjacent EDs (either type A or type B). Define, for each pair, h, of adjacent EDs

$$d_h^2 = \frac{1}{2}(\bar{y}_{h1} - \bar{y}_{h2})^2 \qquad (4.1.1)$$

$$\gamma_h^2 = \frac{1}{2}(\bar{\mu}_{h1} - \bar{\mu}_{h2})^2 \qquad (4.1.2)$$

and

$$v_h = \frac{1}{2}\left[\frac{\sigma_\varepsilon^2(h,1)(c_{h1}\rho_{h1}+1)}{m_{h1}} + \frac{\sigma_\varepsilon^2(h,2)(c_{h2}\rho_{h2}+1)}{m_{h2}}\right] \qquad (4.1.3)$$

where $\bar{y}_{ht}$, $\bar{\mu}_{ht}$, $\sigma_\varepsilon^2(h,t)$, c_{ht}, ρ_{ht} and m_{ht} (t = 1,2) are as previously defined but using the abbreviated notation.

Then, for $h\in A$, that is, for pairs of adjacent EDs lying within the same district

$$E(d_h^2) = \sigma_\delta^2 + \gamma_h^2 + v_h. \qquad (4.1.4)$$

And, for $h\in B$, that is, for pairs of adjacent EDs lying in different districts

$$E(d_h^2) = \sigma_\delta^2 + \sigma_\alpha^2 + \gamma_h^2 + v_h. \qquad (4.1.5)$$

We now estimate the components of variance in (3.2.2). Let A_h be the set of all pairs in A having an ED in common with the h-th ED pair in B. Define

$$\hat{\sigma}_e^2 = \underset{h\in A}{\text{Avg}}\, d_h^2 \qquad (4.1.6)$$

and

$$\hat{\sigma}_\delta^2 = \underset{h\in B}{\text{Avg}}\,[d_h^2 - \underset{p\in A_h}{\text{Avg}}\, d_p^2]. \qquad (4.1.7)$$

It can be shown that, under the stated assumptions,

$$E\hat{\sigma}_e^2 = \sigma_\alpha^2 + \underset{h\in A}{\text{Avg}}\, \gamma_h^2 + \underset{h\in A}{\text{Avg}}\, v_h \qquad (4.1.8)$$

and

$$E\hat{\sigma}_\delta^2 = \sigma_\delta^2 + \Delta_\gamma + \Delta v \qquad (4.1.9)$$

where

$$\Delta_\gamma = \underset{h\in B}{\text{Avg}}\,(\gamma_h^2 - \underset{p\in A_h}{\text{Avg}}\,\gamma_p^2)$$

and

$$\Delta_v = \underset{h\in B}{\text{Avg}}\,(v_h - \underset{p\in A_h}{\text{Avg}}\, v_p).$$

The following upper and lower bound estimators of V(Y) are proposed. Define

$$\hat{V}_L = \hat{\sigma}_\delta^2 \sum_i m_i^2 \qquad (4.1.10)$$

and

$$\hat{V}_U = \hat{V}_L + \hat{\sigma}_e^2 \sum\sum m_{ij}^2. \qquad (4.1.11)$$

4.2 Conditions under which V_L and V_U are Bound Estimators for V(Y)

The estimator $\hat{\sigma}_\delta^2$ is an average of between district office deviations squared minus the average of within district office deviations squared. The contrasts between EDs in different district offices estimate the combined effects of district office operations including the enumerator or ED effect and an effect of population characteristic differences for adjacent EDs. The contrasts between adjacent EDs within district office attempt to estimate the effects of different enumerators and different EDs so that the district office effect can be isolated by subtraction. Choosing the EDs in the estimates of the enumerator and ED effects, which surround the two EDs used for the between district office contrasts, attempts to reduce the bias in the estimate of the between district office variance by taking into account the possibility that the expected differences between adjacent EDs may depend upon the proximity of the EDs to district office borders and other other political boundaries.

Thus, $\hat{\sigma}_\delta^2$ is an estimator of σ_δ^2 with negligible bias if it can be assumed that the district office boundaries are drawn without regard to the value of the characteristic of interest or of variables highly correlated with the characteristic of interest. For example, if it assumed that the set $\mathcal{B}$ is a random sample from the set

$$S_\delta = \bigcup_{h\in\mathcal{B}} A_h \qquad (4.2.1)$$

that is if the district office boundary is essentially drawn in a random fashion through the set of EDs S, the estimator $\hat{\sigma}_\delta^2$ will be unbiased for α_δ^2; that is

$$E(\hat{\sigma}_\delta^2) = \sigma_\delta^2$$

where E includes the expectation over all possible samples of EDs of the same size as the set $\mathcal{B}$ from the set S_δ.

Therefore in expectation, the lower bound estimator, V_L, will underestimate V(Y) by

$$B_L = \sigma_\alpha^2 \sum\sum m_{ij}^2 + \sum\sum m_{ij}\sigma_\varepsilon^2(i,j)(1 + c_{ij}\rho_{ij}) \qquad (4.2.2)$$

Now consider $\hat{V}_U$, the upper bound estimator. As an estimator of V(Y), V_U has a bias

$$B_U = \operatorname*{Avg}_{h\in A} \gamma_h^2 \sum\sum m_{ij}^2 + \sum\sum m_{ij}^2 \left(\operatorname*{Avg}_{h\in A} v_h - \frac{\sigma_\varepsilon^2(i,j)}{m_{ij}}\right) \qquad (4.2.3)$$

The term $\operatorname*{Avg}_{h\in A} v_h$ will, in general, provide a close approximation to $\operatorname*{Avg}_{h\in S} \frac{\sigma_\varepsilon^2(h)}{m_h}$ where here S is the set of all EDs. Therefore, if it is assumed that $\sigma_\varepsilon^2(h) = \sigma_\varepsilon^2$, a constant, for all $h\in S$, it can be shown that the second term on the right of (4.2.3) is always positive. Furthermore, this term will be negligible if, in addition, the EDs have approximately the same size.

The first term on the right, however, is expected to be large. By restricting the estimation to contrasts of adjacent EDs, we attempt to minimize this term, which is the average difference between ED true population means.

Therefore, under the conditions stated above,

$$E(V_L) \le V(Y) \le E(V_U) \qquad (4.2.4)$$

and, the expected range of the bounds is given approximately by

$$R = (\sigma_\delta^2 + \operatorname*{Avg}_{h\in A} \gamma_h^2) \sum\sum m_{ij}^2 + \sum\sum m_{ij}\sigma_\varepsilon^2(i,j)(1 + c_{ij}\rho_{ij}) \qquad (4.2.5)$$

4.3 Estimation Based on a Sample

The previous results assume that all type A and type B EDs were used in the estimation. Because of the cost involved in matching and pairing EDs for all census districts, this approach may not be feasible. However, it is easily shown that the preceeding results also apply when either a random sample of district office pairs are selected from all possible pairs of adjacent district offices or a random sample of type A and type B ED pairs is selected from the sets A and $\mathcal{B}$. The extension of the general theory to allow for this sampling is straightforward and does not afford any difficulties.

1/ In [1], the model has been extended to the full organizational structure of the 1980 Census in which the data processing was performed outside the district offices in processing centers and several types of district office structures were used.

5. REFERENCES

[1] Biemer, P.P., "Methodology for Estimating Upper and Lower Bounds on Census Variance - Draft" memorandum for Distribution List, U.S. Bureau of the Census, May 29, 1981.

[2] Cochran, W.G., Sampling Techniques, 3rd ed., New York: John Wiley and Sons (1977).

6. ACKNOWLEDGEMENTS

The author wishes to thank John Thompson and Robert Fay for their helpful comments and suggestions in this research.