A TOTAL ERROR MODEL FOR THE 1990 CENSUS

Henry F. Woltman and Robert Johnson
Bureau of the Census

I. INTRODUCTION

This paper presents a 1990 Census total error model. The term "total error" implies an integrated treatment of different sources of error in census or survey statistics. The error sources in 1990 Census data include, but are not limited to, errors of noncoverage (frame), sampling, geocoding, editing, coding, interviewing, mail response, nonmail response, and imputation.

Two important applications of total error models have long been recognized: 1) estimation of the total error of a statistic; and 2) estimation of the relative impacts of different kinds of errors on the total error. Using total error models, statistical inferences can be based on the overall accuracy of statistics, rather than just the sampling precision. Moreover, estimation of the relative impacts of different kinds of errors might lay the basis for efficient allocation of resources to different operations, so as to minimize total error for fixed cost.

The model of this paper falls short of treating all sources of error in a single model. Hence, the term "total error model" is not used in a precise sense. The model seeks to account only for noncoverage, sampling, editing, response, enumerator, and imputation errors.

II. OPERATIONS

The total error model seeks to account for the processes giving rise to error in a single census statistic. Let P denote the population proportion to be estimated, and i denote the census questionnaire item that provides the data for computing the census statistic. The model applies to characteristics of households (housing items) as well as to characteristics of individuals (population items).

The model assumes that a proportion C of N total units (i.e., either households or persons, depending on the item i) are covered by the census and that a proportion (1 - C) are not covered. The number of units in the covered and noncovered subpopulations are $N_C = C \times N$ and $N_{NC} = (1 - C) \times N$, respectively.

The model assumes that P is estimated using the available data, i.e., information for the N_C units in the covered population. Figure 1 presents a flow diagram for selected operations affecting the total error of 1990 Census statistics. The model assumes that a proportion R of the N_C covered units, i.e., units represented on mail-out questionnaires, are mail returns and that a proportion (1-R) are nonmail returns. The numbers of units represented on mail returns and on nonmail returns are $N_1 = R \times N_C$ and $N_2 = (1-R) \times N_C$ respectively.

A proportion β^1 of the N_1 mail return units do not have complete data for item i, while a proportion $(1-\beta^1)$ have complete data for item i. The number of mail return units with complete data on item i is $N_{13} = (1-\beta^1) \times R \times N_C$.

Among the mail return units without complete data for item i, a proportion π^1 should fail edit and be sent to follow-up, and a proportion $(1-\pi^1)$ should pass edit, according to questionnaire decision rules and tolerances. The numbers of units represented by "true" fail-edit mail returns and "true" pass-edit mail returns, among those without complete data for item i, are $N_{11} = \beta^1 \times \pi^1 \times N_1$ and $N_{12} = \beta^1 \times (1-\pi^1) \times N_1$ respectively.

Like C, the proportions R, β^1, and π^1 are treated as fixed constants in the model. R is the mail response rate, β^1 is the mail return item nonresponse rate for item i, and π^1 is the conditional fail-edit rate, given missing or incomplete data on item i.

By implication, the population of N_C units represented on mail-out questionnaires is divided into four strata: 1) Stratum 11, the stratum of mail return fail edits with missing data for item i (size = N_{11} units); 2) Stratum 12, the stratum of mail return pass edits with missing data for item i (size = N_{12}); 3) Stratum 13, the stratum of mail return edits with complete data on item i (size = N_{13}); and 4) Stratum 2, the stratum of nonmail returns (size = N_2).

The N total units in the population are partitioned into five strata, the four strata of the covered population together with the noncovered stratum:

$$N = N_C + N_{NC}$$
$$= N_{11} + N_{12} + N_{13} + N_2 + N_{NC}$$

We will develop separate approximations for the error components in each stratum, and then combine the expressions, using weights determined by C, R, β^1, and π^1, in order to derive the mean square error of an estimated proportion.

Many of the remaining symbols in Figure 1 are subscripted by both k and j. These symbols denote random variables that are defined for each combination of an "operator" k (i.e., an enumerator), and a "unit" j (i.e., a household or person) of operator k's assignment. Exceptions occur in the following places: in Stratum 12, where no enumerator is defined since all responses are imputed; at the editing stage of Stratum 11, where units have not yet been assigned to enumerators; and in Stratum 13, where the subscript h denotes the household and j denotes the

eligible unit in the household to which item i applies (the household itself, if i is a housing item, or a household member, if i is a population item). Each subscripted variable in Figure 1 is also conditional upon the item i, but the i is omitted for notational simplicity. For each k and j, δ_j, λ_{kj}, and α_{kj} denote random indicator variables, i.e., random variables that take on either the value 0 or the value 1. Each indicates whether a census operation was carried out successfully (value = 0) or unsuccessfully (value = 1):

δ_j = 0 if editor correctly sends unit j of the failed-edit stratum (Stratum 11) to follow-up;

= 1 if editor does not send unit j of the failed-edit stratum to follow-up.

λ_{kj} = 0 if enumerator k obtains a response from unit j in the failed-edit follow-up.

= 1 if enumerator k does not obtain a response from unit j in the failed-edit follow-up.

α_{kj} = 0 if enumerator k obtains a response from unit j in the non-response (nonmail return) follow-up;

= 1 if enumerator k does not obtain a response from unit j in the nonresponse follow-up.

In this model, we assume that follow-up enumerator assignments have been interpenetrated, i.e., that the units have been split into random subsamples of equal size and assigned to operators. This simplifying assumption allows us to estimate certain covariances in the model that would otherwise be inestimable.

Figure 1 shows that, depending upon the stratum in which a covered unit falls and upon the successful or unsuccessful outcomes of census operations for that unit, the recorded value of item i is subject to one of three kinds of error: 1) response error, denoted by ϵ^R_{hj}; 2) enumerator error, denoted by ϵ^E_{kj}; or 3) imputation error, denoted by ϵ^I_{kj} or by ϵ^I_j. For example, if an editor correctly sends unit j in Stratum 11 to follow-up (i.e., $\delta_j = 0$), and if enumerator k does not obtain a response to item i for unit j (i.e., $\lambda_{kj} = 1$), then an imputation error, denoted ϵ^I_{kj}, results. This combination of stratum and outcomes is called "Path 4" in Fig. 1.

In all, seven operational "paths" are shown in Figure 1. Paths 2, 4, 5, and 7 result in imputation error; paths 3 and 6 result in enumerator error; and path 1 results in response error. Complete specification of the model requires: 1) assumptions about noncoverage; 2) assumptions about the sampling design; 3) assumptions about the assignment of operators to covered units; and 4) assumptions about the joint probability distributions of the indicator variables δ_{kj}, λ_{kj}, and α_{kj} and of the nonsampling errors ϵ^R_{hj}, ϵ^I_{kj}, and ϵ^E_{kj}.

III. MODEL ASSUMPTIONS

This model treats the census as a random sample from the population covered by the census. If item i is a 100% item (i.e., a question on the census short form distributed to every U.S. household), then it is a 100% sample, whereas if item i is a less-than-100% item (i.e., a question found only on the long form distributed to only a sample of U.S. households), then the sample is less than 100%.

Let y denote the observed value of a binary random variable, e.g., whether a household has complete kitchen facilities (y = 1) or does not have complete kitchen facilities (y = 0). Let x denote the unobserved "true" value of the same unit.

For each of the four covered strata, y can be written as the sum of x and a nonsampling error. The nonsampling error added to x can be a response, enumerator, or imputation error, depending on the stratum and on the realized values of the random indicator variables defined in Section II; hence the nonsampling errors can be expressed as weighted sums of nonsampling errors:

Stratum 11, units that fail edit:

$$y^{(11F)}_{kj} = x^{(11F)}_{kj} + \lambda_{kj}\epsilon^I_{kj} + (1-\lambda_{kj})\epsilon^E_{kj} = x^{(11F)}_{kj} + e^{(11F)}_{kj}$$

Stratum 11, units that pass edit:

$$y^{(11P)}_j = x^{(11P)}_j + \epsilon^I_j$$

Stratum 12: $y^{(12)}_j = x^{(12)}_j + \epsilon^I_j$

Stratum 13: $y^{(13)}_{hj} = x^{(13)}_{hj} + \epsilon^R_{hj}$

Stratum 2: $y^{(2)}_{kj} = x^{(2)}_{kj} + (1-\alpha_{kj})\epsilon^E_{kj} + \alpha_{kj}\epsilon^I_{kj} = x^{(2)}_{kj} + e^{(2)}_{kj}$

In Strata 11 and 2, $y^{(11)}_{kj}$ and $y^{(2)}_{kj}$ denote the jth unit in the kth enumerator's assignment. In Stratum 12, and among the pass-edits of Stratum 11, responses are always imputed so there is no enumerator (no subscript k). In Stratum 13, the mail response stratum, a single household respondent is assumed to respond for all units in the hth household. That is, $y^{(13)}_{hj}$ denotes the observed response for the jth eligible unit in the hth household.

Note that, if x = 0, each of the nonsampling errors, ϵ^I, ϵ^E, and ϵ^R, can equal either 0, producing a correct response, or 1, producing a false positive response. If x = 1, each of

the nonsampling errors can equal either 0, producing a correct response, or -1, producing a false negative response. Because $e^{(11F)}$ and $e^{(2)}$ are a random selection of enumerator and imputation errors, they also take the possible values 0 or 1 if x = 0, and 0 or -1 if x = 1.

In order to derive the total mean square error of a sample proportion (Section IV), we made the following assumptions about noncoverage, about the sample design, about the assignments of units to operators, and about the joint distributions of random variables.

A. Noncoverage: We assume that a fixed proportion P_{NC} of the $N_{NC} = (1 - C) \times N$ units in the noncovered stratum have the characteristic (i.e., x = 1).

B. Sample Design: We regard the census as a simple random sample of n units from the total of N_C units in the covered population. The sampling fraction is $f = n/N_C$. There are n_{11} sample units in Stratum 11, n_{12} sample units in Stratum 12, n_{13} sample units in Stratum 13, and n_2 sample units in Stratum 2. It follows that $n = n_{11} + n_{12} + n_{13} + n_2$.

Under simple random sampling, n_{11}, n_{12}, n_{13}, and n_2 are random variables; but in the calculations that follow, they are regarded as constants fixed before the sample was drawn. By this simplification, we treat our poststratified sample as if it were stratified, i.e., as if the correct stratum to which each unit belonged were known before the sample was drawn.

C. Operator Assignments: We assume that the n_{11} sample units in Stratum 11 were sent to editors, and of these n_P passed edit (were not sent to follow-up) and n_F failed edit (were sent to follow-up). We assume that the n_F units sent to follow-up were randomly split into K_{11} equal subsamples (assume that n_F divides by K evenly). One subsample was assigned to each of K_{11} enumerators. Let m_{11} denote the size of an enumerator's assignment; it follows that $n_F = K_{11}m_{11}$.

In Stratum 13, we assume that a single respondent provided answers for all eligible units in the household. Let h_{13} denote the number of households in the sample. For simplicity, we also assume a constant household size of m_{13} eligible units. It follows that $n_{13} = h_{13}m_{13}$. If item i is a population characteristic, characteristic, m_{13} equals the average number of eligible persons in a household. If item i is a housing characteristic, $m_{13} = 1$.

For Stratum 2, assume the n_2 sample units were randomly split into K_2 equal subsamples and one subsample was assigned to each of K_2 nonresponse follow-up enumerators. Let m_2 denote the size of a nonresponse enumerator assignment; it follows that $n_2 = K_2m_2$.

D. Joint Distribution of Random Variables: We propose a hierarchical model for the indicators of census events, δ_j, λ_{kj}, and α_{kj}, as well as for the nonsampling errors ϵ^R, ϵ^E, and ϵ^I. This model treats the individual operators' error rates as random variables, drawn from a probability distribution of possible error rates.

For the indicators of census events, assume the following (wp = "with probability"):

$$\delta_j = \begin{cases} 1 \text{ wp } \delta \\ 0 \text{ wp } 1-\delta \end{cases} \quad \text{independently, } j=1,\ldots,n_{11},$$

$$\lambda_{kj}|\lambda_k = \begin{cases} 1 \text{ wp } \lambda_k \\ 0 \text{ wp } 1-\lambda_k \end{cases} \quad \text{independently, } j=1,\ldots,m_{11}, \; k=1,\ldots,K_{11},$$

$$\alpha_{kj}|\alpha_k = \begin{cases} 1 \text{ wp } \alpha_k \\ 0 \text{ wp } 1-\alpha_k \end{cases} \quad \text{independently, } j = 1,\ldots, m_2, \; k = 1, \ldots, K_2,$$

where δ is the average "fail-edit error rate" among all editors, λ_k is the "fail-edit follow-up failure rate" for the kth enumerator, and α_k is the "nonresponse follow-up failure rate" for the kth enumerator.

The follow-up failure rates are assumed to be independently and identically distributed random variables, drawn from infinite populations, as follows:

$\lambda_1, \ldots, \lambda_{K_{11}}$ are iid with mean λ and variance σ_δ^2.

$\alpha_1, \ldots, \alpha_{K_2}$ are iid with mean α and variance σ_α^2.

The sets of random variables above are assumed to be independent of each other, and independent of the values of x.

For the nonsampling errors, we assume another hierarchical model, with differing rates of false positives and false negatives among operators. For response errors, we have:

$$[\epsilon^R_{hj}|x_{hj}=1,\theta^R_h] = \begin{cases} -1 \text{ wp } \theta^R_h \\ 0 \text{ wp } 1-\theta^R_h \end{cases}$$

$$[\epsilon^R_{hj}|x_{hj}=0,\phi^R_h] = \begin{cases} 1 \text{ wp } \phi^R_h \\ 0 \text{ wp } 1-\phi^R_h \end{cases}$$

independently, $j = 1, \ldots, m_{13}$, $h = 1, \ldots, h_{13}$.

Following the model for qualitative data found in Bailar and Biemer (1984) and U.S. Bureau of the Census (1985), we assume that the false negative rate θ^R_h and the false positive rate ϕ^R_h for each household are random variables, drawn from a bivariate distribution: (θ^R_h, ϕ^R_h), $h = 1, \ldots, h_{13}$, are iid with means (θ_R, ϕ_R), variances $(\sigma_{\theta R}^2, \sigma_{\phi R}^2)$, and covariance $\sigma_{\theta\phi R}$.

The same structure applies to the enumerator errors:

$$[\epsilon^E_{kj}|x_{kj}=1,\theta^E_k] = \begin{cases} -1 \text{ wp } \theta^E_k \\ 0 \text{ wp } 1-\theta^E_k \end{cases}$$

$$[\epsilon^{E}_{kj} | x_{kj}=0, \phi^{E}_{k}] = \begin{cases} 1 & \text{wp } \phi^{E}_{k} \\ 0 & \text{wp } 1-\phi^{E}_{k} \end{cases}$$

independently, j = 1, ..., m_{11} (or m_2), k = 1, ..., K_{11} (or K_2).

The false negative rate θ^{E}_{k} and the false positive rate ϕ^{E}_{k} are drawn from a bivariate distribution: $(\theta^{E}_{k}, \phi^{E}_{k})$, k = 1, ..., K_{11} (or K_2) are iid with means (θ_E, ϕ_E), variances $(\sigma_{\theta E}^2, \sigma_{\phi E}^2)$, and covariance $\sigma_{\theta\phi E}$.

The model for imputation errors however, has a simpler structure. Because there is only one operator, the imputation algorithm itself, there is only a single false positive rate and false negative rate; there is no between-operator variation in these rates. Hence the following model is assumed:

$$[\epsilon^{I}_{kj} | x_{kj}=1] = \begin{cases} -1 & \text{wp } \theta_I \\ 0 & \text{wp } 1-\theta_I \end{cases}$$

$$[\epsilon^{I}_{kj} | x_{kj}=0] = \begin{cases} 1 & \text{wp } \phi_I \\ 0 & \text{wp } 1-\phi_I \end{cases}$$

independently for all k and j.

Under these model assumptions, the conditional expectations and variances of the three types of nonsampling errors, given the associated true values, are as follows:

$$E[\epsilon^{R}_{hj} | x_{hj}] = -x_{hj}\theta_R + (1-x_{hj})\phi_R$$

$$E[\epsilon^{E}_{kj} | x_{kj}] = -x_{kj}\theta_E + (1-x_{kj})\phi_E$$

$$E[\epsilon^{I}_{kj} | x_{kj}] = -x_{kj}\theta_I + (1-x_{kj})\phi_I$$

$$V[\epsilon^{R}_{hj} | x_{hj}] = x_{hj}\theta_R(1-\theta_R) + (1-x_{hj})\phi_R(1-\phi_R)$$

$$V[\epsilon^{E}_{kj} | x_{kj}] = x_{kj}\theta_E(1-\theta_E) + (1-x_{kj})\phi_E(1-\phi_E)$$

$$V[\epsilon^{I}_{kj} | x_{kj}] = x_{kj}\theta_I(1-\theta_I) + (1-x_{kj})\phi_I(1-\phi_I)$$

It can also be shown that, within a stratum s, errors from the same operator are correlated, whereas errors arising from different operators are independent. Specifically, for response errors,

$$\mathrm{Cov}[\epsilon^{R}_{hj}, \epsilon^{R}_{h'j'} | x_{hj}, x_{h'j'}] = x_{hj}x_{h'j'}\sigma_{\theta R}^2 + x_{hj}(1-x_{h'j'})\sigma_{\theta\phi R} + x_{h'j'}(1-x_{hj})\sigma_{\theta\phi R} + (1-x_{hj})(1-x_{h'j'})\sigma_{\phi R}^2$$

for h = h′ and j ≠ j′, and 0 for h ≠ h′. For enumerator errors,

$$\mathrm{Cov}[\epsilon^{E}_{kj}, \epsilon^{E}_{k'j'} | x_{kj}, x_{k'j'}] = x_{kj}x_{k'j'}\sigma_{\theta E}^2 + x_{kj}(1-x_{k'j'})\sigma_{\theta\phi E} + x_{k'j'}(1-x_{kj})\sigma_{\theta\phi E} + (1-x_{kj})(1-x_{k'j'})\sigma_{\phi E}^2$$

for k = k′ and j ≠ j′, and 0 for k ≠ k′.

In other words, for mail responses, errors are correlated within but not between households. For enumerator returns, errors are correlated within but not between enumerator assignments. For imputations, however, all errors are uncorrelated:

$$\mathrm{Cov}[\epsilon^{I}_{kj}, \epsilon^{I}_{k'j'} | x_{kj}, x_{k'j'}] = 0 \quad \text{for } k \neq k' \text{ and/or } j \neq j'.$$

From the expressions above, it is possible to derive the conditional means and variances of $e^{(2)}_{kj}$ and $e^{(11F)}_{kj}$, which are random mixtures of enumerator and imputation errors. (The derivation is excluded due to space limitations.)

We assume that the true values, x_{kj}'s, are uncorrelated with the $\delta_{j'}$'s, $\alpha_{k'j'}$'s, and $\lambda_{k'j'}$'s for all k, k′, j, and j′ and uncorrelated with the nonsampling errors $\epsilon^{R}_{k'j'}$, $\epsilon^{I}_{k'j'}$, and $\epsilon^{E}_{k'j}$, for all k ≠ k′ and/or j ≠ j′. Because the data are qualitative, however, there is a negative correlation between a sample unit's true value and its nonsampling error; this arises because a unit whose true value is 1 can only have a nonsampling error of 0 or -1, while a unit whose true value is 0 can only have an error of 0 or 1 (see U.S. Bureau of the Census, 1985).

IV. DECOMPOSITION OF MEAN SQUARE ERROR

Let P = the proportion of the population having a specified characteristic and let p = the census estimate of P. Using the symbols of Section II (suppressing the superscript i for convenience), we can write

$$P = CR\beta\pi P_{11} + CR\beta(1-\pi)P_{12} + CR(1-\beta)P_{13} + C(1-R)P_2 + (1-C)P_{NC},$$

where P_{11}, P_{12}, P_{13}, P_2, and P_{NC} are the proportions of units in the population having the characteristic (x = 1) in Strata 11, 12, 13, 2, and the noncovered stratum, respectively.

Similarly, assuming known values of R, β, and π, the census estimate p of P can be written

$$p = R\beta\pi p_{11} + R\beta(1-\pi)p_{12} + R(1-\beta)p_{13} + (1-R)p_2, \qquad (2)$$

where the p's are the sample estimates in Strata 11, 12, 13, and 2 respectively.

We now provide expressions for the mean square error of p,

$$\mathrm{MSE}(p) = E[(p-P)^2],$$

where the expectation is taken over an infinite number of repetitions of response errors and census operations (i.e., over the probability distributions of Section III-D), over all possible enumerator assignments (under the interpenetration assumptions of Section III-C), and over repeated

samples (under the sampling assumptions of Section III-B).

Assuming all nonsampling covariances among strata are zero, the mean square error of p can be written

$$MSE(p) = BIAS^2(p) + VAR(p),$$

where

$$BIAS(p) = CR\beta\pi BIAS(p_{11}) + CR\beta(1-\pi)BIAS(p_{12}) + CR(1-\beta)BIAS(p_{13}) + C(1-R)BIAS(p_2) + (1 - C)(E(p)-P_{NC})$$

and

$$VAR(p) = R^2\beta^2\pi^2VAR(p_{11}) + R^2\beta^2(1-\pi)^2VAR(p_{12}) + R^2(1-\beta)^2VAR(p_{13}) + (1-R)^2VAR(p_2) .$$

The biases of p_{11}, p_{12}, p_{13}, and p_2 in are biases relative to P_{11}, P_{12}, P_{13}, and P_2 respectively. Note from the last term that noncoverage contributes a bias equal to the product of the proportion of noncoverage and the difference between the expectation of the census estimate p and the true proportion having the characteristic in the noncovered stratum (cf. Cochran, 1963, p. 357). The individual stratum biases can be derived from the conditional expectations of the errors given their associated true values (in Section III-D), by averaging over the interpenetration and sampling designs. For Strata 12, 13, and 2, the resulting expressions are:

$$BIAS(p_{12}) = -\theta_I P_{12} + \phi_I(1-P_{12})$$
$$BIAS(p_{13}) = -\theta_R P_{13} + \phi_R(1-P_{13})$$
$$BIAS(p_2) = -[\alpha\theta_I + (1-\alpha)\theta_E]P_2 + [\alpha\phi_I + (1-\alpha)\phi_E](1-P_2).$$

The bias for Stratum 11 is more complicated, because the number of cases that pass edit n_P and the number that fail edit n_F are random variables. To simplify matters, we derive the bias assuming that the sample proportion of pass edits, n_P/n_{11}, is equal to the probability of passing edit δ. Under this assumption, the bias in Stratum 11 is given by

$$BIAS(p_{11}) = -[\delta\theta_I + (1-\delta)(\lambda\theta_I + (1-\lambda)\theta_E)]P_{11} + [\delta\phi_I + (1-\delta)(\lambda\phi_I + (1-\lambda)\phi_E)](1-P_{11}).$$

Following U.S. Bureau of the Census (1985), we decompose each of the variances in (4) into three parts, which we denote by SV ("sampling variance"), SRV ("simple response variance"), and CC ("correlated component"). For Stratum 13, the decomposition is

$$VAR(p_{13}) = \frac{1}{n_{13}}(1-f)\,SV(p_{13}) + \frac{1}{n_{13}}SRV(p_{13}) + \frac{(m_{13}-1)}{n_{13}}CC(p_{13})$$

where $f_{13} = n_{13}/N_{13}$ is the sampling fraction in Stratum 13. The sampling variance $SV(p_{13})$ is defined as

$$SV(p_{13}) = V\,E[\epsilon^R_{hj}|x_{hj}]$$

where the expectation is taken over response errors, census operations, and the interpenetration design, and the variance is taken over the sample design. Note that when the sampling fraction equals unity (as for a 100% questionnaire item), the SV term makes no contribution to the variance.

The simple response variance and correlated component are defined as

$$SRV(p_{13}) = E\,V[\epsilon^R_{hj}|x_{hj}]$$
$$CC(p_{13}) = E\,Cov[\epsilon^R_{hj},\epsilon^R_{h'j'}|x_{hj},x_{h'j'}]$$

where the variance and covariance are taken over response errors and census operations, and the expectations are taken over the interpenetration design and sample design.

The variance in Stratum 2 is decomposed in an analagous fashion,

$$VAR(p_2) = \frac{1}{n_2}(1-f_2)\,SV(p_2) + \frac{1}{n_2}SRV(p_2) + \frac{(m_2-1)}{n_2}CC(p_2$$

$$SV(p_2) = V\,E[e^{(2)}_{kj}|x_{kj}]$$
$$SRV(p_2) = E\,V[e^{(2)}_{kj}|x_{kj}]$$
$$CC(p_2) = E\,Cov[e^{(2)}_{kj},e^{(2)}_{k'j'}|x_{kj},x_{k'j'}]$$

where the expectations, variances, and covariances in the expressions above have the same meaning as in Stratum 13.

In Stratum 12, there is no operator assignment, and all errors are conditionally uncorrelated given their true values; hence, the variance is decomposed as follows:

$$VAR(p_{12}) = \frac{1}{n_{12}}(1-f_{12})\,SV(p_{12}) + \frac{1}{n_{12}}SRV(p_{12})$$

$$SV(p_{12}) = V\,E[\epsilon^I_{kj}|x_{kj}]$$
$$SRV(p_{12}) = E\,V[\epsilon^I_{kj}|x_{kj}]$$

where, in the expression for SV, the expectation is over response errors and the variance is over the sample design; and, in the expression for SRV, the variance is over response errors and the expectation is over the sample design.

Stratum 11 is more complicated because the split of n_{11} into n_P and n_F is random. Again, if we assume that the proportion n_P/n_{11} is equal to δ, we can derive simpler expressions for the variance. We decompose the variance as follows:

$$VAR(p_{11}) = \frac{1}{n_{11}}(1-f_{11})\,SV(p_{11}) + \frac{1}{n_{11}}SRV(p_{11}) + \frac{(m_{11}-1)}{n_F}CC(p_{11})$$

The variance components are

$$SV(p_{11}) = \delta\ V\ E[\epsilon^{I}_{j} | x_j] + (1-\delta)\ V\ E[e^{(11F)}_{kj} | x_{kj}]$$

$$SRV(p_{11}) = \delta\ E\ V[\epsilon^{I}_{j} | x_j] + (1-\delta)\ E\ V[e^{(11F)}_{kj} | x_{kj}]$$

$$CC(p_{11}) = (1-\delta)^2\ E\ Cov[e^{(11F)}_{kj}, e^{(11F)}_{k'j'} | x_{kj}, x_{k'j'}]$$

where expectations, variances, and covariances have the same meaning as before.

We now give expressions for the individual SV's, SRV's and CC's. These can be derived through straightforward but tedious calculation. They are presented below, in order of increasing complexity.

$$SV(p_{12}) = [1 - (\theta_I+\phi_I)]^2\ P_{12}\ (1-P_{12})$$

$$SV(p_{13}) = [1 - (\theta_R+\phi_R)]^2\ P_{13}\ (1-P_{13})$$

$$SV(p_2) = [1 - (\alpha(\theta_I+\phi_I) + (1-\alpha)(\theta_E+\phi_E))]^2\ P_2\ (1-P_2)$$

$$SV(p_{11}) = [\ \delta(1-(\theta_I+\phi_I))^2 + (1-\delta)[1-(\lambda(\theta_I+\phi_I) + (1-\lambda)(\theta_E+\phi_E))]^2\ ]\ P_{11}\ (1-P_{11})$$

$$SRV(p_{12}) = \theta_I(1-\theta_I)\ P_{12} + \phi_I(1-\phi_I)\ (1-P_{12})$$

$$SRV(p_{13}) = \theta_R(1-\theta_R)\ P_{13} + \phi_R(1-\phi_R)\ (1-P_{13})$$

$$SRV(p_2) = [\ \alpha\theta_I(1-\theta_I) + (1-\alpha)\theta_E(1-\theta_E) + \alpha(1-\alpha)(\theta_I-\theta_E)^2\ ]\ P_2 + [\ \alpha\phi_I(1-\phi_I) + (1-\alpha)\phi_E(1-\phi_E) + \alpha(1-\alpha)(\phi_I-\phi_E)^2\ ]\ (1-P_2)$$

$$SRV(p_{11}) = \delta\ \{\ \theta_I(1-\theta_I)\ P_{11} + \phi_I(1-\phi_I)\ (1-P_{11})\ \} + (1-\delta)\ \{\ [\ \lambda\theta_I(1-\theta_I) + (1-\lambda)\theta_E(1-\theta_E) + \lambda(1-\lambda)(\theta_I-\theta_E)^2\ ]\ P_{11} + [\ \lambda\phi_I(1-\phi_I) + (1-\lambda)\phi_E(1-\phi_E) + \lambda(1-\lambda)(\phi_I-\phi_E)^2\ ]\ (1-P_{11})\ \}$$

$$CC(P_{13}) = \sigma_{\theta R}^2\ P_{13}^2 - 2\ \sigma_{\theta\phi R}^2\ P_{13}\ (1-P_{13}) + \sigma_{\phi R}^2\ (1-P_{13})^2$$

$$CC(P_2) = \{\ \sigma_{\theta E}^2[(1-\alpha)^2 + \sigma_\alpha^2] + \sigma_\alpha^2(\theta_I-\theta_E)^2\ \}\ P_2^2 - 2\ \{\ \sigma_{\theta\phi E}^2[(1-\alpha)^2 + \sigma_\alpha^2] + \sigma_\alpha^2(\theta_I-\theta_E)(\phi_I-\phi_E)\ \}\ P_2(1-P_2) + \{\ \sigma_{\phi E}^2[(1-\alpha)^2 + \sigma_\alpha^2] + \sigma_\alpha^2(\phi_I-\phi_E)^2\ \}\ (1-P_2)^2$$

$$CC(P_{11}) = [\ \{\ \sigma_{\theta E}^2[(1-\lambda)^2 + \sigma_\lambda^2] + \sigma_\lambda^2(\theta_I-\theta_E)^2\ \}\ P_{11}^2 - 2\ \{\ \sigma_{\theta\phi E}^2[(1-\lambda)^2 + \sigma_\lambda^2] + \sigma_\lambda^2(\theta_I-\theta_E)(\phi_I-\phi_E)\ \}\ P_{11}(1-P_{11}) + \{\ \sigma_{\phi E}^2[(1-\lambda)^2 + \sigma_\lambda^2] + \sigma_\lambda^2(\phi_I-\phi_E)^2\ \}\ (1-P_{11})^2\ ]\ (1-\delta)^2$$

VI. REFERENCES

Bailar, B.A. and Biemer, P.P., 1984. "Some Methods for Evaluating Nonsampling Error in Household Censuses and Surveys," pp. 253-273 in W.G. Cochran's Impact on Statistics, edited by S.R. Rao and J. Sedransk, New York: Wiley.

Biemer, P.P., 1980. "A Survey Error Which Includes Edit and Imputation Error," pp. 610-615 in Proceedings of the Survey Research Methods Section, American Statistical Association.

Cochran, W.G., 1963. Sampling Techniques, 2nd ed., New York: Wiley.

Hansen, M.H., Hurwitz, W.N., and Bershad, M., 1961. "Measurement Errors in Censuses and Surveys," Bulletin of the International Statistical Institute 38: 359-374.

Katzoff, E.B. and Biemer, P.P., 1980. "Estimation of Nonsampling Error Due to Selected Office Operations of the 1980 Census," pp. 616-621 in Proceedings of the Survey Reseach Methods Section, American Statistical Association.

Lessler, J.T., 1983. "An expanded Survey Error Model," pp. 259-270 in Incomplete Data in Sample Surveys. Vol 3 Proceedings in the Symposium, edited by W.G. Madow and I. Olkin, New York: Academic Press.

Platek, R. and Grey, G.B., 1983. "Imputation Methodology: Total Survey Error," pp. 249-333 in Incomplete Data in Sample Surveys, Vol 2. Theory and Bibliographies, edited by W.G. Madow, I. Olkin, and D.B. Rubin, New York: Academic Press.

U.S. Bureau of the Census, 1985. Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, Appendix by P.P. Biemer, Washington, DC: U.S. Government Printing Office.

FIGURE 1

FLOW DIAGRAM FOR CENSUS TOTAL ERROR MODEL