MULTIPLE-SYSTEM METHODS FOR CENSUS COVERAGE EVALUATION

Alan M. Zaslavsky, M.I.T. and Harvard University
Department of Statistics, Harvard University, Cambridge, MA 02139

KEYWORDS: Undercount, dual system estimator, random-effects model.

1. Introduction.

In many contexts, it is desired to obtain an estimate of the size of a population for which a complete and accurate census is impossible. Likewise, it may be desirable to assess the degree of coverage of an incomplete census. Among the situations that fit this description are:

(1) Censuses of animal populations, in which it is usually impossible to obtain more than a sample of the population. (Seber 1973)
(2) Censuses in underdeveloped countries, for which resources and infrastructure limitations may make it impossible to construct complete address lists or to carry out an enumeration.
(3) Estimation of a rare population, such as people with a rare disease, for which there are two or more incomplete lists. (Wittes 1970)

Even the U.S. Decennial Census, certainly one of the best-funded and most sophisticated censal operations in the world, does not have complete coverage of the population, according to estimates from recent decades. (Citro and Cohen 1983) Furthermore, undercoverage varies substantially between groups defined by demographic and geographic variables.

Improved estimates may often be obtained by combining information from two sources (for example, a census and a sample survey such as the Post Enumeration Survey). These "dual system estimates" may be challenged where the underlying assumption of independence between the two sources cannot be justified. The dual system estimator (DSE) is described in Section 2, together with criticisms of the DSE for "correlation bias," that is, bias due to an unjustified assumption of independence between sources.

One method of reducing correlation bias is to obtain additional sources of information on the population, thereby reducing the fraction of the population that is omitted from all sources. In Section 3, estimates that use more than two sources ("multiple system estimators" or MSE) are described. While these can improve upon the DSE, it may be too expensive to collect additional sources on the same scale as the second source. Collecting the third source in a subsample of the two-source area is a compromise that limits costs while still providing a check on the accuracy of the DSE. In Section 4, a subsampled MSE is proposed, and in Section 5, methods are described for estimating the census coverage and total population with a subsampled MSE design.

The exposition of this paper will assume that the primary source is a Census, the secondary source is a coverage evaluation survey, and the additional sources are further surveys or administrative records lists. However, the methods described here are equally applicable in other situations where it is desired to estimate the coverage of a list. For example, the primary source might be a list of participants in a particular program, the other sources might be surveys and administrative records that are used to identify the pool of potential participants, and the objective might be to estimate the coverage rate of the program. Indeed, these methods would be even more useful in these non-Census settings, where coverage rates are relatively poor and therefore the number of units omitted from both the first and second sources may be substantial.

2. The Dual System Estimator and correlation bias.

One of the main statistical tools for population estimation and coverage evaluation where a complete census is not possible is the Dual System Estimator (DSE). In dual system estimation, two surveys or censuses are carried out on the same population. Each observed unit is identified as included in the first survey only, the second survey only, or both surveys. Thus each observation falls into one cell of a 2×2 contingency table like Table 1 where the cell count n_{00} (corresponding to units omitted from both surveys) is unobserved.

The usual assumptions of this methodology are that (1) the population is closed (no births, deaths, or migration) and of fixed size, (2) the two censuses can be matched to determine which units appeared in both, (3) inclusion in the first census is statistically independent of inclusion in the second, and (4) the numbers of units in the cells of Table 1 are multinomially distributed. Under these assumptions, the number of units in the unobserved cell may be estimated as $n_{00} = n_{10}n_{01}/n_{11}$. Using this estimate, the total population of the sampled area may be estimated; the population estimate obtained by summing the cells is

$$n_{++} = n_{1+}n_{+1}/n_{11}. \qquad (1)$$

DSE methodology has been used in Decennial Census coverage evaluation since 1950; the most ambitious effort to date was in the 1980 Census. In every case the first survey source was the Census itself. The second source

First source	Second source		
	Observed	Omitted	Total
Observed	n_{11}	n_{10}	n_{1+}
Omitted	n_{01}	n_{00} Unobserved cell	n_{0+}
Total	n_{+1}	n_{+0}	n_{++} Total population

Table 1: The Dual System Estimator

is a sample survey. In 1980, the second survey was based on the Current Population Survey sample; in 1990 there will be a special survey in a sample of Census blocks, the Post Enumeration Survey (PES), for purposes of coverage evaluation. The second source is only collected for a sample of the population (the CPS sample in 1980 and a specially drawn sample of blocks in 1990). Each stratum in the sample represents an estimation class, that is a subset of the entire population defined by the same variables.

The concept in this effort is somewhat different than that underlying the Peterson estimator, because the second survey consists of only a sample of geographical areas. The population of the survey area is no longer of primary interest. Rather, the total population (for each estimation class) must be estimated. Populations for local areas may also be of interest.

Focusing attention for the moment on a single estimation class, the quantity calculated from the DSE is the coverage rate $r = n_{1+}/n_{++}$ of the first source (the Census). Under the assumption of independence, this is consistently estimated by the observable quantity n_{11}/n_{+1}. (More precisely, the corresponding ratios of expectations are equal.) If the sample for which the second survey is carried out is a properly drawn sample (for each stratum) of the corresponding estimation class, then the sample estimate of the coverage rate is an estimate of the average coverage rate for that class across the population. An estimate of the national population in that class is then N_{1+}/r, where N_{1+} is the enumerated total in the Census for that class. Under the assumption that the undercount rate in any class is constant across geographical areas, the estimated population belonging to a particular class in any local area is $n_{1+}^{(i)}/r$, where $n_{1+}^{(i)}$ is the enumerated total in the Census for that class in area i. The "synthetic estimate" of total local area population is calculated by summing the local-area estimates across all classes.

In practice the factor that would be used for estimation would not be simply $1/r$ but a factor that takes into account other coverage errors such as erroneous enumerations. For simplicity, this exposition will assume that only undercoverage is of concern. It is reasonable to assume that overcoverage is adequately dealt with by checking against a single additional source since there is no reason to think that erroneous enumerations will be duplicated in the second source.

The important distinction here is between

(1) a *population* DSE, such as that described for animal studies, in which the entire finite population is targeted in each survey, and

(2) a *sample-based* DSE in which a population coverage rate is estimated from a sample of areas.

The simple form of the estimator $r = n_{11}/n_{+1}$ may mask the fact that this is actually an average of coverage rates across geographic areas,

$$r = n_{11}/n_{+1} = \frac{\sum_{i \in \ Sample} n_{+1}^{(i)} r^{(i)}}{\sum_{i \in \ Sample} n_{+1}^{(i)}}, \text{where } r^{(i)} = n_{11}^{(i)}/n_{+1}^{(i)}. \quad (2)$$

Thus, when the design of the second source is more complicated than a simple random sample (as it must be in practice), more complex estimators should be used. (Cowan and Malec 1986)

One of the critical assumptions in dual system estimation is that of independence between the first and second sources. But some models of the dynamics of the census process, often labeled heterogeneity and causal-effect models, suggest that this assumption may not be true. (Hogan and Wolter 1988)

"Correlation bias" refers to any systematic bias in the DSE caused by an incorrect assumption of independence of sources (that is, failure to take account of correlation between omission in the first and second sources). The bias affects estimation of the unobserved quantity n_{00}. Heterogeneity will always cause the estimator of n_{00} to be biased downwards, whereas causal effects models can imply a bias in either direction. Another way of regarding correlation bias is to say that the coverage rate n_{11}/n_{+1} in units observed in the second source is not representative of the coverage rate n_{1+}/n_{++} for all units.

Ericksen and Kadane (1985) suggest parametrizing the degree of correlation bias by k, the ratio of the actual number of twice-unobserved units to that predicted under independence of sources, or equivalently the crossproduct ratio in the 2×2 table (Table 1). Thus $n_{00} = kn_{01}n_{10}/n_{11}$. They further suggest using demographic estimates of the total population (based on births, deaths, and migration for demographically defined groups) together with the known rates n_{11}/n_{+1} and n_{11}/n_{1+} to estimate k for each stratum. In this way they estimate an overall value k=2.1 for blacks in the 1980 Census and Post Enumeration Program.

3. Multiple system estimator methods.

Methods that make use of not just two but a multiplicity of sources are called multiple system estimators. Although some units may remain unobserved in all sources, the additional sources improve the combined coverage and make possible more sophisticated models.

The multiple-system equivalent of the Petersen estimator in animal population estimation is the Schnabel estimator, which also assumes that all sources are independent of each other. Variations of the multiple system estimator use the additional information contributed by successive captures to quantify violations of the assumptions of the Petersen estimator, such as migration, births and deaths of animals. (Seber 1973)

Other estimators (Feinberg 1972; Bishop, Feinberg and Holland 1974), drawing on the framework of log-linear models, relax the assumption of independence between sources. Some a priori assumption is required to make the models identifiable, since there is always one cell count that is unobserved (corresponding to units missed by all sources). Typically, some higher-order interactions among sources (minimally, the highest-order interaction term) in the log-linear model are assumed to be zero. As the combined coverage of the sources improves, assumptions of this sort will presumably introduce less bias since the count in the cell corresponding to the never-observed

part of the population will become smaller. Thus, in the three-source case, a minimal assumption is that there is no three-way interaction.

A simpler multiple-source methodology (called by Kadane and Ericksen (1985) the "mega-list" method in contrast to the Wittes-Feinberg "multi-list" method) combines all the sources into a single list that is assumed to approximate a complete list of the population. The coverage of a single source is estimated as the fraction of units on the "mega-list" that are included in the target source. This is equivalent to assuming that the cell corresponding to units omitted from all lists is empty. Comparisons of the fits of these models given heterogeneous populations are shown in Zaslavsky (1989).

4. A sub-sampled multiple system estimator.

The most direct way of applying multiple system estimator (MSE) methods to undercount estimation would be to extend the Post-Enumeration Program (PEP) to multiple sources, collected and matched in all the PES blocks. Then the ratio of the census enumerated population in an estimation class in the PES blocks to the corresponding estimated population using all sources is the maximum likelihood estimate of the coverage rate. The estimated population could be derived from either the "multi-list" or "mega-list" MSE.

Due to the difficulty of matching more than two sources in a human census, it would probably not be feasible to carry out the MSE on such a broad scale. On the other hand, if multiple sources were used in a *subsample* of the PES blocks, then the additional information derived from the MSE could be used to calibrate the DSE estimates calculated from the full PES sample, by giving a better (less biased) estimate size of the "unobserved" cell n_{00} in the DSE. Typically, n_{00} is small compared to the other cells in the DSE. Thus, even if the sampling variability in the MSE of n_{00} is *relatively* large (as measured by its coefficient of variation), because of the smallness of the MSE sample, it would still not contribute much to the variance of the estimated *total* population. The additional variance of the estimate might well be outweighed by the reduction in bias. Perhaps equally important, the information derived from the MSE would provide information to help settle one of the major controversies surrounding use of the DSE in coverage evaluation, that is the controversy over the extent of correlation bias.

The three survey designs are illustrated in Figure 1. The vertical axis represents the entire list of blocks. Panel (a) shows the standard DSE design, in which all blocks are included in the Census and a sample is included in the Post Enumeration Survey (PES). Panel (b) represents the standard MSE design, in which the PES and a third source ("# 3") are obtained for a single sample of blocks. Panel (c) represents the proposed MSE design, in which a third source is obtained for a subsample of the blocks included in the PES; only for this subsample of blocks must the three-way matching operation be carried out.

Once estimates of n_{00} in the entire DSE sample (for the various estimation classes) have been calculated, the DSE can be used to calibrate the census as described above. Estimates of n_{00} will be derived from models relating multiple-source omission status to two-source omission status to be described in the next section. Two general assumptions are required for valid inferences to be drawn from the information derived from third sample in the three-source sample. First, the three-source sample (the blocks included in the MSE) must be a random sample (with some known probability mechanism) of the two-source sample (the blocks included in the DSE). (This assumption can be relaxed in some models.) Second, the act of collecting the third list must not have a causal effect upon the probability of inclusion in the first two sources (Census and PES). This assumption is made plausible if (1) the process involved in collecting the third list is inherently separate from other Census operations,

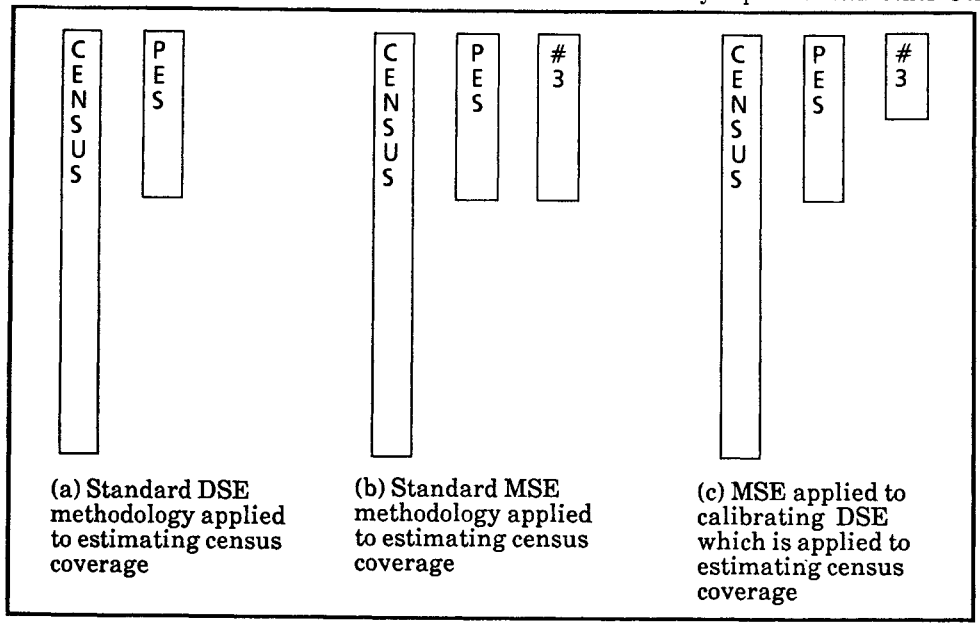

Figure 1: Three designs for calibrating the Census

as would be the case if already existing administrative lists are processed for matching as a third source, or (2) the collection of the third list is later in time than the operations involved in the Census and PES.

If the first source is the Census and the second source is the PES, the third source might be (1) an administrative list, (2) a combination of several administrative lists designed to give broad coverage across different strata of the population, such as a combination of school registration, welfare rolls, income tax lists, and Social Security enrollment (a sub-sample "mega-list"), or (3) some special list derived from a more intensive enumeration procedure than the Census or PES, such as an anthropological participant-observer study. The Census Bureau currently is investigating the practicality of all of these sources for coverage evaluation purposes. Since the third source is only collected on a subsampled basis, problems of cost and difficulty of access to additional sources should be mitigated.

We have thus far assumed a simple monotone pattern of sampling for the three sources, that is, one in which MSE blocks are drawn from DSE blocks, which are drawn from Census blocks. (The term is used in the sense of "monotone patterns of missing data" (Little and Rubin 1986), because Census omission status is more observed than PES status, which is more observed than status in Source #3.) More complicated designs are also possible. Models and estimation procedures for these designs are discussed in Zaslavsky (1989).

5. Models for omissions with multiple sources.

5.1. Monotone sampling pattern, no block effects. Coverage estimates from the DSE can be applied to the Census in a very straightforward manner. The only information available for a non-PES Census block is the observed count (by estimation class) and so the estimated coverage rates for each class can be applied to estimate total population by class and block. (In practice, some form of regression smoothing may be used to combine coverage data from different classes in estimating class coverage rates.)

One might proceed analogously with three sources to calculate an estimated coverage rate for the combined Census and PES using only the three-source blocks. The coverage rate of the Census as a fraction of units captured in either the Census or the PES could be estimated in the usual way from the two-source blocks. The overall coverage rate of the Census could then be estimated by multiplying these two coverage rates. Estimation of the coverage rate may be split into stages in this way because the likelihood for the two coverage rates factors. A simple parametrization of the three-source coverage process makes this argument precise. Suppose $\{m_{ijk}\}$ are expected cell counts in the three-way table for omission with three sources, as in Section 3. The following probabilities correspond to the coverage rates described in the last paragraph:

p_1 = P[unit is included in Census | included in Census and/or PES]
$= m_{1++}/(m_{+++} - m_{00+})$

p_2 = P[unit is included in Census and/or PES | included in Census, PES, and/or third source]
$= (m_{+++} - m_{00+})/(m_{+++} - m_{000})$

p_3 = P[unit is included in Census, PES, and/or third source]
$= (m_{+++} - m_{000})/m_{+++}$

where g=1,2,3 indexes the samples in which 1,2, and 3 sources respectively are observed; x_{ig} is observed for $i \leq g$. Define also the following parameters:

$$m_{1g} = \mathbf{E}X_{1g} = m_{1++}, m_{2g} = \mathbf{E}X_{2g} = m_{01+},$$
$$m_{3g} = m_{001},$$
$$N_g = \text{total count of all units} = m_{+++}.$$

Then $m_{3g} = p_3 N_g, m_{2g} = p_2 p_3 N_g, m_{1g} = p_1 p_2 p_3 N_g$, and the log-likelihood is

$$\begin{aligned} l(p_1,p_2,p_3 \mid x, N) = & \\ & [(x_{12} + x_{13})\log p_1 + (x_{22} + x_{23})\log(1 - p_1)] \\ & + [(x_{12} + x_{22})\log p_2 + x_{33}\log(1 - p_2)] \\ & + f_1(x_{11}, p_1, p_2, p_3, N_1) + f_2(x_{22}, p_2, p_3, N_2) \\ & + f_3(x_{33}, p_3, N_3). \end{aligned} \tag{3}$$

The final set of terms involve the unknown total counts in each sample and therefore are of no help for inference on p (in the absence of strong prior information on these counts). Ignoring those terms, the likelihood factors. The first term is the log-likelihood for p_1 based on the two- and three-source samples, and yields the MLE $p_1 = (x_{12} + x_{13})/(x_{12} + x_{13} + x_{22} + x_{23})$, the fraction observed in the Census in those samples. Similarly, the second term is the log-likelihood for p_2 based on the three-source sample, and yields the MLE $p_2 = (x_{13} + x_{23})/(x_{13} + x_{33})$, the fraction observed in the Census and/or the PES in that sample.

Note that this pattern of observations gives no information on p_3. The sensitivity of coverage estimates should be checked under a range of plausible alternative values for p_3, based on different prior assumptions about the relationship of the coverage rate of the three sources together to the coverage rates of individual sources. Obviously p_3=1 would yield the upper limit for coverage. Various plausible *ad hoc* lower limits might be proposed as functions of p_1, p_2 and other observed proportions. For example, we might suppose that a=P[included in third source]=k_1·P[included in third source | included in Census or PES], and b=P[included in Census or PES | not included in third source]=k_2·P[included in Census or PES | included in third source]=$k_2 p_2$, for some constants k_1, k_2 selected *a priori*. Then $(1-p_3)=(1-a)(1-b)$. Another possible estimate for p_3 would be that derived from the multi-list estimate of the thrice-unobserved cell.

Whatever value is assumed for p_3, the coverage rate of the Census is then estimated as $r = p_1 p_2 p_3$. Our assumption is that all of the estimates of p_3 are close enough

to 1 that sensitivity to the exact form chosen would be minimal.

5.2 Monotone sampling pattern, with block effects. If the DSE of the Census coverage rate in the three-source (MSE) blocks differs from the corresponding estimate in the entire two-source sample, one would have reason for concern that the MSE coverage estimate in the three-source blocks would not be correct for the entire two-source sample. For example, if, due to random variation, the third source was collected from a sample of blocks with an atypically high coverage rate on the Census, the coverage rate of the PES for the units missed by the Census might also be atypically high; this would give misleading coverage estimates when applied to the entire sample of PES blocks.

In this subsection we will extend the model of Section 5.1 to the situation in which the underlying coverage rates differ by blocks, that is, there is block-to-block variation in coverage beyond that due to binomial variation around a single coverage rate. The term "block" here refers to an arbitrarily defined local area whose population is more homogeneous in coverage probability than the entire Census area. Thus a block might be, but would not necessarily be, the same as a Census enumeration block.

The situation is analogous to estimation of the mean of a variable Y which is partially missing when there are completely-observed background variables $\mathbf{W}$. In the latter case, a common approach is to calculate a regression adjustment (Cochran 1977), also known as a covariance adjustment; the linear regression of Y on $\mathbf{W}$ can be estimated from the fully observed cases and then the population mean $\bar{Y}$ can be estimated by substituting the mean of $\mathbf{W}$ into the regression. The regression adjustment may improve the efficiency of the estimate of $\bar{Y}$ when the fully-observed blocks are a simple random sample of all blocks. Furthermore, the regression adjustment may be consistent even when the probability that Y_i is observed depends on $\mathbf{W}_i$, whereas the unadjusted estimate of $\bar{Y}$ almost certainly will not be. The complicating factor here is that the variables of interest p_1, p_2 are not observed directly, but only the counts observed in the different sources, which are multinomially distributed around the underlying rates. Therefore, the covariance structure will be introduced through a heirarchical Bayes model.

Let the argument b index Census blocks, and let B_1, B_2, B_3 be the sets of blocks included in the 1-, 2- and 3-source samples respectively. Let $x_i(b)$ represent the count of units included in any of the first i sources (i=1, 2, or 3) in a particular block b; $x_i(b)$ is observed if $b \in B_g$ and $i \leq g$. Let $p_1(b)$ and $p_2(b)$ be the values of p_1 and p_2 in block b. We will denote by $\mathbf{p}_i(B_g)$ and $x_i(B_g)$ the vectors of values of $p_i(b)$ and $x_i(b)$ respectively in all the blocks in set B_g.

As in the last section, $\mathbf{x}$ and $\mathbf{p}$ are related by the binomial sampling distributions of $\mathbf{x}$, with $X_1(b) \sim$ Binomial$(x_1(b) + x_2(b), p_1(b))$ and $(X_1(b) + X_2(b)) \sim$ Binomial$(x_1(b) + x_2(b) + x_3(b), p_2(b))$.

Suppose further that $p_1(b)$ and $p_2(b)$ are related by a multivariate normal distribution on the logit scale,

$$(z_{1b}, z_{2b})' \sim N(\mu, \sigma)$$
$$\text{where } z_{ib} = \log(p_i(b)/(1 - p_i(b))). \tag{4}$$

In this heirarchical Bayes model, the hyperparameter is (μ, σ), the parameters are $\mathbf{p}$, and the data are $\mathbf{x}$.

The posterior distribution of the probabilities $\mathbf{p}$ is the product of the N(μ, σ) prior and the multinomial likelihood for each component for which there is information in the sample. The likelihood takes the same form as (3) except that each factor has its own p_i(b) and so there must be a separate factor for each block. In the following expression for the posterior distribution of $\mathbf{p}$, each factor corresponds to one of the sets of blocks B_1, B_2, B_3.

$$\begin{aligned} P[\mathbf{p} \mid \mu, \sigma, x] = {} & P[\mathbf{p}_1(B_1), \mathbf{p}_2(B_1) \mid \mu, \sigma] \cdot \\ & P[\mathbf{p}_1(B_2) \mid \mu, \sigma, x_1(B_2), x_2(B_2)] \cdot P[\mathbf{p}_2(B_2) \mid \mu, \sigma] \cdot \\ & P[\mathbf{p}_1(B_3) \mid \mu, \sigma, x_1(B_3), x_2(B_3)] \cdot \\ & P[\mathbf{p}_2(B_3) \mid \mu, \sigma, x_1(B_3), x_2(B_3), x_3(B_3), \mathbf{p}_1(B_3)]. \end{aligned}$$

Those components of $\mathbf{p}$ for which there is no information in the data make no contribution to the likelihood inference for (μ, σ) and may be integrated out to obtain the following probability distribution for the remaining components:

$$\begin{aligned} P[\mathbf{p}_1(B_2 \cup B_3), \mathbf{p}_2(B_3) \mid \mu, \sigma, x] = {} & \\ & P[\mathbf{p}_1(B_2 \cup B_3) \mid \mu_1, \sigma_{11}, x] \cdot \\ & P[\mathbf{p}_2(B_3) \mid a, c, \sigma_{22.1}, \mathbf{x}, \mathbf{b}_1(B_3)] \end{aligned} \tag{5}$$

where $c = \sigma_{12}/\sigma_{11}$ and $a = \mu_2 - c\mu_1$ are the coefficients of the regression of z_{2b} on z_{1b}, and $\sigma_{22.1}$ is the residual variance of that regression.

Since the normal prior is not conjugate to the multinomial likelihood, the posterior distribution of $\mathbf{p}$ is not analytically tractable, nor is that of the logits z. However, each factor of (5) may be approximated by a normal distribution for z, using the derivatives of the logit-multinomial likelihood for z to approximate a quadratic log-likelihood in the neighborhood of the current estimate of the posterior mode.

Inference for μ and σ may now proceed by two approaches. The maximum likelihood approach uses the EM algorithm (Dempster, Laird and Rubin 1977) in two stages to calculate the MLEs of the hyperparameters (μ, σ). The first factor of (5) is the marginal posterior distribution of $\mathbf{p}_1(B_2 \cup B_3)$ and involves only the hyperparameters (μ_1, σ_{11}), whose likelihood depends only on the moments of $\mathbf{p}_1(B_2 \cup B_3)$. Therefore EM can be applied to calculate iteratively the MLE for (μ_1, σ_{11}); at the E step the expected first two moments of $\mathbf{p}_1(B_2 \cup B_3)$ given (μ_1, σ_{11}) are calculated (by extracting the appropriate means and variances from the normal approximation to the posterior distribution), and at the M step the MLE for (μ_1, σ_{11}) is calculated from those sufficient statistics. The second factor of (5) gives the conditional posterior distribution of $\mathbf{p}_2(B_3)$ and involves the hyperparameters $(a, c, \sigma_{22.1})$; the sufficient statistics here are the joint first and second moments of z_{1b} and z_{2b} in B_3. Thus EM may be applied again to calculate the MLE for $(a, c, \sigma_{22.1})$ with

(μ_1, σ_{11}) fixed at its already-calculated MLE.

Another approach generates draws from the joint posterior distribution of the hyperparameters. This method is based on the algorithm of Tanner and Wong (1987). A prior distribution must be assumed for the parameters (μ, σ); the usual non-informative priors would be an improper uniform distribution for μ and the Jeffries prior for σ (probability proportional to $\mid \Sigma \mid^{-1}$). First, draws are taken from the joint marginal distribution of $(\mu_1, \sigma_{11}, \mathbf{p}_1(B_2 \cup B_3))$, by an iterative process of alternately drawing parameters $\mathbf{p}_1(B_2 \cup B_3)$ (conditional on hyperparameters and data) and hyperparameters μ_1, σ_{11} (conditional on parameters). Then draws are taken from the joint distribution of $(a, c, \sigma_{22.1}, \mathbf{p}_2(B_3))$ conditional on $\mathbf{p}_1(B_2 \cup B_3)$ by a similar iterative procedure. By this two-step process, draws from the joint distribution of the hyperparameters and the parameters are obtained. This has the advantage of representing uncertainty properly rather than fixing hyperparameters at their MLEs.

Whichever inferential path is taken, the last step is to estimate the average value of the coverage rate $p_1(b)p_2(b)p_3(b)$. A straightforward method of estimating this would be by multiple imputation. For hyperparameter values fixed at their MLE (if maximum likelihood is used) or drawn from their posterior distribution (in the Bayesian approach), values could be imputed for $p_1(b)$ and $p_2(b)$ in each block of the PES, and for $p_3(b)$ if its estimate is expressed as a function of $p_1(b)$ and $p_2(b)$). Then the average over blocks of $p_1(b)p_2(b)p_3(b)$ (weighted by block population in the corresponding class as in Equation (2)) can be calculated. This would represent the average coverage rate for that class.

The factorization (5) permits a relaxation of the sampling scheme for the second source. The parameters (a, c) are regression coefficients of z_2 on z_1. Thus they can be estimated consistently as long as the probability of inclusion of a block does not depend on $p_2(b)$. The implication of this for the selection of the sample of "three-source" blocks is that the probability of inclusion of a block may be permitted to depend upon the DSE $p_1(b)$ of the coverage rate for that block. For example, it might be desirable to oversample from blocks with extremely high and low values of p_1 in order to include more leverage points and therefore improve efficiency of estimation and make estimates more reliable for the blocks with poor coverage.

5.3. Monotone sampling pattern, with block effects and multiple classes. In Sections 5.1 and 5.2, estimation was assumed to be restricted to a single estimation class. However, where there are block-level random effects, it would be reasonable to assume that the random effects are the same, or at least related, for the different classes. Such an assumption yields a more efficient procedure since the data from all classes are combined in estimating the random effect for a block. In this subsection we consider a model representing this assumption by assuming an additive effect for class on the logit scale.

Let p_{ibs} be the value of p_i corresponding to block b and stratum (or estimation class) s. Then the extended model is

$$\log(p_{ibs}/(1 - p_{ibs})) = \mu_{is} + z_{ib}, \ (z_{1b}, z_{2b})' \sim N(0, \Sigma),$$

where the μ_{is} are the mean logits for the various classes. Note that there are now two parameters for each class (related to the average values of z_1 and z_2 for that class) but that there are only two parameters per block for the random effects regardless of the number of classes.

Estimation of the parameters then proceeds along the same lines described in Section 5.2. The average coverage rate for class s can then be calculated by averaging $p_{1bs}p_{2bs}p_{3bs}$, weighted now by block population *for that class.*

6. References

Bishop, Yvonne M.M., Feinberg, Stephen E., and Holland, Paul W. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge.

Citro, Constance and Cohen, Michael (1983), *The Bicentennial Census: New Directions for Methodology in 1990*, National Academy Press, Washington, D. C.

Cochran, William G. (1977), *Sampling Techniques*, John Wiley and Sons, New York.

Cowan, Charles D. and Malec, Donald (1986), "Capture-Recapture Models when both Sources have Clustered Observations," *Journal of the American Statistical Association* 81:347-353.

Dempster, Arthur P., Laird, Nan M, and Rubin, Donald B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society B* 39:1-38.

Ericksen, Eugene P. and Kadane, Joseph B. (1985), "Estimating the Population in a Census Year: 1980 and Beyond," *Journal of the American Statistical Association* 80:98-109.

Feinberg, Stephen E. (1972), "The Multiple Capture Census for Closed Populations and Incomplete 2^k Contingency Tables," *Biometrika* 59:591-603.

Hogan, Howard and Wolter, Kirk (1988), "Measuring Accuracy in a Post-Enumeration Survey," *Survey Methodology* 14:99-116.

Little, Roderick J.A. and Rubin, Donald B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Seber, George (1973), *The Estimation of Animal Abundance*, Hafner Press, New York.

Tanner, Martin A. and Wong, Wing Hung (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82:528-541.

Wittes, Janet (1970), "Estimation of Population Size: The Bernoulli Census," Ph.D. Thesis, Harvard University, Cambridge.

Zaslavsky, Alan M. (1989), "Multiple-System Methods for Census Coverage Evaluation," *Proceedings, Fifth Annual Research Conference*, Bureau of the Census, Washington.