BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-83-07

COVERAGE ERROR MODELS FOR CENSUS AND SURVEY DATA

by Kirk M. Wolter

Chief, Statistical Research Division
U.S. Bureau of the Census

This paper was prepared for and presented at the 44th Session of the International Statistical Institute, September 1983, in Madrid, Spain.

Recommended by: Barbara A. Bailar
Report completed: July 1, 1983
Report issued: September 1983

July 1, 1983

COVERAGE ERROR MODELS FOR CENSUS AND SURVEY DATA

Kirk M. Wolter

U.S. Bureau of the Census

1. INTRODUCTION

The problem of coverage error in surveys and censuses has become an important statistical issue, supported by the fact that the U.S. Bureau of the Census has been sued in Federal court more than 50 times regarding the completeness of the 1980 census. The main purpose of this article is to discuss certain aspects of coverage error and to provide careful exposition of some alternative statistical models for such error.

Coverage error has been studied for several decades. In the U.S., coverage error has been estimated for each of the past four decennial censuses of population and housing, starting with the 1950 census. In Canada, coverage error has been estimated for each of the past five quinquennial censuses, starting with the 1961 census. Other countries such as Australia, Austria, Finland, and Korea have also produced estimates of the coverage error associated with their population censuses. Despite the apparent vast amount of research on coverage error, previous authors have not, to our knowledge, presented explicit statistical models for such error, although models have been implicit in all of the previous work.

The models we discuss are equivalent to the capture-recapture models employed in estimating the size and density of wildlife populations, and to the dual-system models employed in estimating the number of human vital events. They are also related to the log-linear models employed in the analysis of discrete multivariate data. Capture-recapture models originated in the 17th century, and the modern development dates from Peterson (1896), Lincoln (1930), and Schnabel (1938). Excellent recent reviews are given by Seber (1973) and Otis et al. (1978). The application to human vital events was initiated by the pioneering work of Sekar and Deming (1949). Extensive recent discussion is presented by Marks, Seltzer, and Krotki (1974). Bishop, Fienberg, and Holland (1975) discuss the subject of log-linear models and their relation to the capture-recapture problem.

In Section 2 we present the basic coverage error model and discuss several important special cases that are useful in estimating the level of error. The model denoted M_{th} is the one employed implicitly in several of the previous coverage error studies. The basic model is extended in Section 3 to include the sampling error associated with a postenumeration survey. Statistical adjustments to census data designed to compensate for coverage error are discussed briefly in Section 4. We expose a clear connection between the basic coverage error model and the method of small domain estimation known as synthetic estimation. The paper closes with a general summary in Section 5, where we discuss future research possibilities as well as possibilities for relaxing some of the assumptions imposed in earlier sections.

2. COVERAGE ERROR MODELS

We consider a given human population U, and let N denote the size of U. It is assumed that N is fixed but unknown, and the main problem to be addressed is that of estimating N. The reader will note that this problem differs fundamentally from that treated in the traditional literature on survey sampling, where N is assumed known and the main problem is to estimate other parameters of the finite population U.

We assume that a census of U is conducted at a particular point in time, and that the census, more or less, attempts to enumerate each and every individual in U . Let G_A denote the general conditions present in and during the census enumeration. For a variety of reasons some individuals are missed by the census, however, and the difference between the census count and N is defined as the error of coverage. The census count may be greater than the true N (an overcount), but usually the error mechanism is such that the census count is less than the true N (an undercount).

One of the pernicious features of coverage error is that internal measures cannot be computed from the census data itself. In order to produce measures of coverage error we assume additional information in the form of a sample survey of the same population U. The survey either preceeds (a preenumeration survey) or follows (a postenumeration survey) the census, but in either case employs the same reference period as the census. For a variety of reasons, some individuals in U are missed by the survey procedures. We let G_B denote the general conditions present in and during the survey enumeration.

In the remainder of this section we describe some basic coverage error models that treat jointly the census and sample survey processes. We discuss estimation of the true population size N and make explicit the assumptions used. For notational convenience, we shall refer to the census population as List A and the survey population as List B. For now we shall assume that the survey is in fact a complete enumeration of List B, i.e., a conceptal enumeration of U given general conditions G_B. In Section 3 we shall consider the case where the survey involves observation of only a part of List B.

2.1 M_g: The General Model

The general coverage error model is set forth in the following paragraphs.

(i) (The Closure Assumption) We assume the population U is closed and of fixed size N. In practice, this implies that the census' reference period is well defined and that no recruitment (birth or immigration) or losses (death or emigration) occur during that period.

(ii) (The Multinomial Assumption) Let ξ_i denote the following multinomial distribution:

		List B		
		in	out	
List A	in	p_{i11}	p_{i12}	p_{i1+}
	out	p_{i21}	p_{i22}	p_{i2+}
		p_{i+1}	p_{i+2}	1

We assume that the joint event that the i-th individual is in List A or not and in List B or not is correctly modeled by the distribution ξ_i.

(iii) (Type I Independence) We assume that List A and List B are created as a result of N mutually independent multinomial events, utilizing distributions $\xi_1, \xi_2, \ldots, \xi_N$. The resulting data are

$$\begin{array}{llcccl} & & \multicolumn{2}{c}{\text{List B}} & & \\ & & \text{in} & \text{out} & & \\ \text{List A} & \text{in} & x_{11} & x_{12} & x_{1+} & \\ & \text{out} & x_{21} & x_{22} & x_{2+} & \\ & & x_{+1} & x_{+2} & x_{++} = N & \end{array} \qquad (2.1)$$

where $x_{ab} = \Sigma x_{iab}$, and x_{iab} is an indicator random variable signifying whether or not the i-th individual is in cell (a,b), for a,b = 1,2, +. The census count x_{1+} is considered observable. The cell counts x_{11}, x_{12}, and x_{21} are considered observable on the basis of the survey data and subsequent matching to the census. The cell count x_{22}, and thus the size of the target population N, is considered unknown and to be estimated on the basis of the model.

The reader will note that the census count x_{1+} is regarded as a random-variable under the model, with mean $\mu_{1+} = \Sigma p_{i1+}$ and variance $\sigma^2_{1+} = \Sigma p_{i1+}p_{i2+}$.

(iv) (The Matching Assumption) We assume that it is possible to match correctly the sample survey results to the census results. That is, we assume that it is possible to make a determination, without error, of which individuals recorded in the sample survey are present in the census, and which are not.

(v) (Spurious Events Assumption) We assume that both List A and List B are void of spurious events, or that such are eliminated prior to estimation. This means that all errors are avoided in recording both the census and survey results. In practice, important spurious events that do occur include a) duplicates on the census list, b) "curbstoned" reports in either the census or sample survey, and c) out-of-scope cases, such as an individual born after the reference period, that are recorded erroneously in the census.

(vi) (The Nonreponse Assumption) There will necessarily be some degree of nonresponse. We assume that sufficient identifying information is gathered about the nonrespondents in both the census and the sample survey to permit exact matching from the survey to the census.

(vii) (The Poststratification Assumption) As will become clear in the sequel, it is often desirable to employ some poststratification in the estimation of N. For example, one may wish to poststratify on age, producing age-specific population estimates, then aggregating to give an estimate of the total population N. We assume that any variable employed for poststratification is correctly recorded for all individuals in both the census and the sample survey.

Readers familiar with the capture-recapture literature or the dual-system literature will recognize many of these assumptions. For further discussion see Otis et al. (1978), Seltzer and Adlakha (1974), and Cowan and Bettin (1982).

Unfortunately the model M_g is grossly underidentified and further assumptions are needed to estimate the true population size N. In Sections 2.2 to 2.6 we consider various special cases of the general model, each identified through additional restrictions on the multinomial process. The taxonomy of models is due to Pollock (1974).

2.2 M_0: The Equal Catchability Model

In the equal catchability model, each individual in the population U has the same probability of capture (or coverage) in both List A and B, and the two lists are generated independently. The additional assumptions are

(viii) (Type II Independence) The event of being included in List A is independent of the event of being included in List B. That is, the cross-product ratio satisfies $\theta_i = p_{i11}p_{i22}/(p_{i12}p_{i21}) = 1$, for $i = 1, \ldots, N$.

(ix) The marginal probabilities of capture satisfy $p_{i1+} = p_{i+1} = p$ for $i = 1, \ldots, N$.

The likelihood associated with this model is

$$L_0(N,p) = \binom{N}{x_{11}\ x_{12}\ x_{21}} p^{x_\bullet}(1-p)^{2N-x_\bullet},$$

where $x_\bullet = 2x_{11}+x_{12}+x_{21}$ is the total number of captures in both lists. The sufficent statistic for this problem is $(x_1, x_\bullet)$, where $x_1 = x_{11} + x_{12} + x_{21}$ denotes the total number of distinct captures, and the maximum likelihood estimators are

$$\hat{N}_0 = \left[\left[\frac{x_\bullet^2}{4(x_\bullet - x_1)}\right]\right],$$

and $\hat{p} = 2(x_\bullet - x_1)/x_\bullet$, where $[[\cdot]]$ denotes the greatest integer function.

2.3 M_b: The Trap Response Model

One of the criticisms of the simple model M_0 is that the assumption regarding Type II independence may not obtain in surveys and censuses of human populations. The concern is that both Lists A and B may tend to miss or capture the same individuals. In the capture-recapture literature, this tendency is described as a behavioral response, and individuals are said to be "trap happy" or "trap shy." That is the odds of capture in List B (survey population) given capture in List A (census) are either greater or less than the odds of capture in List B given noncapture in List A. This problem is thought to be more serious in human populations than in wildlife populations.

One possibility for modeling this situation is to impose assumption (x) in place of assumptions (viii) and (ix):

(x) The probability of first capture is the same for each individual in the population. That is,

Pr{i-th individual is captured in List A} = $p_{i1+} = p$
Pr{i-th individual is captured in List B|i-th individual is not captured in List A}
$= p_{i21}/p_{i2+} = p$

for $i = 1, \ldots, N$. The probability of second capture is also the same for each individual in the population. That is,

Pr{i-th individual is captured in List B|i-th individual is captured in List A}
$= p_{i11}/p_{i1+} = c$

for $i = 1, \ldots, N$. The cross-product ratio for this model is $\theta_b = c(1-p)(1-c)^{-1} p^{-1}$, and $\theta_b > 1$ (positive association between lists) whenever $c > p$ and $\theta_b < 1$ (negative association between lists) whenever $c < p$. Thus, the lists are positively associated whenever the odds of second capture are greater than the odds of first capture, and the lists are negatively associated whenever the odds of second capture are less than the odds of first capture. Implicit in this assumption is the condition that List B follows List A in time.

The likelihood associated with this model is

$$L_b(N,p,c) = \binom{N}{x_{11}\ x_{12}\ x_{21}} p^{x_1} c^{x_{11}} (1-c)^{x_{1+}-x_{11}} (1-p)^{2N-x_1-x_{1+}} .$$

The sufficient statistic for this problem is (x_1, x_{1+}, x_{11}), and the maximum likelihood estimators are

$$\hat{N}_b = [[x_1 \{1 - (\frac{x_1-x_{1+}}{x_{1+}})^2\}^{-1}]],$$

$$\hat{p}_b = (2x_{1+} - x_1)x_{1+}^{-1}, \text{ and } \quad \hat{c} = x_{11}x_{1+}^{-1}.$$

2.4 M_t: The Petersen Model

A second extension of the simple model M_0 occurs when we assume time variation in the capture probabilities. We assume (i) - (viii) and

(xi) The capture probabilities satisfy $p_{i1+} = p_{1+}$ and $p_{i+1} = p_{+1}$ for $i = 1, \ldots, N$.

The likelihood associated with model M_t is

$$L_t(N,p_{1+},p_{+1}) = \binom{N}{x_{11}\ x_{12}\ x_{21}} p_{1+}^{x_{1+}} p_{+1}^{x_{+1}} (1-p_{1+})^{N-x_{1+}} (1-p_{+1})^{N-x_{+1}} .$$

The sufficient statistic is now (x_1, x_{1+}, x_{+1}), and the maximum likelihood estimators are

$$\hat{N}_t = [[\frac{x_{1+}x_{+1}}{x_{1+}+x_{+1}-x_1}]] = [[\frac{x_{1+}x_{+1}}{x_{11}}]] ,$$

$$\hat{p}_{1+} = (x_{1+} + x_{+1} - x_1)x_{+1}^{-1}, \quad \text{and} \quad \hat{p}_{+1} = (x_{1+} + x_{+1} - x_1)x_{1+}^{-1}.$$

The estimator $\hat{N}_t$ has a long history, dating back several hundred years. In the modern era it has been called variously the Petersen estimator, the Schnabel estimator, the Lincoln index, the Chandrasekar-Deming method, and the dual-system estimator.

2.5 M_h: The Heterogeneity Model

A third extension of the simple model occurs when the capture probabilities are allowed to vary across individuals but not across time. We assume (i) - (viii) and

(xii) The capture probabilities satisfy $p_{i1+} = p_i$ for $i = 1, \ldots, N$.

The likelihood for model M_h is

$$L_h(N,p_1,p_2, \ldots, p_N) = \prod_{i=1}^{N} p_i^{x_{i.}} (1-p_i)^{2-x_{i.}},$$

where $x_{i.} = 2x_{i11} + x_{i12} + x_{i21}$ denotes the total number of times the i-th individual was captured. It is clear that $x_{i.} = 0, 1,$ or 2 for $i = 1, \ldots, N$. It is also clear that this likelihood is not useful for estimating the population size N.

A possible procedure for estimating N is to divide the population U into L poststrata, such that model M_0 holds within each stratum. To estimate the size of the total population we 1) estimate the size of each stratum utilizing a model M_0 estimator, and 2) aggregate over strata. But for some populations, it may not be possible to formulate an appropriate poststratification scheme such that model M_0 holds within strata.

2.6 M_{tb}, M_{th}, M_{bh}, M_{tbh}: Combination Models

Numerous additional models may be specified by combining the features of the basic models M_0, M_b, M_t, and M_h. We shall discuss these models only briefly because they are not generally useful for the coverage error problem. In capture-recapture studies of wildlife populations, however, it is often possible to employ more than two lists (or captures) in the analysis, and some of the combination models are useful in such applications.

Model M_{tb} combines time variation in the capture probabilities with a behavioral response. The multinomial distribution for this model is specified by $p_{iab} = p_{ab}$, for a,b = 1,2,+, and the distribution is assumed to be homogeneous across individuals i = 1, ..., N. The cross product ratio $\theta = p_{11}p_{22}/(p_{12}p_{21})$ indicates the nature of the behavioral response, with $\theta > 1$ signifying "trap happy" and $\theta < 1$ signifying "trap shy." The simple model M_t arises in the special case of Type II independence, i.e., $\theta = 1$.

Model M_{th} has been used implicitly in most of the previous studies of coverage error. We assume (i) - (viii) and

(xiii) (Type III Independence) The marginal probabilities associated with Lists A and B are uncorrelated across individuals in the sense that

$$\sigma(p_{1+},p_{+1}) = N^{-1} \Sigma (p_{i1+} - \bar{p}_{1+})(p_{i+1} - \bar{p}_{+1}) = 0,$$

where $\bar{p}_{ab} = N^{-1} \Sigma p_{iab}$ for (a,b) = (1,+), (+,1).

Given these assumptions, the Petersen statistic $\hat{N}_t$ is a consistent estimator of N. In the more general case where Type III Independence fails, the estimator N_t is not consistent and the leading term in the bias is given by

$$\text{Bias } \{\hat{N}_t\} = - N \sigma(p_{1+},p_{+1})/\{\sigma(p_{1+},p_{+1}) + \bar{p}_{1+}\bar{p}_{+1}\} .$$

In applications involving human populations, $\sigma(p_{1+},p_{+1})$ is thought to be positive, implying a downward bias in the estimator of population size.

Model M_{th} was first treated in the vital events literature by Sekar and Deming (1949). They referred to the downward bias in N_t as the "correlation bias" and poststratification was suggested as a means of reducing the bias. This terminology and the concept of poststratification have since become standard features of the dual system model in the vital events literature. As we have seen, however, "correlation bias" is somewhat of a misnomer because the bias arises from heterogeniety in the capture probabilities across individuals, not from a behavioral response between Lists A and B.

Model M_{bh} combines a behavioral response with heterogeniety in the capture probabilities. Without imposing additional assumptions, this model is not useful for estimating N.

Finally, we have the model M_{tbh} which allows variation in the capture probabilities by time and individual, as well as a behavioral response. This is the same as the general model M_g introduced in Section 2.1. As was stated in that section, this model is too general to be useful for estimating N, which is unfortunate, because it is probably the most realistic model in the context of surveys and censuses of human populations.

In passing we remark that all of the models defined in this section may be extended to include more than two lists. See Otis et al. (1978) for discussion of such possibilities. For the vital events application and the coverage error application, however, there are rarely more than two lists and, consequently, this article has concentrated specifically on models for the two-list problem.

3. DETAILS OF MODEL M_t

In this section we develop fully one of the coverage error models introduced in Section 2. We choose model M_t for this development because of its historical importance both in the vital events literature and in the previous studies of coverage error. The development also applies to model M_{th}, provided that enough prior information is available to poststratify the population to the point where model M_t holds within strata.

3.1 The Sample Survey

In Section 2 model M_t was discussed in terms of an application where List B was enumerated entirely. This condition is not realistic in most coverage error studies, and we shall now assume that a sample is selected from List B and that only the sample cases are enumerated and matched to List A. We continue to assume that List A is observable. We also continue to impose all other assumptions about M_t described in Section 2, including (i) - (viii) and (xi). Of the population quantities in (2.1), only the census total x_{1+} is considered known. The survey population total x_{+1} is now considered unobservable, as are the totals x_{11}, x_{12}, x_{21}, but all are estimable based on the survey data. The quantities x_{22} and N are to be estimated on the basis of model M_t.

To concentrate on essentials, we consider a simple survey design. Let the survey population (List B) be divided into M areal clusters and assume that a simple random sample of size m is selected without replacement. Let the survey population be enumerated entirely within the selected clusters. The ensuing development easily extends to more complicated sampling designs, however, this simple design will serve well for communicating the main ideas.

The reader will note that the list of clusters is assumed to be complete. Each member of U is assumed to belong to one and only one cluster and there are no members of the population U that are not covered by one of the M clusters. In both the census and sample survey, however, not all true members of the selected clusters are enumerated. Only members who are in List A and B, respectively, are enumerated.

We expand the notation utilized in Section 2 in order to accommodate the clustering. Let

$$x_{ijab} = 1 \quad \text{,if the j-th individual in the i-th cluster is in the cell (a,b) of Table (2.1)}$$
$$= 0 \quad \text{, otherwise,}$$

for a,b = 1, 2, +. Then define

$$x_{ab} = \sum_{i}^{M} x_{iab} = \sum_{i}^{M} \sum_{j} x_{ijab}$$

for a,b = 1, 2, +. The reader will note that

- o x_{i+1} is the size of the i-th cluster for the survey population
- o x_{i1+} is the size of the i-th cluster for the census population
- o x_{i++} is the true size of the i-th cluster.

For the selected clusters, the quantities x_{i11}, x_{i21}, and x_{i+1} are observable on the basis of the survey enumeration and subsequent matching to the census population (List A). x_{i22} is clearly not observable. Also, we assume that x_{i12} and $x_{i1+}$are not observable.

We shall let $\hat{x}_{11}$, $\hat{x}_{21}$, and $\hat{x}_{+1}$ denote the usual design-unbiased estimators of x_{11} , x_{21}, and x_{+1}, respectively. For example

$$\hat{x}_{11} = \frac{M}{m} \sum_{i}^{m} \sum_{j} x'_{ij11} = \frac{M}{m} \sum_{i}^{m} x_{i11}.$$

We shall employ the design-unbiased estimator $\hat{x}_{12} = x_{1+} - \hat{x}_{11}$ of x_{12}. If List A is mapped onto the selected clusters and matched to List B, then an alternative estimator of x_{12} is available. This situation is not assumed here, however.

From Section 2 we recall that the model M_t estimator of N is $\hat{N}_t = x_{1+}x_{+1}x_{11}^{-1}$. This estimator is not available unless List B is enumerated completely (i.e., m = M). Replacing x_{+1} and x_{11} by the corresponding design-unbiased estimators gives the sample-based estimator $\tilde{N}_t = x_{1+}x_{+1}x_{11}^{-1}$. One of the main objectives of this section is to discuss the statistical properties of $\tilde{N}_t$. As we shall see, this estimator is subject to two sources of variability: sampling variability and model variability.

To facilitate the discussion it is convenient to introduce additional notation. We shall use E_ξ and V_ξ to indicate expectation and variance operators with respect to the model M_t distribution ξ; E_p and V_p to indicate expectation and variance operators with respect to the sampling design p; and E and V to denote total expectation and total variance, respectively.

Utilizing this notation we are able to obtain the following properties of $\tilde{N}_t$ and N_t.

Theorem 3.1 The first- order approximate expectations and variances of $\hat{N}_t$are given by

(i) $E_p\{\hat{N}_t\} = \hat{N}_t$ (iv) $V_p\{\hat{N}_t\} = 0$

(ii) $E_\xi\{\hat{N}_t\} \doteq N + \dfrac{p_{2+}p_{+2}}{p_{1+}p_{+1}}$ (v) $V_\xi\{\hat{N}_t\} \doteq N \dfrac{p_{2+}p_{+2}}{p_{1+}p_{+1}}$

(iii) $E\{\hat{N}_t\} \doteq N + \dfrac{p_{2+}p_{+2}}{p_{1+}p_{+1}}$ (vi) $V\{\hat{N}_t\} \doteq N \dfrac{p_{2+}p_{+2}}{p_{1+}p_{+1}}$. //

Theorem 3.2 The first-order approximation expectations and variances of $\tilde{N}_t$ are given by

(i) $$E_p\{\tilde{N}_t\} \doteq E_p\{\hat{N}_t\} + \hat{N}_t\left(\frac{V_p\{\hat{x}_{11}\}}{x_{11}^2} - \frac{C_p\{\hat{x}_{+1},\hat{x}_{11}\}}{x_{+1}x_{11}}\right)$$

where

$$V_p\{\hat{x}_{11}\} = M^2(1-f)S_{11}^2/m,$$

$$C_p\{\hat{x}_{+1},\hat{x}_{11}\} = M^2(1-f)S_{+1,11}/m,$$

$$f = m/M,$$

$$S_{11}^2 = (M-1)^{-1} \sum_i^M (x_{i11}-x_{11}/M)^2,$$

$$S_{+1,11} = (M-1)^{-1} \sum_i^M (x_{i+1}- x_{+1}/M)\ (x_{i11}- x_{11}/M),$$

(ii) $$E_\xi\{\tilde{N}_t\} \doteq E_\xi\{\hat{N}_t\} + \left(\frac{\sum_i^{M-m} x_{i++}}{\sum_i^m x_{i++}}\right) \frac{p_{2+}}{p_{1+}p_{+1}}$$

(iii) $$E\{\tilde{N}_t\} \doteq E\{\hat{N}_t\} + \left(\frac{1-f}{f}\right) \frac{p_{2+}}{p_{1+}p_{+1}}$$

(iv) $$V_p\{\tilde{N}_t\} \doteq V_p\{\hat{N}_t\} + \hat{N}_t^2\left(\frac{V_p\{\hat{x}_{+1}\}}{x_{+1}^2} + \frac{V_p\{\hat{x}_{11}\}}{x_{11}^2} - 2\,\frac{C_p\{\hat{x}_{+1},\hat{x}_{11}\}}{x_{+1}x_{11}}\right)$$

(v) $$V_\xi\{\tilde{N}_t\} \doteq V_\xi\{\hat{N}_t\} + N\left(\frac{\sum_i^{M-m} x_{i++}}{\sum_i^m x_{i++}}\right) \frac{p_{2+}}{p_{1+}p_{+1}}$$

(vi) $$V\{\tilde{N}_t\} \doteq V\{\hat{N}_t\} + N\left(\frac{1-f}{f}\right) \frac{p_{2+}}{p_{1+}p_{+1}} \quad .$$ //

In the statement of Theorem 3.2 note that we have expressed the expectations and variances of $\tilde{N}_t$ as equal to the analogous quantities for N_t plus additional terms. The additional terms represent the statistical effects of employing a sample from List B, rather than a complete enumeration.

It is possible to modify the estimators $\hat{N}_t$ and $\tilde{N}_t$ to reduce or eliminate the bias, though this is usually unnecessary when the sample sizes are large. The only occasion in which an important difference may occur is when the population is poststratified deeply (in order to cope with assumed heterogeneity) and estimates are to be prepared within strata based upon small sample sizes.

As regards estimation of variance, we suggest the following for $\hat{N}_t$:

$$\hat{V}\{\hat{N}_t\} = \hat{V}_\xi\{\hat{N}_t\} = \frac{x_{1+}x_{+1}x_{12}x_{21}}{x_{11}^3} .$$

This is the estimator traditionally presented in the capture-recapture literature. See, e.g., Seber (1973). For $\tilde{N}_t$ we suggest

$$\text{(i)} \quad \hat{V}_p\{\tilde{N}_t\} = M^2 (1-f)s_d^2/m,$$

where

$$s_d^2 = (m-1)^{-1} \sum_i^m d_i^2,$$

$$d_i = \frac{x_{1+}}{\hat{x}_{11}} x_{i+1} - \frac{x_{1+}\hat{x}_{+1}}{\hat{x}_{11}^2} x_{i11},$$

$$\text{(ii)} \quad \hat{V}_\xi\{\tilde{N}_t\} = \frac{x_{1+}\hat{x}_{+1}(\hat{x}_{+1} - \hat{x}_{11})(x_{1+} - \hat{x}_{11})}{\hat{x}_{11}^3} + (\frac{1-f}{f}) \frac{\hat{x}_{+1}x_{1+}^2(\hat{x}_{+1} - \hat{x}_{11})}{\hat{x}_{11}^3}$$

$$\text{(iii)} \quad \hat{V}\{\tilde{N}_t\} = \hat{V}_\xi\{\tilde{N}_t\}$$

or

$$\hat{V}\{\tilde{N}_t\} = \hat{V}_p\{\tilde{N}_t\} + \frac{x_{1+}\hat{x}_{+1}(x_{1+} - \hat{x}_{11})(\hat{x}_{+1} - \hat{x}_{11})}{\hat{x}_{11}^3} .$$

See Marks, Seltzer, and Krotki (1974) for some additional discussion of variance and variance estimation in the context of vital events estimation. In estimating the ξ-variance of $\tilde{N}_t$, we estimate the term

$$(\sum_i^{M-m} x_{i++} / \sum_i^m x_{i++})$$

by (1-f)/f. Thus, $\hat{V}_\xi\{\tilde{N}_t\}$ is a ξp-consistent estimator of $V_\xi\{\tilde{N}_t\}$ in the sense that

$$M^{-1} | \hat{V}_\xi\{\tilde{N}_t\} - V_\xi\{\tilde{N}_t\} | = 0_p(m^{-1/2}) + 0_\xi(m^{-1/2}) ,$$

where the probabilities 0_p and 0_ξ are created by the design and the model, respectively. Obviously, $\hat{V}_p\{\tilde{N}_t\}$ is the usual Taylor series estimator of the design variance, and replication technfques offer alternatives. The estimator $\hat{V}\{\tilde{N}_t\}$ is ξp-consistent for the total variance. The alternative estimator of the total variance is obtained by exchanging the order of expectations in the derivation of $V\{\tilde{N}_t\}$. It can be shown that the difference between the two estimators of total variance, normalized by M^{-1}, differs from zero by terms of order $0_\xi(m^{-1/2})$.

3.2 Relationship to Measurement Error Models

We now develop a connection between coverage error models and the measurement (or response) error model for survey data. The key to the connection is to regard the cluster as the survey reporting unit, rather than the individual or household.

We shall describe the connection in terms of the observations x_{i11}, where similar developments can be given for x_{i21} and x_{i+1}. We may write

$$x_{i11} = x_{i++}p_{11} + e_{i11} \qquad (3.1)$$

for i = 1, ..., M, where e_{i11} is an error with ξ-expectation 0 and ξ-variance $x_{i++}p_{11}(1-p_{11})$. Equation (3.1) is in the form of the response error model for survey data. See Hansen, Hurwitz, and Bershad (1961) for a clear exposition of the model. In the coverage error application, the cluster is treated as the reporting unit and the error arises from the multinomial or ξ-distribution. This differs from the usual survey situation where the individual is the reporting unit and the error is the result of an erroneous response. In both the usual survey situation and the coverage error application, the observations and their error distribution are assumed to be conditional on the general conditions in which the survey or census is conducted. To signify this fact we might have explicitly subscripted our variables by G_A and G_B, but this was not done in order to simplify notation.

The expectation of the response x_{i11} for the i-th cluster, may be written as

$$x_{i++}p_{11} = x_{i++} + \frac{N}{M}(p_{11}-1) + (x_{i++}-\frac{N}{M})(p_{11}-1). \qquad (3.2)$$

Each term on the right side of (3.2) has a specific interpretation in the response error literature. The first term, x_{i++}, is the true value of the i-th unit; the second term, $\frac{N}{M}(p_{11}-1)$, is the fixed bias component; and the third term, $(x_{i++}-\frac{N}{M})(p_{11}-1)$, is the variable bias component associated with the i-th unit.

The total variance of the estimator $\hat{x}_{11}$ may be expressed by

$$V\{\hat{x}_{11}\} = V_p\{\frac{M}{m}\sum_{i}^{m} x_{i++}p_{11}\} + E_p\{\frac{M^2}{m^2}\sum_{i}^{m} x_{i++}p_{11}(1-p_{11})\} \ . \qquad (3.3)$$

The first term on the right side of (3.3) is called the sampling variance in the response error literature, and the second term is called the simple response variance . The response error literature speaks of two additional variance components, the correlated component of response variance and the interaction between response and sampling error, neither of which appear in (3.3). If we allow for a ξ-covariance between the coverage errors in different clusters (e.g., in the case where an interviewer works in two or more clusters), then the total variance $V\{x_{11}\}$ would contain a term analogous to the correlated component for the response error model. This would obviously require a modification to assumption (iii) regarding Type I independence. The interaction term would arise if the selected clusters somehow affect the multinomial capture distribution, so that alternative samples are associated with different distributions. This situation, however, would seem incompatible with the coverage error problem.

3.3 Asymptotic Considerations

In this section we discuss briefly some of the asymptotic issues concerning the estimator $\tilde{N}_t$. First, we define the concept of a sequence of populations. Let $\{u_i\}$ denote a sequence of units, and let x_{i++} denote the true value (unobservable) associated with the i-th unit. Let $\{U_\alpha\}$ denote a sequence of finite populations, with corresponding size $\{M_\alpha\}$, created from the sequence $\{u_i\}$, where $0 < M_1 < M_2 < M_3 < \ldots$. That is, U_1 is composed of the first M_1 elements of $\{u_i\}$, and so on. Note that

$U_1 \subseteq U_2 \subset U_3 \quad \ldots$. See Isaki and Fuller (1982) for some discussion of this type of sequence.

Let $N_\alpha = \Sigma\, x_{i++}$ denote the true size of the α-th population, which is assumed unknown and to be estimated on the basis of model M_t. Let $\{s_\alpha\}$ denote a sequence of samples of corresponding size $\{m_\alpha\}$ created from the sequence of populations by the sampling design "simple random sampling without replacement," where $0 < m_1 < m_2 < m_3 < \ldots$ and $m_\alpha < M_\alpha$ for all α. Note that while the sequence of populations is nested, the sequence of samples is not. Let $f_\alpha = m_\alpha/M_\alpha$ denote the sampling fraction.

For each population U_α in the sequence, we assume that two lists, A_α and B_α, of the individuals in the population are created in accordance with the model M_t. The same multinomial capture distribution is assumed for all populations in the sequence. By analogy with earlier notation, we have

		List B_α in	List B_α out	
List A_α	in	$x_{11\alpha}$	$x_{12\alpha}$	$x_{1+\alpha}$
	out	$x_{21\alpha}$	$x_{22\alpha}$	$x_{2+\alpha}$
		$x_{+1\alpha}$	$x_{+2\alpha}$	$x_{++\alpha} = N_\alpha$

for the α-th population.

List A_α is the census list, and so $x_{1+\alpha}$ is assumed to be observable. List B_α is the survey population, and so $x_{+1\alpha}$ is assumed unobservable but estimable on the basis of the observed sample s_α.

Define the means

$$\hat{\bar{x}}_{11\alpha} = m_\alpha^{-1} \sum_i^{m_\alpha} x_{i11\alpha},\ \hat{\bar{x}}_{+1\alpha} = m_\alpha^{-1} \sum_i^{m_\alpha} x_{i+1\alpha},\ \bar{x}_{1+\alpha} = M_\alpha^{-1} \sum_i^{M_\alpha} x_{i1+\alpha},$$

and the model M_t estimator

$$\tilde{N}_{t\alpha} = M_\alpha \frac{\bar{x}_{1+\alpha}\,\hat{\bar{x}}_{+1\alpha}}{\hat{\bar{x}}_{11\alpha}} .$$

The estimator has the following important properties:

Theorem 3.3 We let model M_t hold and assume all of the other sequence conditions imposed in this section. In addition, we assume

$$\text{(xiv)} \quad \lim_{\alpha\to\infty} M_\alpha^{-1} \sum_i^{M_\alpha} x_{i++}^g$$

exists and is finite for $g = 1, 2, 2+\delta$, for some $\delta>0$;

$$\text{(xv)} \quad \lim_{\alpha\to\infty} f_\alpha = f < 1;$$

and (xvi) (Hájek's (1960) condition)

$$\lim_{\alpha\to\infty} \frac{\sum\limits_{i\,\varepsilon U_{\alpha\tau}} (x_{i++} - \bar{x}_{++\alpha})^2}{\sum\limits_{i\,\varepsilon U_\alpha} (x_{i++}-\bar{x}_{++\alpha})^2} = 0,$$

where $U_{\alpha\tau}$ is the subset of U_α on which the inequality

$$|x_{i++}- \bar{x}_{++\alpha}| > \tau\{(1-f_\alpha)m_\alpha S_\alpha^2\}^{1/2}$$

holds, $\tau > 0$.

Then we have

$$\frac{\sqrt{m_\alpha}}{M_\alpha} (\tilde{N}_{t\alpha} - N_\alpha) \overset{d}{\to} N(0,\ \underset{\sim}{P}'\lim_{\alpha\to\infty} \underset{\sim\alpha}{\Sigma}\ \underset{\sim}{P}),$$

where

$$\underset{\sim}{P} = (p_{1+}^{-1},\ p_{+1}^{-1},\ p_{1+}^{-1}\ p_{+1}^{-1})',$$

$$\bar{x}_{++\alpha} = N_\alpha/M_\alpha\ ,$$

$$S^2_{++\alpha} = (M_\alpha-1)^{-1} \sum_i^{M_\alpha} (x_{i++} - \bar{x}_{++\alpha})^2,$$

and

$$\underset{\sim\alpha}{\Sigma} = \begin{pmatrix} f_\alpha p_{1+}p_{2+}\bar{x}_{++\alpha} & 0 & f_\alpha p_{1+}p_{+1}p_{2+}\bar{x}_{++\alpha} \\ & p_{+1}^2(1-f_\alpha)S^2_{++\alpha}+ p_{+1}p_{+2}\bar{x}_{++\alpha} & p_{1+}p_{+1}^2(1-f_\alpha)S^2_{++\alpha}+ p_{1+}p_{+1}p_{+2}\bar{x}_{++\alpha} \\ \text{symmetric} & & p_{1+}^2p_{+1}^2(1-f_\alpha)S^2_{++\alpha}+ p_{1+}p_{+1}(1-p_{1+}p_{+1})\bar{x}_{++\alpha} \end{pmatrix}$$

//

Theorem 3.3 shows that, $\tilde{N}_{t\alpha}$ is a consistent estimator of the true population size N_α and that the error $(\tilde{N}_{t\alpha} - N_\alpha)$, normalized by $\sqrt{m_\alpha}/M_\alpha$, is asymptotically normally distributed. This result suggests that normal-theory confidence intervals might be constructed for N_α. An open question in this context is whether the lower limit of the the confidence interval should be forced to be larger than the observed sum $x_{1+\alpha} + x_{21\alpha}$? Although, $\tilde{N}_{t\alpha}$ is a consistent estimator of N_α under the model M_t, we note that it is not a design consistent estimator.

3.4 Discussion

In this section we raise some general issues about design-based versus model-based inference, and show how these issues pertain to the problem of coverage error.

In classical survey sampling theory and practice, the sampling design generates the distribution of the statistics of interest. The prediction theory approach to finite population inference, however, treats the design as unimportant and the inference derives almost entirely from an assumed model. Regarding measures of uncertainty, the classical design approach relies upon the sampling variance of the estimator, where the probability is created by the design. For the prediction theory approach, the design is of presampling interest only, and the measure of uncertainty is the model variance (or ξ-variance) given the sample.

From Sections 3.1 to 3.3 it is clear that there are many possibilities for describing the uncertainty in the estimator $\tilde{N}_t$. In what way do the above issues affect inferences about N?

In sharp contrast to most problems of finite population inference, the design is uninformative regarding the estimation of N. The survey estimator $\tilde{N}_t$ is neither design consistent nor unbiased. We regard the lack of consistency as most damaging to the design-based approach. Conversely, the estimator $\tilde{N}_t$ is ξ-consistent for N and its bias is of order m^{-1}. We are forced to conclude, therefore, that the design variance is an inappropriate measure of the uncertainty in $\tilde{N}_t$ as a predictor of N. This conclusion is entirely analogous to the conclusion reached by sampling statisticians regarding survey data that is contaminated by errors of measurement (or response). For the present problem, just as for the response error problem, the appropriate measure of uncertainty is the total variance, including both the sampling variance and the ξ-variance.

Fortunately, the difference between adopting the model variance or the total variance as a measure of uncertainty is not important from a practical point of view. Either way, the estimator of variance and corresponding confidence intervals are identical.

4. THE ADJUSTMENT PROBLEM

We now consider briefly the problem of estimating the true population size of a small geographic area (or other small domain) within the population U. Such an area is generally characterized by one of two problems: 1) either there is no sample from List B in the area or 2) there is a sample but it is of small size. Either way there are difficulties in applying the estimation methodologies discussed in Sections 2 and 3 within the small area. In the first case the estimators are undefined, and in the second case there are problems with both high variance and bias due to the ratio form of the estimators. The general issue of estimating the true population size of small geographic areas is often called census adjustment. In this section we expose a clear connection between the methodology known as synthetic estimation and census adjustment under model M_{th}.

Let D denote a particular area or domain of interest. We wish to estimate x_D , the true population size of D . The only data available to us is the census information, the survey information, and the results of matching the survey to the census.

The natural estimator of x_D is

$$\hat{x}_D = \sum_{i\varepsilon D} x_{i1+}/p_{i1+}$$

where

$$x_{i1+} = 1, \text{ if the i-th individual is in the census}$$

$$= 0, \text{ otherwise,}$$

$$p_{i1+} = \text{probability that the i-th individual is listed in the census,}$$

and the summation is over all individuals in D. This estimator is ξ-unbiased for x_D with variance

$$V_\xi\{\hat{x}_D\} = \sum_{i \varepsilon D} \frac{p_{i2+}}{p_{i1+}} .$$

Unfortunately, the estimator is not workable in practice because the individual probabilities p_{i1+} are unknown.

A workable estimator is obtained by replacing the unknown capture probabilities by sample-based estimators, e.g.,

$$\tilde{x}_D = \sum_{i \varepsilon D} x_{i1+}/\tilde{p}_{i1+} , \qquad (4.1)$$

where $\tilde{p}_{i1+}$ is an estimator of p_{i1+} based upon the available data. Different estimators $\tilde{p}_{i1+}$ are available depending upon which coverage error model is applicable , and even within a given model there are alternatives.

To illustrate these ideas we consider model M_t once again. For this model the reader will recall that $p_{i1+} = p_{1+}$ for all i = 1, ..., N, and that the maximum likelihood estimator given complete enumeration of List B is $\hat{p}_{1+} = x_{11}/x_{+1}$. Replacing x_{11} and x_{+1} by their p-unbiased estimators gives $\tilde{p}_{1+} = \hat{x}_{11}/\hat{x}_{+1}$, which may be employed in (4.1). The first-order Taylor series variance is then given by

$$V\{\tilde{x}_D\} = \underset{\sim}{Q}' \underset{\sim}{\Omega} \underset{\sim}{Q} ,$$

where

$$\underset{\sim}{Q} = (p_{1+}^{-1} , x_D N^{-1} p_{+1}^{-1} , -x_D N^{-1} p_{1+}^{-1} p_{+1}^{-1})'$$

and $\underset{\sim}{\Omega}$ denotes the covariance matrix of $(\sum_{i \varepsilon D} x_{i+1}, \hat{x}_{+1}, \hat{x}_{11})$. All three components of this vector are subject to model variability, and $\hat{x}_{+1}$ and $\hat{x}_{11}$ are also subject to sampling variability. A consistent estimator of the variance may be obtained by replacing both $\underset{\sim}{Q}$ and $\underset{\sim}{\Omega}$ by sample based estimators.

Alternatively, one may define the estimator $\tilde{p}_{1+}$ only in terms of survey data collected within area D, i.e.,

$$\tilde{p}_{1+} = \hat{x}_{D11} / \hat{x}_{D+1}$$

where $\hat{x}_{D11}$ and $\hat{x}_{D+1}$ are design-unbiased estimators of $\sum_{i \varepsilon D} x_{i11}$ and $\sum_{i \varepsilon D} x_{i+1}$, respectively. The disadvantage of this estimator is that it may be based upon extremely small sample sizes, thus leading to problems of large variance and ratio bias.The advantage is that when model M_t fails to some degree, this estimator may be subject to less specification bias than the estimator based upon the full sample from List B.

A third version of the estimator $\tilde{p}_{1+}$ is

$$\tilde{p}_{1+} = \omega \hat{x}_{11}/\hat{x}_{+1} + (1-\omega)\hat{x}_{D11}/\hat{x}_{D+1} ,$$

where $\omega \varepsilon$ (0,1). One can attempt to deal with the tradeoff between variance and bias by varying the parameter ω. See Dempster and Tomberlin (1980) for discussion of related issues in a Bayesian framework.

In U.S. applications of census adjustment, a version of model M_{th} is often thought to be appropriate, where model M_t holds within strata, such as by age, race, and sex. In this case a model M_t estimator is prepared within each stratum, and then aggregated across strata to estimate the total population size N. The adjusted census count $\tilde{x}_D$ for this problem is of the form (4.1), where $\tilde{p}_{i1+}$ is a model M_t estimator (possibly one of the three discussed above) constructed for the stratum of which the i-th individual is a member. The reader will recognize the result as a synthetic estimator.

5. DISCUSSION

In this article we have presented alternative models for coverage error in censuses or sample surveys. For one of the models, M_t, we developed in some detail a theory of inference for the unknown population size N. We showed that the estimators of N are subject to two sources of variability: 1) sampling variability and 2) model variability. The estimators were shown to be asymptotically normally distributed. Finally, we developed clear connections between the coverage error model and the usual survey response error model and between the ideas of synthetic estimation and census adjustment given the coverage error model.

One of the models presented, M_{th}, specifies both a time effect and heterogeniety in the capture probabilities. This is the model that has typically been applied in the historical studies of coverage error. The model M_t theory may be applied to this problem provided the population can be poststratified to eliminate the heterogeniety.

Throughout the article we have necessarily imposed some rather restrictive assumptions in order to facilitate the mathematical developments. The most important challenges for future research involve relaxing the assumptions to the point where the methods are useful in applied settings. Three of the main difficulties involve the matching assumption (iv), the spurious events assumption (v), and the nonresponse assumption (vi). None of these assumptions will obtain entirely in practice. Some progress has been made on (v) and the new methods implemented in a recent study in the U.S. See Cowan and Bettin (1982). Also, we have recently expanded the coverage error models to accommodate a nonresponse mechanism and will report preliminary developments in a future paper. Matching error, however, remains a difficult, unresolved practical problem that can have significant effect on inferences about N.

Finally, we are unaware of any previous work in the coverage error literature that introduces the equivalent of a correlated component into the model. This would also seem an important area for further development.

BIBLIOGRAPHY

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press: Cambridge.

Burnham, K. P. and Overton, W. S. (1978), "Estimation of the Size of a Closed Population when Capture Probabilities Vary Among Animals," *Biometrika* 65, 625-633.

Cowan, C. and Bettin, P. (1982), "Estimates and Missing Data Problems in the Post Enumeration Program," unpublished manuscript, U.S. Bureau of the Census, Washington, D. C.

Dempster, A. P. and Tomberlin, T. J. (1980), "The Analysis of Census Undercount From a Postenumeration Survey," Proceedings of the 1980 Conference on Census Undercount, U.S. Bureau of the Census, Washington, D. C.

Hájek ,J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population,"*Pub. Math. Inst. Hung. Acad. Sci.* 5, 361-374.

Hansen, M. H., Hurwitz, W. M. and Bershad, M. A. (1961), "Measurement Errors in Censuses and Surveys," *Bull. of the Int. Statist. Inst.* 38, No. 2, 359-374.

Isaki, C.T. and Fuller, W.A. (1982), "Survey Design Under the Regression Superpopulation Model," *JASA* 77, 89-96.

Lincoln, F.C. (1930), "Calculating Waterfowl Abundance on the Basis of Banding Returns," Circ. U.S. Dept. of Agric. No. 118, 1-4.

Marks, E.S., Seltzer, W. and Krotki, K.J. (1974). *Population Growth Estimation*, The Population Council: New York.

Otis, D. L., Burham, K.P., White, G.C. and Anderson, D.R. (1978), "Statistical Inference for Capture Data from Closed Populations," *Wildl. Monogr. No. 62*, The Wildlife Society, Allen Press: Lawrence, Kansas.

Peterson, C.G.J. (1896), "The Yearly Immigration of Young Plaice into the Limfjord from the German Sea," *Rep. Danish Biol. Stat.* 6, 1-48.

Pollock, K.H. (1974), "The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments," unpublished Ph.D. dissertation, Cornell Univ., Ithaca, N.Y.

Schnabel, Z.E. (1938), "The Estimation of the Total Fish Population of a Lake," *American Math. Month.* 45, 348-352.

Seber, G.A.F. (1973). *The Estimation of Animal Abundance and Related Parameters*, Griffin: London.

Sekar, C.C. and Deming, W.E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *JASA* 44, 101-115.

Seltzer, W. and Adlakha, A. (1974), "On the Effect of Errors in the Application of the Chandrasekaran-Deming Techniques," Laboratory for Population Statistics Reprint Series No. 14, University of North Carolina, Chapel Hill.

SUMMARY

Alternative models are presented for representing coverage error in surveys and censuses of human populations. The models are related to the capture-recapture models used in wildlife applications and to the dual-system models employed in the vital events literature. Estimation methodologies are discussed for one of the coverage error models. The theoretical foundations of the methodology are developed and distinctions are made between two kinds of error: 1) sampling error and 2) error associated with the model. Clear connections are made between the coverage error model and models for measurement or response error.

Finally, the problem of adjusting census and survey data for coverage error is discussed. A direct connection is exposed between the synthetic estimation methodology and the appropriate methodology given the coverage error model. Properties of the adjustments are discussed.

Résumé

On présente des modèles alternatifs pour représenter les erreurs de couverture dans les enquêtes par sondage et dans le recensement des populations. Ces modèles ont rapport aux modèles de prise - reprise qui sont employés dans leur application aux animaux sauvages et aussi aux modèles du système double collecte qui sont employés dans les enregistrements de l'état civil. On y discute les méthodologies d'estimation pour un des modèles d'erreur de couverture. On développe les fondements théoriques de la méthodologie et on établit des distinctions entre deux types d'erreurs: l'erreur d'échantillonnage et l'erreur qui s'associe avec le modèle. Des rapports évidents s'établissent entre le modèle d'erreur de couverture et les modèles d'erreur empirique.

Enfin on discute le problème d'ajuster les données à partir des recensements et des enquêtes par sondage pour l'erreur de couverture. Un rapport direct est établi entre la méthodologie d'estimation synthétique et la méthodologie appropriée d'ajustement etant donné le modèle d'erreur de couverture. On y discute les caractéristiques des ajustements.