EMAP Status Estimation: Statistical Procedures and Algorithms

V.M. LESSER AND W.S. OVERTON

Department of Statistics, Oregon State University,

Corvallis, Oregon

Project Officer

Anthony R. Olsen

U.S. Environmental Protection Agency

Environmental Research Laboratory

200 SW 35th Street, Corvallis, Oregon

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency under cooperative agreement CR-816721 with Oregon State University at Corvallis. It has been subject to the agency's peer and administrative review. It has been approved for publication as an EPA document.

EMAP Status Estimation: Statistical Procedures and Algorithms

V.M. LESSER AND W.S. OVERTON

Department of Statistics, Oregon State University,

Corvallis, Oregon

Project Officer

Anthony R. Olsen

U.S. Environmental Protection Agency

Environmental Research Laboratory

200 SW 35th Street, Corvallis, Oregon

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency under cooperative agreement CR-816721 with Oregon State University at Corvallis. It has been subject to the agency's peer and administrative review. It has been approved for publication as an EPA document

CONTENTS

SECTION 1

INTRODUCTION

The Environmental Protection Agency (EPA) has initiated a program to monitor ecological status and trends and to establish baseline environmental conditions against which future changes can be monitored (Messer et al., 1991). The objective of this environmental program, referred to as EMAP (Environmental Monitoring and Assessment Program), is to assess the status of a number of different ecological resources, including surface waters, wetlands, Great Lakes, near-coastal waters, forests, arid lands, and agroecosystems.

A design plan and a number of support documents have been prepared to guide design development for EMAP (Overton et al., 1990, Overton and Stehman, 1990; Stehman and Overton, in press; Stevens, in press). The statistical methods outlined in earlier documents, such as those analyzing the EPA National Surface Water Surveys, are also relevant to EMAP (Linthurst et al., 1986, Landers et al., 1987; Messer et al., 1986).

This report presents statistical procedures collected from other EMAP documents, as well as from Oregon State University technical reports describing data analyses for other EPA designs. By integrating this information, this manual and the EMAP design report will serve as reference sources for statisticians who implement an ecological monitoring program based on the EMAP design framework. Spatial and temporal analyses of EMAP data are not covered in this version of the report. A brief discussion of the four-point moving average, which combines data over the interpenetrating sample, is presented in Overton et al. (1990; Section 4.3.7). Algorithms listed in this report cover most design options discussed in the EMAP design report. It is expected that any further realizations of the EMAP design will also include documentation of corresponding variance estimators.

1.1 Overall Design

The EMAP design is a probability sample of resource units or sites that is based on two tiers of sampling. The first tier (Tier 1) primarily supports landuse characterization and description of population structure, while the second tier (Tier 2) supports status assessment by the indicators. The second tier sample is a probability subsample of the first tier sample; such a sample is referred to as a double sample. Across the ecological resource groups, it is expected that discrete, continuous, and extensive populations will be monitored. The statistical methods outlined in this report address these different population types at both sampling tiers. A description of the sampling design is presented in Overton et al. (1990)

1.2 Resources

EMAP is designed to provide the capability of sampling any ecological resource. To achieve this objective, explicit design plans must be specific for a particular resource and all resources to be characterized must be identified Currently, the resources to be sampled within EMAP include surface waters, wetlands, forests, agroecosystems, arid lands, Great Lakes, and near-coastal wetlands. These resources are further divided by major classes to represent the specific 'resource' that will be addressed by the sampling effort. For example, surface waters can be partitioned into classes such as very small lakes, intermediate-sized lakes, very large lakes, very small streams, intermediate-sized streams, rivers, and very large rivers. Because each class potentially generates different sampling issues, each would be considered a different entity. The design structure meets this condition by identifying each such class as a resource, thereby resulting in 6 to 12 surface water resources. Each major resource group may also have as many divisions.

Most resources will be sampled via the basic EMAP grid and associated structures. However, other resources, such as very large lakes and very large rivers, represent unique

ecological entities and cannot be treated as members of a population of entities to be described via a sample of the set. For example, Lake Superior and the Mississippi River are unique, although the tributaries of the Mississippi might be treated as members of a wider class of tributaries.

Resources sampled by the EMAP grid will be associated with an explicit domain in space, within which the resource is confined. This domain should be established early in the design process. Within the defined domain, it is not expected that the resource will occupy all space or that no other resource will occur. Domains of different resources will overlap, but the domain of a particular resource is an important parameter of its design. For purposes of nomenclature, the resource domain is a region containing the resources. The resource universe is either a point set within one point for each resource unit (discrete resource) or the continuous space actually occupied by the resource (for extensive resources) A resource class will be a subset of its universe. Such a class may or may not be treated as a sampling stratum and may or may not have an identified subdomain.

1.3 Response Variables of Interest

The term response variable is used generally for the measured characteristic of interest in the sample survey. In EMAP, a special class of response variables is referred to as indicators, such as indicators of ecological status (Hunsaker and Carpenter, 1990). These indicators are the environmental and ecological variables measured in the field on resource units or at resource sites; they may be measured directly or modified via formulae or analytic protocols.

The term, indicators, should not be applied to the structural variables defined at Tier 1. The Tier 1 variables are used to estimate population structure and to partition the Tier 1 sample into the necessary structural parts for Tier 2. Then the indicator variables are determined on the Tier 2 sample. When the Tier 2 sample includes the entire Tier 1

sample, it is still appropriate to make the distinction between indicator and structural variables, both of which are response variables. Because of this distinction, it is sometimes appropriate to distinguish Tier 1 and Tier 2 in terms of the variables, rather than strictly in terms of a subsample.

1.4 Statistical Methods

The primary statistic used to summarize population characteristics is the estimated distribution function. This distribution function estimates the number or proportion of units or sites for which the value of the indicator is equal to or less than y. For discrete resources, the estimated distribution function for numbers is designated as $\hat{N}(y)$, while the estimated distribution function for the proportion of numbers is designated as $\hat{F}(y)$. The estimated distribution function of size-weighted totals in discrete resources is designated as $\hat{Z}(y)$, while a size-weighted proportion is designated as $\hat{G}(y)$. Examples of size-weights are lake area, stream miles, and stream discharge. There are no distribution functions comparable to $\hat{Z}(y)$ and $\hat{G}(y)$ in the continuous and extensive populations because there are no objects in these resources. Therefore, there are no object sizes to use as weights. In extensive resources, the estimated distribution function representing actual areal extent for which the value of the indicator is equal to or less than y is designated as $\hat{A}(y)$, while the proportion of areal extent is designated as $\hat{F}(y)$. Thus $\hat{A}(y)$ is analogous to $\hat{N}(y)$; A is the size of an extensive resource and N is the size of a finite resource.

A number of estimates of interest, which can be obtained from the distribution function, have been used quite successfully in the National Surface Water Surveys (NSWS) (Linthurst et al., 1986; Landers et al., 1987; Kaufmann et al., 1988). For example, any quantile, including the median of the distribution, can be interpolated easily from the distribution function. In addition, the distribution function can be supplemented with tables of means, quantiles, or any other statistics of particular interest, providing greater accuracy than can be obtained from the plotted distribution function.

The basic formula for estimated distribution functions is

$$\widehat{F}_a(y) = \frac{\left(\sum_{S_{ay}} w_i\right)}{\left(\sum_{S_a} w_i\right)} = \frac{\widehat{N}_a(y)}{\widehat{N}_a} \quad , \tag{1}$$

where S is the sample of units representing the universe ($\mathcal{U}$) and the variable y represents any response variable. The subscript a denotes a subpopulation of interest; S_a is the portion of the sample in subpopulation a, and S_{ay} is the portion of the sample in subpopulation a having values $\leq y$. We associate the inclusion probability, π_i, with each i^{th} sampling unit. Each sampling unit is a representation of a subset of the universe, and the weight ($w_i = \frac{1}{\pi_i}$) accounts for the size of the subset. $\widehat{N}_a$ denotes the estimated population size for the subpopulation a. $\widehat{F}_a(y)$ is calculated for each value of y appearing in the sample.

As given, $\widehat{F}_a(y)$ is a step function not suitable to general EMAP needs; a smoothed version is desirable. Thus, we propose the following method. In this method, $\widehat{F}_a(y)$ is replaced by $\frac{\widehat{F}_a(y) + \widehat{F}_a(y')}{2}$, where y' is the next lesser value to y. For the minimum values of y, $\widehat{F}_a(y)$ is replaced by $\frac{\widehat{F}_a(y)}{2}$. A linear interpolation is then made between these points to generate the plot or to determine quantiles. For each of the distribution function algorithms provided in this report, two successive values are averaged in this manner and used to develop an interpolated distribution function. Confidence bounds are constructed on $\widehat{F}_a(y)$ and then averaged and interpolated in the same manner. We rest justification for this procedure on the interpretation of the initial and final values of the resulting distribution function. The initial value is our best estimate of the proportion of the population below the minimum observed value, and similarly, one minus the last point is our best estimate of the proportion of the population above the maximum observation.

Computation of the distribution function and the associated confidence bounds differs slightly for specific resource groups, reflecting differences in detail of the sampling design.

For example, in some cases simplifications of the computing algorithms result from equal probability designs. Some algorithms are presented in this report to accommodate the range of conditions and objectives anticipated across the resource groups. These algorithms have been previously discussed in greater detail in other documents; references are given for those requiring a more in-depth approach. Table 1 provides an outline of the distribution functions, which are presented in greater detail throughout this report. Table 2 provides a table of notation used throughout this document.

1.5 Use of this Manual

The body of this manual has been separated into two sections. Section 2 includes the general theoretical development upon which the algorithms are based. Formulae for discrete, continuous, and extensive resources are presented, the mathematical notation is introduced and defined; and both design-based and model-based approaches for computing the distribution functions are discussed. Other issues relevant to the analysis of EMAP data, such as handling of missing data, are also discussed in this section.

Section 3 includes the algorithms used to produce the distribution functions, the conditions that provide for the application of the algorithms, and the rationales that support the choice and derivation of the algorithms presented. References will be made to the general formulae (discussed in Section 2) used to develop these algorithms.

The following list outlines a step-by-step sequence for obtaining the distribution functions:

1. Determine whether the data represent a discrete or extensive resource.
2. Determine the type of distribution function to compute. For example, for discrete resources, the distribution of numbers and/or proportions of numbers will be of interest.
3. Determine whether the sampling units were collected with equal or variable probability of selection. The inclusion probabilities, π_{1_i} and $\pi_{2.1_{ij}}$, discussed in Section 2.1.1, are to be a permanent part of a datum record, as are the identification code of the sampling unit and the variable of interest. In some cases, it is also necessary to identify the grid point, which can be included as part of the identification code.
4. Determine whether the size of the subpopulation of interest is known or unknown

The subpopulation is the group of population units about which one wishes to draw inference.

5. Using the conditions from steps 1-4, refer to the example of that specific algorithm in Section 3.
6. Optional, but suggested: Refer to the formulae referred to in Section 2 for a description of the formulae and for clarification of any notational problems.
7. Optional: An algorithm to obtain specific quantiles is presented in Section 3.

This manual is expected to be updated as research continues in the development of statistical procedures for EMAP, as EMAP adapts to changing concerns and orientation, and as EMAP makes and accumulates more in-depth frame materials. For example, efforts to date have been focused on design-based approaches to confidence bound estimation, therefore this version reflects a fairly in-depth approach to design-based estimation over all resources Further discussion of model-based approaches currently under development are expected in future versions of this manual.

SECTION 2

GENERAL THEORETICAL DEVELOPMENT

Two approaches are commonly used to draw inferences from a sample survey relative to a real population. In the design-based approach, described in Section 2.1, the probabilities of selection are accounted for by the estimators and the properties of inferences are derived from the design and analytical protocol. In contrast, the model-based approach (Section 2.2) assumes a model and requires knowledge of auxiliary variables for inference. Properties of model-based inference are derived from the assumed model and analytical protocol A model-based estimator takes into account only model features, while a model-assisted estimator takes into account both model and design features. For a discussion of the relationship between these two approaches and the way they are used together, refer to Hansen et al. (1983). The paper by Smith (1976) also provides useful insight.

2.1 Design-Based Estimation Methods

2.1.1 Discrete Resources

A population of natural units readily identified as objects is defined as a discrete resource. For example, lakes, stream segments, farms, and prairie potholes are all considered discrete resources. Populations of a large number of discrete resource units that can be described by a sample are considered for EMAP representation. It is suggested, for example, that lakes less than 2,000 hectares be characterized as populations of discrete resources. Distribution functions of the numbers of units or proportions of these numbers may be of interest. On the other hand, very large lakes are unique, and less usefully characterized as members of populations of lakes.

2.1.1.1 General Estimator and Approximations of Design-Based Formulae

Because the EMAP design is based on a probability sample, design-based estimators, which account for this structure, are applicable. The Horvitz-Thompson theorem (Horvitz and Thompson, 1952) provides general estimators of the population attributes for general probability samples and for estimators of variance of these estimators (Overton and Stehman, 1993a; Overton et al., 1990).

In Horvitz-Thompson estimation, the probability of inclusion, π_i, is associated with the i^{th} sampling unit. Each sampling unit is a representation of a subset of the universe, and the weight ($w_i = \frac{1}{\pi_i}$) accounts for the size of the subset Therefore, estimates of population parameters, such as totals or means, simply sum the variables collected over the sampling units, expanding them by the sampling weights. The Horvitz-Thompson estimator proposed for EMAP is unbiased for the population (and subpopulation) totals and means, if $\pi_i > 0$ for all units in the population.

The general form of the Horvitz-Thompson estimator is

$$\hat{T}_y = \sum_S y_i w_i \quad , \tag{2}$$

where S is the sample of units representing the universe ($\mathcal{U}$), w_i is the weight, and the variable y represents any response variable. The total of y on the universe is defined as $T_y = \sum_{\mathcal{U}} y$ and is generally referred to as the population total. This estimator (Equation 2) yields estimates of many parameters simply by the definition of y. For example, if $y_i=1$ for all units in the population, then $T_y = N$, the population size; it follows that $\hat{N} = \sum_S w_i$.

Suppose further that we are interested in a subpopulation, a. The portion of the sample, S_a, that came from this subpopulation is also a probability sample from this subpopulation. To obtain parameter estimates for a subpopulation, Equation 2 is simply summed over the subpopulation sample,

$$\hat{T}_{ya} = \sum_{S_a} y_i w_i \tag{3}$$

The Horvitz-Thompson theorem also provides for an unbiased estimator of the variance of these estimators under the condition that $\pi_{ij}>0$ for all pairs of units in the population. The quantity π_{ij} is the probability that unit i and unit j are simultaneously included in the sample. The estimator of the variance is designated in lower case as var, and w_{ij} is the inverse of the pairwise inclusion probability. The variance of $\hat{T}_{ya}$ or $\hat{N}_a$ is obtained by the choice of y_i:

$$\text{var}(\hat{T}_{ya}) = \left\{ \sum_{S_a} y_i^2 w_i (w_i - 1) + \sum_{S_a} \sum_{\substack{S_a \\ j \neq i}} y_i y_j (w_i w_j - w_{ij}) \right\} , \tag{4a}$$

$$\text{var}(\hat{N}_a) = \left\{ \sum_{S_a} w_i (w_i - 1) + \sum_{S_a} \sum_{\substack{S_a \\ j \neq i}} (w_i w_j - w_{ij}) \right\} . \tag{4b}$$

This presentation shows that the form of the variance estimator does not change when estimating variance for the estimator based on a full sample or a subset of the sample. The subsetting device, with summation over the appropriate subset of the sample, will always represent the appropriate estimator. The principal reason for using the Horvitz-Thompson form (Equation 4) is its subsetting capability; the commonly used form for the Yates-Grundy variance estimator does not permit the convenience of subsetting.

The EMAP design is based on a systematic sampling scheme. The Horvitz-Thompson theorem does not provide a design-unbiased estimator of variance based on this design, because some pairwise inclusion probabilities are zero. The following sections include a discussion of assumptions and approximations applied to Equation 4 in order to apply this variance estimator in EMAP.

Systematic sampling

Because EMAP units are selected by a systematic sampling design, many of the pairwise inclusion probabilities ($\pi_{i,j}$) equal zero and an unbiased variance estimator is not available. However, it has been established that in many cases the variance of a systematic design can be satisfactorily approximated by the variance that applies to a sample taken on a randomly ordered list (cf., Wolter, 1985) A common systematic sample selected on a randomly ordered list is a simple random sample. Therefore, a simple random sample is an approximate model for an equiprobable systematic sample. The randomized model proposed here provides approximate variance estimation for a systematic variable probability design.

A modification of the randomized sampling model provides only for 'local' randomization of the position of the population units, rather than global randomization. Good behavior of the variance estimator results from this assumption (Overton and Stehman, 1993b). As a consequence, we can justify use of the suggested pairwise inclusion probability with less restriction as compared with the global randomization assumption We will refer to the local randomization model as the weak randomization assumption.

The Horvitz-Thompson estimator of variance, Equation 4, is thus proposed for EMAP indicators under the weak randomization assumption. The simulation studies conducted on the behavior of this estimator suggested that this assumption was adequate in most situations expected for EMAP (Overton, 1987a; Overton and Stehman, 1987; Overton and Stehman, 1992; Stehman and Overton, in press). In a few situations the estimator overestimated the true variance, thus providing for a conservative estimate of variance. In certain circumstances, as discussed in Section 2.1.2.3, it is appropriate to modify the estimation methods to account for the spatial patterns.

Pairwise inclusion probability

Approximations for the pairwise inclusion probability under the randomized model have been proposed in the literature (Hartley and Rao, 1962). A major disadvantage with these approximations, as discussed by Stehman and Overton (1989), is the requirement that all inclusion probabilities for the population must be known. For large populations such as those studied in EMAP, it is practically impossible to obtain inclusion probabilities for all units in the populations. Another approximation for this pairwise inclusion probability requires that the inclusion probabilities be known only on the sample (Overton, 1987b). The formula for the inverse of this pairwise inclusion probability is

$$w_{i,j} = \frac{2nw_i w_j - w_i - w_j}{2(n-1)} \tag{5}$$

where n is sample size.

Investigation of this approximation indicates that it performs at least as well as other commonly recommended approximations (Overton and Stehman, 1992, Overton, 1987a). Therefore, this pairwise inclusion probability will be used in the approximation of the variance estimator for the population parameter estimates collected in EMAP, for those circumstances in which the randomization assumption is justified.

This variance estimator (Equation 4) accommodates variable probability of selection, but it is also appropriate for equal probability designs. The approximation for the pairwise weight given in Equation 5 is also appropriate for randomized equal probability designs. As a consequence, Equation 4 with 5 is valid for either equal or variable probability selection, under the weak assumption of a randomized model, as discussed above under systematic sampling.

When the randomization model is not acceptable, alternative variance estimators, based on the mean square successive difference, have been developed for use with an equal

probability systematic design and regular spacing (Overton and Stehman, 1993a). The conditions and assessment of these and other variance estimators are presently under investigation; subsequent versions of this document will discuss these alternate methods. Extension must be made to account for irregular spacing. It should also be noted that in some circumstances the methods of spatial statistics may provide adequate assessment of variance.

The confidence bounds obtained using the Horvitz-Thompson variance estimator (Equation 4) are based on normal approximation. This approximation may be inadequate for estimating confidence bounds at the tails of the distribution, even for moderate sample sizes. In the special case of equal probability of selection and the randomization assumption, confidence bounds can be obtained by exact methods (see Section 3). However, exact methods may also yield inadequate confidence bounds at the tails of the distribution (also discussed in Section 3).

2.1.1.2 Tiered Structure

The following description of the tiered structure was summarized in the EMAP design report (Overton et al., 1990).

The Tier 1 sample

The EMAP sample design partitions the area of the United States into hexagons, each comprising approximately 635 square kilometers (Overton et al., 1990), and selects a point at an identical position in each hexagon; selection of this one position is random (equiprobable) over the hexagon. This method results in a triangular grid of equally spaced points. An areal sample of a 40-km^2 hexagon (40-hex) is imposed on each point, with the sampled hexagonal area containing $\frac{1}{16}$ of the total area of the larger hexagon. This fraction, $\frac{1}{16}$, therefore represents a constant inclusion probability, π, and 16 represents a constant

weight, w, to be applied to each fixed-size areal sample. Because other enhancements of the grid are expected, possibly with different sized areal samples, the following formulae will incorporate general notation.

No detailed characterization of indicators is collected at Tier 1, so no distribution functions will be computed based on the Tier 1 data. It is of interest, however, to estimate the total number of discrete resource units in specific populations at Tier 1. This estimation is possible for any resource class for which units can be uniquely located by a position point. The following formula can be used to estimate the total number of units for a particular resource (r) at Tier 1:

$$\widehat{N}_r = w \sum_{\mathfrak{D}_r} n_{ri} \quad , \tag{6}$$

where $\mathfrak{D}_r$ is the domain for resource r. A domain of a resource is a feature of the spatial frame that delineates the entire area within which a sample might encounter the resource (Section 1.1). In these formulae, the quantity n_{ri} represents the number of units for the particular resource at grid point i. The variance can be estimated using Equation 4b, as follows:

$$\text{var}(\widehat{N}_r) = \frac{n_r w(w-1)}{n_r - 1} \left\{ \sum_{\mathfrak{D}_r} n_{ri}^2 - \left(\sum_{\mathfrak{D}_r} n_{ri} \right)^2 / n_r \right\} \quad , \tag{7}$$

where n_r is the number of grid points for which the areal sample hexagon includes part of the domain of the resource. It is worth noting again that the estimates of variance are often expected to slightly overestimate variance if the systematic design results in greater precision than would a randomized design, thus providing for a conservative estimate of variance.

The reduced Tier 1 sample

In preparation for selecting the Tier 2 sample, resource classes are identified. Some of these classes will be treated as sampling strata, and hence be designated as 'resources'. The Tier 1 sample for such a 'resource' is reduced so that it contains only one unit at each 40-hex at which that resource is represented. This condition effectively changes the sample from a set of systematic areal samples to a spatially well-distributed subset of units from the population of units for the particular resource.

A consequence of this sample reduction step is the introduction of variable inclusion probabilities in the Tier 1 sample, reflecting the scheme used to reduce n_{ri} to 1. For example, if a random sample of size 1 is selected from the n_{ri} units of hexagon i, then the selected unit will have $\pi_{1ri} = \frac{1}{wn_{ri}}$. A consequence of this is that $\hat{N}_r = \sum_{S_{1r}} w_{1ri} = w\sum_{S_{1r}} n_{ri}$, where S_{1r} is now the sample of points for which $n_{ri}>0$, at each of these points, the sample now consists of one unit of resource r. Because this estimate, $\hat{N}_r$, is identical to the original Tier 1 estimate, it has the same variance. This sample, S_{1r}, is then subsampled to generate the Tier 2 sample, S_{2r}. Again, note that it is now a resource-specific sample of units, not the original areal sample.

The Tier 2 sample

The Tier 2 sample, S_2, is a probability subsample, a double sample, of the Tier 1 sample of resource units. At this tier, a specific resource has been identified and the reader should remember that subsequent equations are for specific resources. The reader should also be aware that the subscript i will now index a resource unit, not the grid point. All Tier 2 samples for discrete resources consist of individual units from the universe of discrete resource units.

With these changes, the estimator presented in Equation 2 is appropriate for the sample collected at the second tier. The indicator values are summed over the samples

surveyed at the second tier by the assigned weights. The inclusion probabilities account for the probability structure of this double sample. Overton et al. (1990) identified the probability of the inclusion of the i^{th} unit in the Tier 2 sample as the product of the Tier 1 inclusion probability and the conditional Tier 2 inclusion probability. The conditional Tier 2 inclusion probability is defined as the probability of inclusion at Tier 2, given that the unit was included at Tier 1. This product is still conditional on the Tier 1 sample and leads to conditional Horvitz-Thompson estimation.

In subsequent equations, the subscripts 1 and 2 represent the first and second tiers, respectively. The weighting factor for unit i at Tier 2 is defined as

$$w_{2ri} = w_{1ri} w_{2.1ri} \quad , \tag{8}$$

where w_{1ri} is the weighting factor for the i^{th} unit in the Tier 1 reduced sample and $w_{2.1ri}$ is the inverse of the conditional Tier 2 inclusion probability for resource r.

Selection of the Tier 2 sample from the reduced Tier 1 sample and calculation of the conditional Tier 2 inclusion probabilities are discussed in Section 4.0 of the EMAP design report (Overton et al., 1990) This procedure generates a list in a specific order, based on spatial clusters. Clusters of 40-hexes are arbitrarily constructed with uniform size of the initial Tier 1 sample of the specific resource. The reduced Tier 1 sample is sorted at random within clusters, and then the clusters are arranged in an arbitrary order. A subsample of fixed size, n_{2r}, is selected from S_{1r} by ordered variable probability systematic sampling from this list. The purpose of this elaborate procedure is to generate a spatially well-distributed Tier 2 sample.

The Tier 2 conditional inclusion probabilities are proportional to the weights at Tier 1:

$$\pi_{2.1ri} = \frac{n_{2r} w_{1ri}}{\sum_{S_{1r}} w_{1ri}} = \frac{n_{2r} w_{1ri}}{\hat{N}_r} \quad . \tag{9}$$

where $\hat{N}_r$ was defined for a specific resource in Equation 6. However, for some units $w_{1ri} > \frac{\hat{N}_r}{n_{2r}}$. To obtain conditional inclusion probabilities ≤ 1, these units are placed into an artificial 'certainty' stratum, all having $\pi_{2.1ri}=1$. This step takes place prior to the cluster formation. For the remaining units, the selection protocol and achieved probabilities are modified to adjust for the number of units having probability 1. These remaining units now have conditional inclusion probabilities:

$$\pi_{2.1ri} = \frac{n^*_{2r} w_{1ri}}{\sum_{S^*_{1r}} w_{1ri}} \quad , \tag{10}$$

where n^*_{2r} equals n_{2r} less the number of units entering S_{2r} with probability 1, and S^*_{1r} equals S_{1r} less these same units that were included with probability 1.

Note that this selection protocol is designed to create Tier 2 inclusion probabilities as nearly equal as possible:

$$\pi_{2ri} = \begin{cases} \pi_{1ri}, & \text{if } i \text{ is in the artificial stratum with } \pi_{2.1ri}=1 \\ \dfrac{n^*_{2r}}{\sum_{S^*_{1r}} w_{1ri}}, & \text{otherwise,} \end{cases} \tag{11}$$

and if no units are in the artificial stratum,

$$\pi_{2ri} = \frac{n_{2r}}{\hat{N}_r} \quad , \tag{12}$$

where $\hat{N}_r$ is the Tier 1 estimate of the total number of population units in resource r. For generality, we will retain the variable probability notation, but ideally the sample will now be equal probability. If there is great deviation from equiprobability, then consideration should be given to enhancement of the grid, perhaps by reducing the size of the Tier 1 areal sample, in order to better achieve the goal of equiprobability.

The variance estimator presented in Equation 4 is also appropriate for estimating variance from the Tier 2 sample, using inclusion probabilities defined above for Tier 2.

When no units enter with $\pi_{2.1r_i}=1$, then

$$w_{2r_ij} = \frac{2n_{2r}w_{2r_i}w_{2r_j} - w_{2r_i} - w_{2r_j}}{2(n_{2r}-1)} \tag{13}$$

However, if unit i enters with $\pi_{2.1r_i}=1$, then,

$$w_{2r_ij} = \left\{ \frac{2n_{1r}w_{1r_i}w_{1r_j} - w_{1r_i} - w_{1r_j}}{2(n_{1r}-1)} \right\} w_{2.1r_j} \tag{14}$$

Because the term in the bracket equals w_{1r_ij}, Equation 14 simplifies to $w_{2r_ij} = w_{1r_ij}w_{2.1r_j}$.

Special case: The Tier 2 point sample of lakes

We assume a stratified design with equiprobable selection within strata. If a quasi-stratified design is used instead, appropriate analysis can condition on the realized sample sizes in the classes and use post-stratification.

Special case: The Tier 2 point sample of streams

A point sample of streams at Tier 2, rather than a sample of stream reaches, has been proposed. With a few simple changes, that point sample will be a rigorous equiprobable point sample of the stream population with a very simple estimation algorithm. A probability sample of stream reaches, on which the sample points are represented and from which other estimates of population structure can be obtained, will also be provided. The protocol provided will apply to the sample of stream reaches and the point sample design proposed to us.

We assume a stratified design with equiprobable selection within strata. If a quasi-stratified design is used instead (as has been proposed), appropriate analysis can condition on the realized sample sizes in the classes and use post-stratification.

S_{1r} is the Tier 1 collection of reaches in a resource stratum identified via the 40-hexes. S_{2r} is generated by selecting n^*_{2r} points from this set using the frame representation of stream length This process results in (1) the selection of n^*_{2r} frame reaches with probability proportional to frame length, and (2) the random selection of 0, 1, 2, ... points in each selected reach with inclusion density, given reach selection, inversely proportional to length. The resultant point sample is equiprobable on the population of stream reaches. Then, in terms of the sample of reaches,

$$\hat{L}_r = \frac{wD_r}{n^*_{2r}} \sum_{i=1}^{n^*_{2r}} \frac{\sum_{j=1}^{k_i} l_{rij}}{l^*_{ri}} = \frac{wD_r}{n^*_{2r}} \sum_{i=1}^{n^*_{2r}} \frac{z_{ri}}{l^*_{ri}} \tag{15a}$$

estimates the total length of population reaches, where for resource r, l_{rij} represents the length of the j^{th} actual reach in the i^{th} sampling unit, l^*_{ri} represents the length of the i^{th} frame reach, and z_{ri} represents the sum of length of all reaches in the i^{th} sampling unit. Recall that a sampling unit is a frame reach. Also, D_r is the total frame reach length in the Tier 1 sample of resource r and $\hat{L}^*_r = wD_r$ is the Tier 1 estimate of L^*_r. Because L^*_r is known on the frame, wD_r is replaced by L^*_r, resulting in·

$$\hat{L}_r = L^*_r \hat{R} \tag{15b}$$

where $\hat{R} = \frac{1}{n^*_{2r}} \sum_{i=1}^{n^*_{2r}} \frac{z_{ri}}{l^*_{ri}}$. Also,

$$\hat{N}_r = w \sum_{i=1}^{n^*_{2r}} \frac{k_{ri}}{\pi_{ri}} = \frac{wD_r}{n^*_{2r}} \sum_{i=1}^{n^*_{2r}} \frac{k_{ri}}{l^*_{ri}} \ , \tag{16a}$$

where k_{ri} represents the number of actual reaches in the i^{th} sampling unit. Again, wD_r can be replaced by L^*_r, which is known for the frame, resulting in:

$$\hat{N}_r = \frac{L_r^*}{n_{2r}^*} \sum_{i=1}^{n_{2r}^*} \frac{k_{ri}}{l_{ri}^*} = L_r^* \hat{R} \,. \qquad (16b)$$

For these estimates, the variance estimators for $\hat{L}_r$ and $\hat{N}_r$, are given by $L_r^{*2}\text{var}(\hat{R})$, where the variance of the ratio can be approximated by:

$$\text{var}(\hat{R}) = \frac{1}{n_{2r}^*(n_{2r}^* - 1)} \sum_{i=1}^{n_{2r}^*} (v_{ri} - \hat{R})^2 \quad , \qquad (17)$$

where $v_{ri} = \frac{z_{ri}}{l_{ri}^*}$, when computing $\text{var}(\hat{L}_r)$, or $v_{ri} = \frac{k_{ri}}{l_{ri}^*}$, when computing $\text{var}(\hat{N}_r)$. Note that this formula is different from most ratio variance estimators in this report.

The distribution function is estimated from the data collected at the sample points, not from the set of sample reaches, as in the above estimation of N and L. Recall that each selected frame reach will have an associated sample point; this may result in 0, 1, 2, or more sample points for the actual streams represented by the frame reach. Let:

$$\hat{F}_r(y) = \frac{n_r(y)}{n_r} \,, \qquad (18)$$

where n_r is the total number of sample points in the resource, as realized in stream reaches, and $n_r(y)$ is the number of these for which the observed indicator value is less than y; $n_r(y) = \sum_S I(y_{ij} < y)$, with summation over all frame reaches, i, and all points, j, for each frame reach. Then, rewrite $n_r(y) = \sum_{i=1}^{n_{2r}^*} z_{ri}(y)$, so that $z_{ri}(y) = \sum_{j|i} I(y_{ij} < y)$.

For $\hat{F}_r(y)$, under the randomization assumption, it is appropriate to treat $n_r(y)$ as conditionally binomial(n_r, $F_r(y)$) and to use the binomial algorithm. Alternatively, the variance of $\hat{F}_r(y)$ can be estimated in the manner of ratio estimators. For the equiprobable point sample, this is:

$$\text{var}(\hat{F}_r(y)) = \sum_{S_2} d_{ri}^2 w_{2i}^2 + \sum_{S_2} \sum_{S_2} d_{ri} d_{rj} (w_{2i} w_{2j} - w_{2ij}) / n_{2r}^{*2} \,, \qquad (19a)$$

where $d_{rt}(y)=(z_{rt}(y) - \hat{F}_r(y)x_{rt})$, $w_{2t}= \frac{wD_r}{n^*_{2r}}$, and $w_{2tj}= \frac{w^2D_r^2}{n^*_{2r}(n^*_{2r}-1)}$. Here, x_{rt} equals the number of sample points for the t^{th} frame reach. This then simplifies to:

$$\text{var}(\hat{F}_r(y)) = \frac{w^2}{n^*_{2r}(n^*_{2r}-1)}[\sum_{S_2} d^2_{rt} - T^2_y/n^*_{2r}], \qquad (19b)$$

Then it is necessary to estimate L(y) as a product, $\hat{L}_r(y) = \hat{L}_r\hat{F}_r(y)$, with variance estimator, $\text{var}(\hat{L}_r(y))=\hat{L}^2_r \text{ var}(\hat{F}_r(y)) + \hat{F}^2_r(y) \text{ var}(\hat{L}_r)$.

This analysis presumes that there are no strata for stream reaches For two strata (1^{st} and 2^+ order), simple modification of these formulae will suffice. The numbers and length of reaches in the cross-classified strata are estimated and then combined. For $\hat{F}$, sample points from units in the wrong stratum are simply combined with the correct stratum If more than these two strata are desired, the general method of frame correction via sample unit correction is not feasible, and the method prescribed here is not appropriate.

2.1.2 Extensive Resources

The universe of an extensive resource is a continuous spatial region. If the domain is correctly identified, the universe of the resource will be a subset of the domain and may be fragmented over that domain Extensive resources may have populations of two kinds, continuous or discontinuous. Because these discontinuous populations are defined on a continuous universe, they are referred to as extensive resources. Continuous populations are referred to as extensive resources as well. Section 3.3.4 of the EMAP design report (Overton et al., 1990) describes two methods for sampling extensive populations, via a point or areal sample. For each resource, the design provides for the classification of a large areal sample (40-hex) at each grid point; these areal samples are also subject to subsampling via points or areal subsamples.

At Tier 2, two distinct directions are available, depending on the nature of the resource. Specifically, if the domain of the resource is well known from existing materials, as are boundaries of the Chesapeake Bay or Lake Superior, then the Tier 1 areal sample is of little value either in estimating extent or in obtaining a sample at Tier 2. In these cases, the domain should correspond to the universe. Conversely, if the spatial distribution or pattern of a resource is poorly known, as it will be for certain arid land types or for certain types of wetlands, then the Tier 1 areal sample may provide the best basis for obtaining a well-distributed sample at Tier 2. Other factors, such as size of the domain and degree of correspondence of universe and domain, will influence the sampling design. In the first circumstance, the Tier 2 sample will be selected from the areal sample obtained at Tier 1. In the other, the Tier 2 sample will be selected from the known universe by a higher resolution point sample that contains the base Tier 1 sample.

2.1.2.1 Areal Samples

Tier 1

All Tier 1 areal samples are expected to be collected with equal probability. Enhancement of the grid may be made for any resource, but any resource should have uniform grid density over its domain. Further, the areal sample imposed on the grid points will be of the same size for any resource, so that algorithms are presented only for equal probability sampling. The following formula estimates the total areal extent of a particular resource (r) over its domain $\mathcal{D}_r$:

$$\hat{A}_r = w \sum_{\mathcal{D}_r} a_{ri} \tag{20}$$

where the domain was discussed in Section 2.1.1.2. The value a_{ri} defines the area of resource r in the areal sample at grid point i, and w is the inverse of the density of the grid divided by the size of the areal sample. For equal probability sampling, the variance

estimator is

$$\text{var}(\hat{A}_r) = \frac{n_r w(w-1)}{n_r - 1} \left\{ \sum_{\mathfrak{D}_r} a_{rs}^2 - \left(\sum_{\mathfrak{D}_r} a_{rs} \right)^2 / n_r \right\} \tag{21}$$

where n is the number of whole or partial areal sample hexagons located in $\mathfrak{D}_r$. As with the discrete resources, even though the sample is selected by a systematic grid, we assume, in order to estimate variance, that the sample was taken from a locally randomized scheme. The justification of this assumption is similar to that for discrete resources. Alternate procedures are available when the assumption is questioned (see Section 2.1.2.3).

Tier 2

At the second stage of sampling for extensive resources, the distribution function for a particular resource is estimated. To identify the objective of Tier 2 sampling, we can write estimating equations as though a complete census were made at Tier 2. The general conceptual formula for the distribution function of areal extent for a specific resource (r) over its domain $\mathfrak{D}_r$ is

$$\hat{A}_r(y) = w_r \sum_{s \in \mathfrak{D}_r} a_{rs}(y) \quad , \tag{22}$$

where $a_{rs}(y)$ is the area of the resource in areal sample s such that the value of the indicator is less than y. The estimated variance follows Equation 21 as

$$\text{var}(\hat{A}_r(y)) = \frac{n_r w_r (w_r - 1)}{n_r - 1} \left\{ \sum_{\mathfrak{D}_r} a_{rs}^2(y) - \left(\sum_{\mathfrak{D}_r} a_{rs}(y) \right)^2 / n_r \right\} \tag{23}$$

The estimate of areal proportion for an extensive population divides Equation 22 by the estimate of total areal extent:

$$\widehat{F}_r(y) = \frac{\widehat{A}_r(y)}{\widehat{A}_r} \quad . \tag{24}$$

In the rare instance in which A_r is known, then an improved estimator of $A_r(y)$ is given by

$$\tilde{A}_r(y) = A_r \widehat{F}_r(y) \quad . \tag{25}$$

Ordinarily, these distribution functions will be calculated at each distinct value of y appearing in the sample. The variance associated with the areal proportion is the general form for a ratio estimator (Särndal et al., 1992, Equation 7.2.11) In writing this expression, it is necessary to identify the specific value, y_t, at which $\widehat{F}_r(y)$ is being assessed.

$$\mathrm{var}(\widehat{F}_r(y_t)) = [\sum_j d_j^2 w_j(w_j - 1) \; + \; \sum_j \sum_{k \neq j} d_j d_k (w_j w_k - w_{jk})]/\widehat{A}_r^2 \quad , \tag{26}$$

where $d_j = [a_{rj}(y_t) - a_{rj}\widehat{F}_r(y_t)]$, a_{rj} is the area of sample j in resource r, w_{jk} is defined as in Equation 5, and $\widehat{A}_r^2$ is replaced by A_r^2 when A_r is known

In practice, the Tier 2 assessment will be based on a subsample of some kind, and the above ideal estimation will not be available. The only method proposed for subsampling is use of a Tier 2 point sample.

2.1.2.2 Point Samples

Two methods of directly sampling objects from the grid points are discussed in the EMAP design report (Overton et al., 1990, Section 4.3.3.2). A Tier 1 reduced sample of discrete resource units can be selected by choosing the units into whose areas of influence the points fall; this method is not currently scheduled for use, but it is a viable method for several discrete resources. The same procedure can be used to select areal sampling units from an arbitrary spatial partitioning of the United States. The agroecosystem component

of EMAP provides such an example: the units selected for the sample are secondary sampling units of the National Agricultural Statistics Service (NASS) frame, and estimates are of totals over subsets of the frame units. Each selected unit is a mosaic of fields and other land use structures. These structures are then classified and sampled to provide ecological indicators for characterizing the sampling unit. Essentially, this areal sample is analyzed exactly like the 40-hex fixed areal unit discussed in the previous section, with the exception that inclusion probabilities are now proportional to the size of the unit, and the general formulae (e.g., Equations 2-4) must be used.

An alternate use of the point sample can be applied to an extensive resource, with the ecological indicators of the resource measured at the grid points. For continuous populations, such as temperature or pH, the response can be measured exactly at a selected point. For other populations, it is necessary to make observations on a support region surrounding the point, like a quadrat. For example, the wetlands resource group could obtain an indicator, such as plant diversity, from a quadrat sample centered on a grid point. The indicator measured in the quadrat can be treated like a point measurement. A cluster of quadrats centered on the grid point provides yet another method for sampling extensive resources.

This point sample will be applied at Tier 2 in either of two ways. For resources that depend on the Tier 1 areal sample to provide a sample frame, a high-resolution sample of points is to be imposed on each 40-hex containing the resource; this arrangement will generate an equiprobable point sample of the areal fragments of all resources that were identified at Tier 1. For a resource in which the universe is clearly identified, such as Lake Superior, a better spatial pattern of sample points will be obtained by imposing an enhanced grid on the entire universe. In the latter case, the universe is known, whereas in the former case, the Tier 1 sample provides a sample of the universe, which is then sampled by a Tier 2 point sample.

In either case, an equiprobable sample of points is obtained from which resource indicators will be measured, and the estimation equations will differ only by the weights. Variance estimators will differ, as one is a single-stage sample and the other is a double sample.

Point sample for a universe with well-defined boundaries

For a resource in which the universe is known (e.g., the Chesapeake Bay), the general formula for equiprobable point samples for a resource class is presented. A resource class is defined as a subset of the resource. For example, two classes of substrate, sand and mud, can be defined in the Chesapeake Bay. The distribution function of the proportion of a specific class of a specific resource (rc) having the indicator $\leq$ y reduces to

$$\hat{F}_{rc}(y) = \frac{n_{rc}(y)}{n_{rc}} \quad , \tag{27}$$

where $n_{rc}(y)$ is the number of points in resource class rc with the indicator equal to or less than a specific value, y, and n_{rc} is the total number of sample points in the resource class rc. Under the randomization assumption, the conditional distribution of $n_{rc}(y)$, given n_{rc}, is Binomial(n_{rc}, $F_{rc}(y)$), so that confidence bounds are readily set by the binomial algorithm in those instances in which spatial patterns indicate adequacy of the randomization model (Overton et al., 1990, Section 4.3.5). Alternate protocols are available when the randomization model cannot be assumed (Section 2.1.2.3).

Estimation of the area occupied by an extensive resource class is provided by

$$\hat{A}_{rc} = A_r \frac{n_{rc}}{n_r} = A_r \hat{p}_{rc} \quad , \tag{28}$$

where n_r is the number of grid points falling into the domain of the resource, and A_r is the area of the resource. Under the randomization assumption, n_{rc}, conditional on n_r, is a

binomial random variable; bounds on p_{rc} are again set by the binomial algorithm, as are bounds on A_{rc}.

Point sample for universe with poorly defined boundaries

When the universe of the resource is not known and one must use the Tier 1 areal sample as a base for the Tier 2 sample, then Equation 19 provides the estimates of A_r at Tier 1. Then the Tier 2 sample is an equiprobable sample of points selected from the area of the resource class contained in the 40-hexes. This procedure is implemented as a tessellation stratified sample in each 40-hex, with k=1 to 6 sample points per 40-hex. With only 1 point per 40-hex, the binomial algorithm will be appropriate under the randomization assumption; multiple points per 40-hex will require an explicit design-based expression for variance. In all cases,

$$\hat{A}_{rc} = \hat{A}_r \frac{n_{rc}}{n_r} = \hat{A}_r \hat{p}_{rc} \quad , \tag{29a}$$

$$\hat{F}_{rc}(y) = \frac{n_{rc}(y)}{n_{rc}} \quad , \tag{29b}$$

$$\hat{A}_{rc}(y) = \hat{A}_{rc}\hat{F}_{rc}(y) = \hat{A}_r \frac{n_{rc}(y)}{n_r} = \hat{A}_r \hat{R} \quad . \tag{29c}$$

It should be recognized that Equation 29a is a special case of Equation 29c.

When k>1, the following variance formula is appropriate:

$$\mathrm{var}(\hat{F}_{rc}(y)) = \frac{1}{k(k\text{-}1)n_{rc}^2} \sum_{hex} \left\{ \sum_{i=1}^{k} d_{ij}^2 + \left(\sum_{i=1}^{k} d_i\right)^2 / k \right\} \tag{30}$$

The outside summation is over the 40-hexes and $d_{ij} = (I(rc, y_{ij} < y) - \hat{F}_{rc}(y)I(rc))$. This expression is derived from the general Horvitz-Thompson formula used with ratio estimators. The formula can be recognized as the usual stratified random sample variance

formula, applied to $d_{i,j}$.

In addition,

$$\text{var}(\hat{A}_{rc}(y)) = \text{var}(\hat{A}_r)\,\hat{R}^2 + \hat{A}_r^2\,\text{var}(\hat{R}) \tag{31a}$$

where $\text{var}(\hat{R})$ follows,

$$\text{var}(\hat{R}) = \frac{1}{k(k\text{-}1)n_r^2} \sum_{hex} \left\{ \sum_{i=1}^{k} d_{i,j}^2 + \left(\sum_{i=1}^{k} d_i \right)^2 / k \right\} \tag{31b}$$

where $d_{i,j} = [I(rc, y_{i,j} < y) - I(r)\hat{R}]$.

Note that $\text{var}(\hat{A}_r(y)) = \text{var}(\hat{A}_r)\hat{F}_r^2(y) + \hat{A}_r^2\text{var}(\hat{F}_r(y))$; $\hat{F}$ replaces $\hat{R}$ in Equation 31a as well as in Equation 29c.

2.1.2.3 Alternative Variance Estimators

Confidence bounds for distribution functions based on point samples of continuous and extensive populations can be computed by several methods. The choice of a method is determined by the pattern of the resource area. First, the binomial approach is suggested for fragmented area distributed randomly across the domain. When this condition has been met, the randomization assumption holds and the binomial model is appropriate for computing confidence bounds.

If the area, $A_r(y)$, is in an entire block, rather than fragmented, then the binomial algorithm will overestimate variance, and alternative estimators will be needed. Other methods allow for a nonfragmented area and the randomization assumption is not required. The mean square successive difference (MSSD) is suggested for a strict systematic sampling scheme. Another method, the probability sampling method using the Yates-Grundy variance estimator, requires that the design have all positive pairwise inclusion probabilities.

One such design that provides this structure is the two-stage tessellation stratified model. The MSSD is discussed by Overton and Stehman (1993a) and the probability estimator is discussed by Cordy (in press). Methods of spatial statistics are also available for estimating this variance.

Mean square successive difference estimator

The variance estimator based on the mean square successive difference is intended to provide an estimate of variance for either the mean of values from a set of points on a triangular grid or obtained from a random positioning of the tessellation cells of the hexagonal dual to the triangular grid. In the latter case, the data are analyzed as though the values were taken from the center of the tessellation cell. The data set consists of all points falling in the target resource The MSSD has not been developed for this tessellation formed by triangular decomposition of the hexagons

Smoothing

Smoothing often results in improved variance estimation (Overton and Stehman, 1993a). The following method is from that report. For each datum, y, calculate a 'smoothed' value, y^*, as a weighted average of the datum and its immediate neighbors (i.e.: distance of one sampling interval). Weighting for this procedure is provided below. As a result, two new statistics are generated at each point: y^* and λ.

Number of Neighbors	y^* values	λ values
6	$(6y_i + \sum y_j)/12$	7/24
5	$(7y_i + \sum y_j)/12$	5/24
4	$(8y_i + \sum y_j)/12$	5/36
3	$(9y_i + \sum y_j)/12$	1/12
2	$(10y_i + \sum y_j)/12$	1/24
1	$(11y_i + \sum y_j)/12$	1/72
0	y_i	0

Given these smoothed values, summing over all data points,

$$\hat{\sigma}^2 = \frac{\sum (y - y^*)^2}{\sum \lambda} . \tag{32}$$

Mean Square Successive Difference

Identify the data along the three axes of the triangular grid; each point will appear once in the analyses of each axis. Analyze the y^*, not the original y. For each axis, calculate

$$\delta = \sum (y_a^* - y_b^*)^2 , \tag{33}$$

where y_a and y_b represent members of a pair of adjacent points, and where the summation is over all adjacent pairs identified on this axis. Also, calculate for each axis,

$$S^2 = [\sum (y_a^* - y_b^*)]^2 , \tag{34}$$

where it is necessary that all pair differences be taken in the same direction. From these statistics, calculate for each axis,

$$s^2 = \frac{(\delta^2 - \frac{S^2}{m})}{2(m-1)} \tag{35a}$$

and

$$\Delta^2 = \frac{(S^2 - \delta^2)}{m(m-1)} \quad , \tag{35b}$$

where m denotes the number of pairs in the above summations.

These statistics are then combined over the three axes, where summation is over all successive pairs in the k^{th} axis.

$$v_1 = 2\sum_{k=1}^{3} \frac{\Delta_k^2}{43.2} \quad , \tag{36a}$$

$$s^2 = \frac{\sum_{k=1}^{3} (m_k - 1)s_k^2}{\sum_{k=1}^{3} (m_k - 1)} \quad . \tag{36b}$$

Lastly, the following are computed to provide estimates of the variance of the mean values.

$$\hat{V}(\bar{y}_{sys}) = v_1 + \frac{\hat{\sigma}^2}{n_r} \tag{37a}$$

and

$$\hat{V}(\bar{y}_{str}) = \frac{\hat{\sigma}^2}{n_r} \quad , \tag{37b}$$

where n_r equals the number of sample points in resource r. This method has not been extended to distribution functions, but the extension is straightforward.

Yates-Grundy variance estimator for tessellation stratified probability samples

Investigation of this variance estimator is continuing. The method will be included in the next version of this manual.

2.2 Model-Based Estimation Methods

The previous section was devoted to design-based methods used to derive population estimates, distribution functions, and confidence bounds. Model-based estimation is another common approach to computing population estimates. In this approach, certain assumptions with regard to the underlying model are made, and the information provided by auxiliary variables often provides greater precision of the estimates.

Within EMAP, these model-based methods have not been developed to the same degree as the design-based methods. No algorithms for confidence bounds of distribution functions using model-based methods are presented in this report, although they are under development. The purpose of including this section is to provide a brief description of currently available model-based methods. Further, application of the model-based methods has so far been restricted to discrete populations. Investigation of the applicability of these methods in continuous and extensive populations is under way.

Three ways in which model-based methods can be used within EMAP are discussed: (1) data collected on the full frame across the population can be incorporated into the estimation process using prediction estimators to improve precision; (2) because the EMAP design is a double sample (Section 2.1.2.2), auxiliary variables on the first-stage sample can be used to improve the precision at the second stage; and (3) a calibration method is described for modifying an indicator variable to adjust for changes in instrumentation or protocol such methods are needed to maintain the viability of a long-term monitoring program.

The strategy is to begin with the basic design-based methods and to incorporate

model-based methods as the opportunity to do so becomes apparent and the necessary frame materials are developed. The design-based methodology will be enhanced by the use of models whenever feasible.

2.2.1 Prediction Estimator

If auxiliary data that can be used to predict certain indicators are available on the entire frame, model-based prediction techniques can be used to obtain predictions of the response variable for the population. These predictions then become the base for population inference

These methods require a vector of predictor variables defined on the frame, while the response variable is measured on the Tier 2 sample. A model is postulated for the relationship between the response variable, y, and the vector of predictor variables, x:

$$y=g(x) + \epsilon, \quad \text{with } \mathrm{Var}(\epsilon) = h(x) \quad . \tag{38}$$

Based on this model, a predictor equation, $\hat{y}=\hat{g}(x)$, is estimated from the Tier 2 sample.

The equation for the basic estimator, which is referred to as the general regression estimator, is defined as

$$\hat{T}_y = \sum_{\mathcal{U}} \hat{y}_i + \sum_{S_2} w_{2i} (y_i - \hat{y}_i) \quad , \tag{39}$$

where $\mathcal{U}$ and S_2 designate the universe of units and sample units at Tier 2, respectively. The variance of this estimator is estimated by

$$\mathrm{var}(\hat{T}_y) = \sum_{S_2} d_i^2 w_{2i} (w_{2i} - 1) + \sum_{S_2} \sum_{\substack{S_2 \\ j \neq i}} d_i d_j (w_{2i} w_{2j} - w_{2ij}) \quad , \tag{40}$$

where $d_i = (y_i - \hat{y}_i)$ (Sarndal et al., 1992, Equation 7.2.11). Our experience (Overton and Stehman, 1993b) suggests that this equation slightly underestimates the variance; this result is not unexpected because Equation 40 is based only on the variance of the second term of Equation 39.

One model-based estimator of the distribution function of the proportion of numbers, as established by Rao et al. (1990), is based on the general regression estimator and defined as

$$\hat{F}(y) = \frac{1}{N}\left\{\sum_{\mathcal{U}} \hat{P}[(\hat{y}_i+\epsilon_i) \le y] + \sum_{S_2}[I(y_i \le y) - \hat{P}\left((\hat{y}_i+\epsilon_i) \le y\right)]w_{2i}\right\} , \quad (41)$$

where N is the target population size, and $I(y_i \le y)=1$ if $y_i \le y$, 0 otherwise. Equation 41 gives an unsatisfactory representation of F(y); Dr. Dave Thomas (personal communication) has developed an improved version of this estimator, along with an adequate representation of confidence bounds that will be available in the near future. Other estimators, such as one based on the simple regression estimator, are also available; the form proposed by Chambers and Dunstan (1986) is more satisfactory than Equation 41, but not as suitable as Thomas' modification of Equation 41. The Chambers-Dunstan estimation is identical to Equation 41, with the w_{2i} deleted. Preliminary investigation indicates great improvement in precision using either of the model-based estimates of distribution functions.

Estimation of the regression coefficients deserves some attention. For Equations 39 and 42, it has been established that ordinary least squares predictors are adequate (Overton and Stehman, 1993b). For Equations 41 and 44, it is necessary to use generalized least squares predictors. In all of these methods, it seems unnecessary to use π-weighted estimates of regression coefficients.

2.2.2 Double Samples

As mentioned previously, the EMAP design is a double sample with Tier 1 representing the first stage (or phase) and Tier 2 the second stage. Through most of this document, design-based methods are provided for the Tier 2 sample; these methods are similar to those described for single-stage samples. However, where model-based methods are used, double sampling formulae can be quite different from single-stage formulae. An elementary discussion of double sampling with model-based methods is presented in Cochran (1977).

Existence of an auxiliary variable on the Tier 1 sample will enable model-based double-sample methods at Tier 2. EMAP does not require a resource-specific frame, but it does allow for acquisition of more detailed information for many resources. There is a Tier 1 sample for all resources, and for most resources, the Tier 2 sample is a subset of the Tier 1 sample, thus providing the basis for model-based double-sample methods.

The model specification follows the developments under the general prediction model (Equation 38). The basic estimator, derived from the general regression estimator, is defined as

$$\hat{T}_y = \left\{ \sum_{S_1} w_{1,i}\hat{y}_i + \sum_{S_2} w_{2,i}(y_i - \hat{y}_i) \right\} \quad , \tag{42}$$

where S_1 and S_2 define the sample at Tiers 1 and 2, respectively. The form of this estimator allows equal or variable probability at Tier 1. The variance estimator for Equation 42 follows Sarndal et al. (1992, p.365, Eq. 9.7.28):

$$\mathrm{var}(\hat{T}_y) = \sum_{S_2}\sum_{S_2}(w_{1,i}w_{1,j} - w_{1,i,j})y_i y_j w_{2.1,i,j} + \sum_{S_2}\sum_{S_2}(w_{2.1,i}w_{2.1,j} - w_{2.1,i,j})d_i w_{1,i} d_j w_{1,j} \, , \tag{43}$$

where $d_i = (y_i - \hat{y}_i)$.

The estimate of the distribution function of the proportion of numbers is developed as an extension of Equation 41,

$$\hat{F}(y) = \frac{1}{\hat{N}} \left\{ \sum_{S_1} \hat{P}[(\hat{y}_i + \epsilon_i) \le y] w_{1i} + \sum_{S_2} \left[I(y_i \le y) - \hat{P}\left((\hat{y}_i + \epsilon_i) \le y\right) \right] w_{2i} \right\} . \quad (44)$$

When N is unknown, $\hat{N}$ is a suitable replacement. Smoothed versions of Equations 41 and 44, along with confidence bound algorithms, are under development.

2.2.3 Calibration

Calibration is defined as the replacement of one variable in the data set by a function of that variable representing another variable. For example, in a long-term monitoring program such as EMAP, it is expected that some laboratory or data management protocols will change over time. Using this analytical tool, data from old protocols can be calibrated to represent data from the new protocols, thereby allowing assessment of trends across the transition.

Overton (1987a, 1987b, 1989a) described the application of calibration issues for the National Surface Water Surveys. In that instance, protocols were unchanged but the extensive data of 1984 were calibrated to the same variable in 1986 to take advantage of the strong predictive relationship through the double sample. The algorithms are provided for this calibration in Technical Report 130 (Overton, 1989a). Tailoring of these methods to the specific needs of EMAP will be required in certain instances. However, each application is likely to present some unique issues and properties, so that general development does not appear feasible.

2.3 Other Issues

2.3.1 Missing Data

Two types of missing data are expected to arise in EMAP. One type is a missing sampling unit, such as a missing lake. The other type of missing value occurs within a sampling unit, such as a missing observation on a specific chemical variable or a missing suite of chemical variables for a lake. In this situation, information is available on some, but not all, indicators for a specific unit or site.

2.3.1.1 Missing Sampling Units

There appears to be no basis for imputation of a missing sampling unit where no Tier 2 information is available to predict that observation. Therefore, missing sampling units should be considered as representing a subset of the subpopulation of interest that is unavailable for measurement. All procedures outlined in this document accommodate data sets that contain missing units. No adjustments to the weighting factors are necessary; summation is over the observed portion of the sample, and the estimates produced apply to the subpopulation represented by the sample analyzed. When Yates-Grundy estimation of variance is used, it will be necessary to modify the equation; this requirement is the primary reason for using the Horvitz-Thompson variance estimator when possible.

In a long-term program, this approach of classifying missing units with the subpopulation not represented by the sample is clearly appropriate; such units can be sampled in subsequent years without having to modify sample weights again. This approach is also consistent with the practice of allowing sampling units to change subpopulation classes from time to time. Comparisons must take this into account, but such class changes will always be a feature of long-term monitoring programs

A general problem remains when a substantial number of resource sites cannot be measured; EMAP must find a way to provide indicator values for such sites. When the

problem is severe, it might be possible to develop an alternate indicator suite that can be obtained via aerial television or photography. Perhaps it will be possible to impose a higher (lower resolution) sample level that will provide for model-based methods and predictors of the indicator. (This option will be difficult because the predictor relation must be developed specifically for the subpopulation of concern.) But whatever the solution, some method is required to provide representation of these sites. Until then, it is appropriate for these to be identified in the subpopulation for which no sample has been generated and about which nothing is known.

2.3.1.2 Missing Values within Sampling Units

It is advantageous to use information collected on a specific sampling unit to impute any missing observations for that sampling unit. To minimize error, a multivariate analysis is suggested, utilizing the data collected for that particular unit. No specific procedure is suggested for this analysis, because most standard analyses will impute similar values, and because the method must be tailored to the circumstances. Some multivariate procedures are discussed in statistics books that concentrate on imputation of missing values (cf., Little and Rubin, 1987).

2.3.2 Censored Data

For certain measurements, values for indicators will be less than the identified detection limit; exact values cannot be measured for such units or sites. This problem is not uncommon and has been discussed frequently in the literature applying to water quality management programs (cf., Porter et al., 1988). Caution is prescribed when characterizing data that consist of many observations below the detection limit. Proper analysis and reporting can prevent improper inference for these data; specifically, it must be noted that although reliable values are not provided, a great deal is known about the site that has a value at or below the detection limit.

To guide the data analyst in the treatment of the indicator that contains censored observations, the proximity of the detection limit to the critical value of the indicators needs to be considered. Indicators, such as chemical variables, that have detection limits near or above the critical value should not be considered meaningful indicators; the information supplied by such an indicator is too fuzzy to justify inferences. In such cases, the most meaningful parameters are those whose estimates are not affected by censoring. Other indicators have a detection limit well below the critical value. For these indicators, it is suggested that values below the detection limit should be scored to the detection limit and analyzed with the uncensored data.

The mean is a poorly defined statistic to describe censored data. However, the scored mean can be interpreted, even though it is slightly biased. Another statistic, the scored mean minus the detection limit, is unbiased for the mean in excess of the detection limit, which is a well-defined population parameter. If the distribution below the detection limit is modeled, and the mean value below the detection limit is calculated, then the scored mean can be converted into an unbiased estimate of the true mean, given the model.

On the other hand, the median is less ambiguous than the mean and is more appropriate for characterizing these indicators. Usually the median will not be affected by scoring. Distribution functions also should not be described below the detection limit. This restriction is another reason for scoring; standard analyses of the scored data yield the desired distribution function, emphasizing that the shape of the curve below the detection limit is unknown. Because the critical level changes with circumstance, it is desirable to present the truncated (scored) distribution function, to be interpreted as the situation dictates. In fact, the capacity to truncate the distribution function without impairing inferences is one of the strong arguments for choosing this parameter to characterize these data.

Modeling the function below the detection limit is one method proposed in the statistical literature to modify estimates from censored data (Cox and Oakes, 1984; Miller, 1981). However, a hypothetical distribution must be assumed to represent the censored data. In EMAP, distributions are defined on real populations and are unlikely to follow any distributional law. We propose that the distribution function reflect the data alone and that the unsupported portion of the distribution function is not described. Use of the scored mean is somewhat less justifiable, but generally consistent with this position.

2.3.3 Combining Strata

The strata that form the structure of the Tier 2 sample are established from classes of resources identified at Tier 1, on the Tier 1 sample. The seven basic resources are the foundation of this structure, but there is provision for further classification leading to several strata for lakes, several for forests, and so on. These strata are referred to as resources in this report.

Tier 2 selection is then stratum (resource) specific and independent among strata. This structure is chosen to provide inferences within strata, with the thought that few occasions will arise for inferences involving combined strata. For example, a distribution function [F(y)] combining small and middle-sized lakes will be dominated by small lakes. If the population of large lakes is of interest, it must be characterized separately. Further, a wide range of sizes makes the frequency distribution less useful in characterizing the population. Still, because there may be interest in a population consisting of the largest of the small lakes and the smaller of the middle-sized lakes, analysis of combined strata is needed.

2.3.3.1 Discrete Resources

Samples are combined across strata to compute the Tier 2 estimates Weights will not be uniform, so the Horvitz-Thompson algorithms using weights are needed. Estimation of $N_a(y)$ and $F_a(y)$ is identical to the estimation algorithms for a single stratum, but estimation of variance requires modification. The basic formula for estimating variance is also unchanged, only the $w_{2,ij}$ must be modified. Specifically,

if i and j are from the same stratum, then

$$w_{2,ij} = \frac{2n_2 w_{2,i} w_{2j} - w_{2,i} - w_{2j}}{2(n_2 - 1)} \quad ; \tag{45}$$

or if i and j are from different strata, then

$$w_{2,ij} = w_{1,ij} w_{2.1,i} w_{2.1j} \quad , \tag{46}$$

where, if i and j are from different 40-hexes, then

$$w_{1,ij} = \frac{2n_1 w_{1,i} w_{1j} - w_{1,i} - w_{1j}}{2(n_1 - 1)} \quad , \tag{47}$$

or, if i and j are from the same 40-hex, then

$$w_{1,ij} = \frac{w_{1,i} w_{1j}}{w} \quad , \tag{48}$$

where w is the weight associated with the basic Tier 1 areal sample.

In the case of the quasi-stratified design used for lakes and streams, the recommendation is that the sample be conditioned on the realized sample sizes in the several distinct classes having equal inclusion probabilities (within class). This approach leads to a

post-stratified sample that can be analyzed exactly like the sample from a stratified design. The gain in precision will carry over into analysis of combined strata in the manner discussed in this section.

2.3.3.2 Extensive Resources

Procedures for combining strata for point samples in extensive resources are the same as those outlined for discrete resources (Section 2.3.3.1). Methods to combine strata for areal samples in the extensive resources are still under consideration and will be addressed at a later time.

2.3.4 Additional Sources of Error

Other potential sources of error can be expected in the process of developing the distribution function and confidence bounds Some of these have been discussed after evaluation of the Eastern and Western Lake Surveys (Overton, 1989a, 1989b). These additional sources of error add to the uncertainty and bias of the estimated distribution function. Research is presently under way to investigate methods, such as deconvolution, to correct for these added components of error and bias. Preliminary methods are unsatisfactory, and two different approaches are being followed to improve results. These methods will be introduced to EMAP analyses as they become available.

The rounding of measurements reduces precision in quantiles and distribution function estimation. Analyses of the National Surface Water Surveys suggested that reporting data at two decimal places beyond the inherent accuracy of the indicator satisfactorily reduces bias attributed to rounding error (Overton, 1989b). It is recommended that additional decimal places be carried into the data set if they are provided by the instrumentation. Additional rounding should be made only at the reporting step, and the

rule for rounding should take into account gain in precision from averaging and other statistical practices.

2.3.5 Supplementary Units or Sites

Supplementary units, in addition to the yearly EMAP grid points, have been selected and measured or remeasured by some resource groups. For example, a set of supplementary units can be selected as a subset of one of the interpenetrating replicates. The remeasurement of these supplementary units is directed at specific issues, such as estimation of variance, and the selection procedure is likely to be influenced by this purpose. If data from supplementary probability samples are combined with the general EMAP sample, it is necessary to use a protocol for combining two probability samples. If the supplementary data are not from a probability sample, then it is necessary to use a protocol for combining found data with probability sample data (Overton et al., 1993). Ordinarily, a good strategy will be to use these supplementary data only for analyses initially intended. The effort necessary to satisfactorily combine supplementary data within the general sample analysis, such as the distribution functions, is sufficiently great that one should be reluctant to attempt this combination. On the other hand, there will be certain circumstances in which this effort is justifiable.

SECTION 3

DISTRIBUTION FUNCTION ALGORITHMS

The types of distribution function algorithms, along with their associated conditions for application, are presented in Table 1. The first part of this table (A) presents the various cases yielding the distribution of numbers, $\widehat{N}(y)$. Part B presents the various cases discussed in this report yielding the distribution functions for the proportions of numbers. The methods of obtaining the distribution functions for size-weighted statistics are presented in Part C.

To explain the notation presented in the following algorithms, some terminology is introduced. The target population size, N, is the size of the target subset of the universe of units, defined as $\mathcal{U}$. The following algorithms are written to obtain estimates over a particular subpopulation of interest. For a particular subpopulation (a), the distribution of numbers is denoted as $\widehat{N}_a(y)$ and the distribution of the proportion of numbers is denoted as $\widehat{F}_a(y)$. N_a denotes the subpopulation size over the subpopulation, a. In addition, the n and n_a refer to the sample size from the population and subpopulation, respectively.

The variance estimator discussed in Section 2 is based on the Horvitz-Thompson theorem and is appropriate for both equal and variable probability sampling, independent of a known population or subpopulation size. The confidence bounds using this variance estimator are then based on the normal approximation. Therefore, for any condition, the general Horvitz-Thompson algorithms for $N_a(y)$ and $F_a(y)$, as presented in the following subsections under variable probability sampling, are appropriate.

Estimation of these bounds simplifies under equal probability sampling when the size of either the population or the subpopulation is known. For example, an exact confidence bound for $\widehat{F}_a(y)$ can be based on the hypergeometric distribution in the case of equal

probability sampling when the subpopulation size is known. When the subpopulation size is unknown, these bounds can be based on the binomial distribution.

It should be emphasized that there are no differences in the distribution functions obtained from the alternative design-based approaches discussed in this report. Further, the distribution functions obtained under the same conditions based on the Horvitz-Thompson, the binomial, or the hypergeometric algorithm are the same. The differences occur in the computation of the confidence bounds. Note, however, that model-based distribution functions will be different from those obtained from design-based methods.

In all situations, the algorithms in this report provide two one-sided 95% confidence bounds. The combined upper and lower confidence bounds enable two-sided 90% confidence bounds on the distribution function The Horvitz-Thompson algorithm estimates standard errors from which the confidence bound is based on a normal approximation. The alternative methods directly provide confidence bounds based on the exact binomial or hypergeometric distributions. All design-based methods suggested for discrete populations assume the randomized model, as discussed in Section 2. Because exact methods are usually preferred over approximate methods, the exact methods are suggested for those cases by which the conditions justify their use.

A test data set was applied to the following algorithms. Any resource group interested in comparing their versions of these algorithms to the ones provided in this report are encouraged to contact the authors. A copy of the test data set will be provided in order to compare results from other programs.

3.1 Discrete Resources

In this section, examples are provided for each of the possible approaches to obtaining $\widehat{N}_a(y)$ and $\widehat{F}_a(y)$ for discrete resources. For each of these approaches, the conditions and assumptions of the selection of the sampling units are defined. For quick reference, Table 1

(Section 4) presents this information in condensed form. An interest in obtaining the distribution function of numbers and proportion of numbers across the subpopulation is expected for all resource groups. For example, the lakes and streams resource group can compute the numbers or proportions of numbers of lakes with some attribute based on this algorithm.

3.1.1 Estimation of Numbers

A number of algorithms are presented for computing the distribution function for numbers. The choice of the algorithm is dependent on whether the units were chosen by either equal or variable selection. The first three cases (algorithms) in this section derive the distribution functions based on an equal probability selection of units and the latter two cases (algorithms) are based on an unequal probability selection of units.

Equal Probability of Selection

In this subsection, three cases are provided based on information that is known or unknown. For the first algorithm, N is either known or unknown and N_a is known; this algorithm produces confidence bounds based on the hypergeometric distribution. For the second algorithm, N is known, but N_a is unknown; this algorithm is also based on the hypergeometric distribution. For the third algorithm, both N and N_a can be either known or unknown; this algorithm produces confidence bounds based on the Horvitz-Thompson variance estimator and the normal approximation.

Case 1– Estimation of $N_a(y)$: Discrete Resource, Equal Probabilities, N_a known, $n=n_a$. Confidence Bounds by Hypergeometric Algorithm.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is known.
3. There is an equal probability of selection of units from the subpopulation.
4. Sample size condition: $n=n_a$.

Outline for Algorithm

Under the given conditions, the confidence bounds can be obtained by either the exact hypergeometric distribution or by the normal approximation. This case provides the confidence bounds for $N_a(y)$ by the hypergeometric distribution, when N_a is known. The normal approximation bounds are provided in the next subsection (see Examples 4 and 5).

This algorithm computes the confidence bounds for each point along the curve using the hypergeometric distribution. In the following formula, N_a is the subpopulation size; n_a is the sample size from the subpopulation; $N_a(y)$ refers to the number of units, u, in the subpopulation, $\mathcal{A}$, for which $y_u \leq y$; and $n_a(y)$ refers to the number of units in the sample from N_a, S_a, for which $y_u \leq y$. Under these conditions, $n_a(y)$ has the following hypergeometric distribution. Let X represent the random variable for which $n_a(y)$ is a realization:

$$\frac{\binom{N_a(y)}{X}\binom{N_a - N_a(y)}{n_a - X}}{\binom{N_a}{n_a}} \tag{49}$$

The upper confidence bound is computed by obtaining the largest value of $N_a(y)$ for which $\text{Prob}[X \leq n_a(y)] \geq 0.05$. The lower confidence bound is computed by obtaining the smallest value of $N_a(y)$ for which $\text{Prob}[X \geq n_a(y)] \geq 0.05$.

A GAUSS program is presented here that derives the confidence bounds based on the hypergeometric distribution. Comments in capital letters in braces explain the programming steps. Under the conditions of Case 1, the upper and lower halves of the confidence bounds are symmetric.

CALCULATION OF CONFIDENCE BOUNDS ON $N_a(y)$ BY THE HYPERGEOMETRIC DISTRIBUTION

```
load x[a,b] = data;  {LOADS DATA FILE WHICH INCLUDES LABEL CODE AND
                      VARIABLE TO BE ANALYZED. HERE a DESIGNATES
                      THE SAMPLE SIZE, n_a, AND b DESIGNATES THE
                      NUMBER OF COLUMN VECTORS}
x=x[.,2];  {IDENTIFIES THE VARIABLE OF INTEREST IN SECOND COLUMN}
nm=rows(x);  {NUMBER OF ELEMENTS OF INTEREST IN SUBPOPULATION, n_a.
              IN THIS ALGORITHM, n=nm=n_a}
n=rows(x);
NN=N_a;  {DEFINES TOTAL SUBPOPULATION SIZE HERE, N_a }

x=sortc(x,2);          {SORTS VARIABLE OF INTEREST}
y=seqa(1,1,nm),        {CREATES SEQUENCE OF NUMBERS}
x2=x[.,2];             {DEFINES VARIABLE OF INTEREST AS X2}
x=y~x2;                {CREATES MATRIX x}
zz=x;                  {DEFINES MATRIX x as zz}

{THE FOLLOWING COMBINES RECORDS WITH DUPLICATE VALUES
     OF THE VARIABLE}
xx=zeros(1,2);
q=0;
 i=1;
 do while i < nm;
    if x[i,2]==x[i+1,2];
    q=q+1;  I;
       else;
       xx=xx|x[i,.];
    endif;
 i=i+1;
 endo;
xx=xx|x[nm,.];
x=xx;
```

```
{THE FOLLOWING STEPS BEGIN CONFIDENCE BOUND ESTIMATION}
r=rows(x),
z=zeros(r,1);
x1=x[.,1],
x2=x[.,2];
x=x2~x1~(NN*x1/nm)~z;


{THE FOLLOWING STEPS GENERATE THE UPPER CONFIDENCE BOUND}
i=1;
do while i <= r;   {BEGINS INITIAL DO LOOP}
    rr=x[i,2];
    mm=trunc(NN*rr/nm);
       if mm >= NN;
          goto three,
       endif;

one:,
 mm=mm+1;
    if NN <= 160,
       aa=n!*mm!/NN!*(NN-mm)!*(NN-n)!;
    else,
       aa=lnfact(n) + lnfact(mm) - lnfact(NN) + lnfact(NN-mm) + lnfact(NN-n);
    endif;
 j=0;
    if (NN-mm-n) < 0;
       j=-(NN-mm-n);
    endif,
 s=0;
 do while j <= rr,
    if NN <= 160,
       a=aa/j!/(mm-j)!/(n-j)!/(NN-mm-n+j)!;
    else;
       a=aa - lnfact(j) - lnfact(mm-j) - lnfact(n-j) - lnfact(NN-mm-n+j);
       a=exp(a);
    endif,
    s=s+a;
    j=j+1;
 endo;
    if s>= .05;
       goto one;
    endif;

three:;
 if mm>=NN;
    x[i,4]=NN;
 else;
    x[i,4]=mm-1;
 endif;
 i=i+1;
ENDO;   {ENDS INITIAL DO LOOP}
```

```
{THE FOLLOWING STEPS ADD AN EXTRA LINE TO DATA MATRIX NEEDED IN
  CONFIDENCE BOUND ADJUSTMENT COMPUTED AT END OF ALGORITHM}
r=rows(x);
y=zeros(r,1)
x=x~y;
y=zeros(1,5);
y[1,2:4]=x[r,2:4];
x=x|y;


{THE FOLLOWING STEPS GENERATE THE LOWER CONFIDENCE BOUND}
r=rows(x);
i=1;
do while i <= r;   {BEGINS SECOND DO LOOP}
 rr=x[i,2];
 mm=trunc(NN*rr/n);
   if mm==0;
      goto six;
   endif;

four:;
mm=mm-1;
 if NN <= 160;
   aa=n!*mm!/NN!*(NN-mm)!*(NN-n)!;
 else;
   aa=lnfact(n) + lnfact(mm) - lnfact(NN) + lnfact(NN-mm) + lnfact(NN-n);
 endif;
 j=rr;
 mnm=minc(n|mm);
 s=0;
do while j <= mnm;
 if NN <= 160;
   a=aa/j!/(mm-j)!/(n-j)!/(NN-mm-n+j)!;
 else;
   a=aa- lnfact(j) - lnfact(mm-j) - lnfact(n-j) - lnfact(NN-mm-n+j);
   a=exp(a),
 endif;
   s=s+a;
   j=j+1;
endo;
 if s>= .05;
   goto four;
 endif;

six:;
 if mm==0;
   x[i,5]=0;
 else;
   x[i,5]=mm+1;
 endif;
 i=i+1;
ENDO;   {ENDS SECOND DO LOOP}
```

```
{ASSIGN LABELS}
 "N= " NN ", n = " n;
x;
OUTPUT OFF;


{ADJUST N̂a(y) and CONFIDENCE BOUNDS - AVERAGE SUCCESSIVE VALUES}
r=rows(x);
xx=x;
i=2;
do while i <= r-1;
 xx[i,3:5]=(x[i,3:5] + x[i-1,3:5])/2;
 i=i+1;
endo;

{OUTPUT FILE AND PRINT MATRIX x}
OUTPUT FILE=NAME;
OUTPUT reset,
"  x  "  " Sequence #"  "  F(x)  "  "  F-lower(x)  "  "  F-upper(x) ";
format /m1/rd 12,7;
print x;
OUTPUT OFF;

end;
```

Case 2– Estimation of $N_a(y)$: Discrete Resource, Equal Probabilities, N known, N_a unknown, $n>n_a$. Confidence Bounds by Hypergeometric Algorithm.

Conditions for approach

1. The frame population size, N, is known.
2. The subpopulation size, N_a, is unknown.
3. There is an equal probability of selection of units from the subpopulation.
4. Sample size condition: $n>n_a$.

Outline for Algorithm

Under the given conditions, the confidence bounds can be obtained by either the exact hypergeometric distribution or by the normal approximation. This example provides the confidence bounds for $N_a(y)$ by the hypergeometric distribution, when N is known, but N_a is unknown. Normal approximation bounds are provided in the next subsection (see Examples 4 and 5).

This algorithm computes the confidence bounds for each point along the curve using the hypergeometric distribution. In the following formula, N is the frame population size; n is the sample size from the frame population; $N_a(y)$ refers to the number of units, u, in the subpopulation, $\mathcal{A}$, for which $y_u \le y$; and $n_a(y)$ refers to the number of units in the sample from N_a, S_a, for which $y_u \le y$. Under the conditions, $n_a(y)$ has the following hypergeometric distribution. Let X represent the random variable for which $n_a(y)$ is a realization. Note that $n_a(y) \le n_a$ and that $N_a(y) \le N_a < N$.

$$\frac{\binom{N_a(y)}{X}\binom{N - N_a(y)}{n - X}}{\binom{N}{n}} \tag{50}$$

The upper confidence bound is computed by obtaining the largest value of $N_a(y)$ for which $\text{Prob}[X \leq n_a(y)] \geq 0.05$. The lower confidence bound is computed by obtaining the smallest value of $N_a(y)$ for which $\text{Prob}[X \geq n_a(y)] \geq 0.05$.

To obtain the distribution function, the data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of numbers, $\widehat{N}_a(y)$, denotes the number of units in the target population that have the value less than or equal to the specific y. Conversely, if it is of interest to obtain bounds on the number of units in the target population with indicator values greater than or equal to y, the data file must be sorted and analyzed in descending order on this variable. The distribution function generated by the analysis in descending order is $[\widehat{N}_a - \widehat{N}_a(y)]$.

A GAUSS program provided in Case 1 derives the confidence bounds based on the hypergeometric distribution. However, under the conditions discussed here, the sample size and population sizes are defined as follows.

CALCULATION OF CONFIDENCE BOUNDS ON $N_a(y)$ BY THE HYPERGEOMETRIC DISTRIBUTION

```
load x[a,b] = data;  {LOADS DATA FILE WHICH INCLUDES LABEL CODE AND
                      VARIABLE TO BE ANALYZED. HERE a DESIGNATES
                      THE OBSERVED SAMPLE SIZE, n_a, AND b DESIGNATES
                      THE NUMBER OF COLUMN VECTORS}
x=x[.,1];
nm=rows(x);   {NUMBER OF ELEMENTS OBSERVED, n_a. IN THIS ALGORITHM,
               n ≠ n_a. }
n=#;          {FULL SAMPLE SIZE}
NN=N;         {DEFINES TOTAL POPULATION SIZE HERE }

REFER TO CASE 1 (AFTER LINE 13) FOR THE REMAINING STEPS
   IN THIS PROGRAM.
```

Case 3– Estimation of $N_a(y)$: Discrete Resource, Equal Probabilities. Confidence Bounds by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, can be known or unknown.
3. There is an equal probability of selection of units from the subpopulation. [Note that this algorithm can also be applied to those cases presented in Examples 1 and 2.]

Outline for Algorithm

The algorithm recommended, given the foregoing conditions, is based on the Horvitz-Thompson formulae, which were discussed in Section 2. The algorithm presented for the general case of variable probability of selection (the following subsection) is appropriate to use given the foregoing conditions.

Equal probability selection is a special case of variable probability selection. In equal probability of selection of units, the weighting factors are equal for all units, $w_i=w_j=w$. If the weights, w_{1i} and $w_{2.1i}$ are appropriately identified, then the general algorithm presented in Example 4 will not need any modification. The Tier 2 weight, w_{2i}, computed by Equation 4 is the same for all units.

Variable Probability Selection

In this subsection, two examples are provided to demonstrate variable probability of selection. For both cases, the frame population size can be known or unknown. In Case 4, N_a can be unknown or known and equal to $\widehat{N}_a$. For Case 5, N_a is known and not equal to $\widehat{N}_a$. Both algorithms produce confidence bounds based on the Horvitz-Thompson variance estimator and the normal approximation.

Case 4– Estimation of $N_a(y)$: Discrete Resource, Variable Probabilities, N_a unknown or N_a known and equal to $\hat{N}_a$. Confidence Bounds by by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is unknown or known and equal to $\hat{N}_a$
3. There is a variable probability of selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied for this example is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest. It is useful to identify the estimator of N_a from Tier 2. The design-based estimator of N_a is

$$\hat{N}_a = \sum_{S_a} w_{2i} \quad , \tag{51}$$

where S_a is the portion of the sample from the subpopulation, $\mathcal{A}$, over which the weighting factors (w) are summed. The variance estimator for $\hat{N}_a$ is presented in Equation 3b of Section 2.

Calculation of confidence bounds on $N_a(y)$ by the Horvitz-Thompson formulae .

For each indicator, the following algorithm derives the distribution function and the confidence bound for N(y) or $N_a(y)$. This algorithm is similar to the algorithm defined for the National Surface Water Surveys (Overton, 1987a,b). The Horvitz-Thompson variance estimator, discussed in Section 2.1, is used to compute the variance in this algorithm. The confidence bounds are computed based on a normal approximation.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, w_{1i}
   c. Tier 2 conditional weighting factor, $w_{2\,1i}$
   d. Indicator of interest (y)

2. Sorting of data

The data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of numbers, $\widehat{N}_a(y)$, denotes the number of units in the target population that have a value less than or equal to the y for a specific indicator Conversely, if it is of interest to estimate the number of units in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable. The distribution function generated by the analysis in descending order is $[\widehat{N}_a - \widehat{N}_a(y)]$.

3. Computation of weighting factors

The Tier 1 and Tier 2 weights are included for each observation in the data set. These weights are used to compute the total weight of selecting the i^{th} unit in the Tier 2 sample. Compute the following weight for each observation:

$$w_{2i} = w_{1i}w_{2.1i}$$

where w_{1i} is the weighting factor for the i^{th} unit in the Tier 1 sample (the inverse of its Tier 1 inclusion probability) and $w_{2.1i}$ is the inverse of the conditional Tier 2 inclusion probability.

4 Algorithm for $\widehat{N}_a(y)$

   a. Define a matrix of q column vectors, which will be defined as the following. There is one row for each data record and five statistics for each row.
      q_1 = value of y variable for the record
      $q_2 = \widehat{N}_a(y)$
      $q_3 = \text{var}[\widehat{N}_a(y)]$
      q_4 = upper confidence bound for $\widehat{N}_a(y)$
      q_5 = lower confidence bound for $\widehat{N}_a(y)$

   b. Index rows using i from 1 to n; the i^{th} row will contain q-values corresponding to the i^{th} record in the file, as analyzed.

   c. Read first observation (first row of data matrix), following with the successive observations, one at a time. Accumulate the q-statistics as each observation is read into file. Continue this loop until the end of file is reached. At that time, store these vectors and go to d. This algorithm is calculating the distribution for the number of units $[\widehat{N}_a(y)]$ in the subpopulation. It is necessary to identify the records for which $w_{2.1i}=1$.

i. $q_1[i] = y[i]$

ii. $q_2[i] = q_2[i-1] + w_{2,i}$

iii. $q_3[i] = q_3[i-1] + w_{2,i}*(w_{2,i}-1) + 2\sum_{j<i}(w_{2,i}w_{2,j} - w_{2,j})$

where, if neither $w_{2.1,i}$ nor $w_{2.1,j}=1$:

$$w_{2,i,j} = \frac{2n_2 w_{2,i}w_{2,j} - w_{2,i} - w_{2,j}}{2(n_2 - 1)}$$

where, if either $w_{2.1,i}$ or $w_{2.1,j}=1$:

$$w_{2,i,j} = w_{1,i,j}w_{2.1,i}w_{2.1,j}$$

where:

$$w_{1,i,j} = \frac{2n_1 w_{1,i}w_{1,j} - w_{1,i} - w_{1,j}}{2(n_1 - 1)}$$

iv. $q_4[i] = q_2[i] + 1.645*\sqrt{q_3[i]}$

v. $q_5[i] = q_2[i] - 1.645*\sqrt{q_3[i]}$

Multiple observations with one y value create multiple records in the above analysis for one distinct value of y. The last record for that y contains all the information needed for $\widehat{N}_a(y)$. Therefore, at this stage of the analysis, eliminate all but the last record for those y values that have multiple records.

d. Output of interest

From the last entry of the row of q-vectors just computed:

i. q_1 = largest value of y (or smallest if analysis is descending)
ii. $q_2 = \widehat{N}_a$
iii. $q_3 = \text{var}(\widehat{N}_a)$
iv. Standard error of $\widehat{N}_a = \sqrt{q_3}$

From the q column vectors:

i. q_1 represents the ordered vector of distinct values of y
ii. q_2 represents the estimated distribution function, $\widehat{N}_a(y)$, corresponding to the values of y
iii. q_4 represents the 95% one-sided upper confidence bound of the distribution function, $\widehat{N}_a(y)$
iv. q_5 represents the 95% one-sided lower confidence bound of the distribution function, $\widehat{N}_a(y)$

Case 5– Estimation of $N_a(y)$: Discrete Resource, Variable Probabilities, N_a known and not equal to $\widehat{N}_a$. Confidence Bounds by by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is known and not equal to $\widehat{N}_a$
3. There is a variable probability of selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest. The algorithm for the distribution function for the proportion of numbers, $\widehat{F}_a(y)$, given exactly the same conditions listed above, is presented in Case 8 To compute the distribution function of numbers, $\widehat{N}_a(y)$, first use the algorithm in Case 8 to compute the distribution function with the corresponding confidence bounds for the proportion of numbers Then, compute the following:

$$\widehat{N}_a(y) = \widehat{F}_a(y) * N_a, \tag{52}$$

where N_a is the known subpopulation size. To compute the confidence bounds for $N_a(y)$, simply multiply the upper and lower confidence limits of $\widehat{F}_a(y)$ by N_a.

3.1.2 Proportions of Numbers

A number of algorithms are presented to compute the distribution function for the proportion of numbers. For any case in a resource group, the choice of the algorithm is first determined by the method by which the units were selected. The first two algorithms in this section derive the distribution functions based on an equal probability selection of units and the latter two algorithms are based on an unequal probability selection of units.

Equal Probability of Selection

In this subsection, two examples are provided based on whether or not the subpopulation size is known or unknown. For the first algorithm, N_a can be known or unknown; this algorithm produces confidence bounds based on the binomial distribution. For the second algorithm, N_a is known, this algorithm is based on the hypergeometric distribution

Case 6— Estimation of $F_a(y)$: Discrete Resource, Equal Probabilities, N_a known or unknown. Confidence Bounds by Binomial Algorithm.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, can be known or unknown.
3. There is an equal probability of selection of units from the subpopulation.

Outline for Algorithm

Under the given conditions in which N_a may not be known, the confidence bounds can be based on the binomial distribution In addition, Example 8 provides the normal approximation approach to the confidence bound estimation.

A program, based on the binomial distribution and written in the GAUSS language, is presented in this section. We assume X has the binomial distribution, $X \sim \text{Binomial}[n_a, F_a^*(y)]$, where $n_a(y)$ is the observed realization of X, n_a represents the number of "trials", $F_a^*(y) = \frac{N_a(y)}{N_a}$ represents the true finite population proportion of "successes", and $F_a(y)$ is the infinite population parameter. The estimated distribution function is denoted as $\widehat{F}_a(y) = \frac{n_a(y)}{n_a}$, where n_a is the sample size from the subpopulation and $n_a(y)$ refers to the number of units in the sample for which $y_u \leq y$. The upper confidence bound is computed by obtaining the largest value of $F_a(y)$ for which $\text{Prob}[X \leq n_a(y)] > 0.05$. The lower confidence bound is computed by obtaining the smallest value of $F_a(y)$ for which $\text{Prob}[X \geq n_a(y)] > 0.05$. As written, the algorithm calculates the upper and lower confidence bounds to three decimal places.

Comments in capital letters in braces explain the programming steps. Under these conditions, the upper and lower halves of the confidence bounds are symmetric.

CALCULATION OF CONFIDENCE BOUNDS ON $F_a(y)$ BY THE BINOMIAL DISTRIBUTION

```
load x[a,b] = data;   {LOADS DATA FILE FOR THE TARGET SUBPOPULATION
                       WHICH INCLUDES LABEL CODE AND VARIABLE TO
                       BE ANALYZED. HERE a DESIGNATES THE SAMPLE
                       SIZE, n_a, AND b DESIGNATES THE NUMBER OF
                       COLUMN VECTORS}
n=rows(x);            {SAMPLE SIZE IN TARGET SUBPOPULATION, n_a}
x=sortc(x,2);         {SORTS VARIABLE OF INTEREST}
y=seqa(1,1,nm);       {CREATES SEQUENCE OF NUMBERS}
x2=x[.,2];            {DEFINES VARIABLE OF INTEREST AS X2}
x=y~x2,               {CREATES MATRIX x}

{THE FOLLOWING STEPS COMBINE RECORDS WITH COMMON y-VALUES}
xx=zeros(1,2);
q=0;
i=1;
do while i < n;
 if x[i,2]==x[i+1,2];
    q=q+1;   I;
    else;  xx=xx|x[i,.];
 endif;
 i=i+1;
endo;

xx=xx|x[n,.];
r=rows(xx);
x=xx;

{THE FOLLOWING STEPS FORM DATA MATRIX - x}
r=rows(x);
z=zeros(r,1);
x1=x[.,1];
x2=x[.,2];
x=x2~x1~(x1/n)~z;

{THESE STEPS GENERATE BINOMIAL COMBINATION TERMS}
f=zeros(n+1,1);
i=0;
if n ≤ 160;
 do while i<=n;
    f[i+1,1]=n!/i!/(n-i)!;
```

```
    i=i+1;
 endo,
else;
    f[i+1,1]=lnfact(n) - lnfact(i) - lnfact(n - i);
endif;


{THE FOLLOWING STEPS GENERATE UPPER CONFIDENCE BOUND}
i=1;
do while i <= r; {BEGINS INITIAL DO LOOP}
 rr=x[i,2];
 p=(trunc(100*x[i,3]))/100;
    if p==1.0;
       p=p - .001,
       goto three;
    endif;

one:;
p=p+.01;
j=0,
s=0;
do while j <= rr,
 a=f[j+1,1]*p^j*(1-p)^(n-j);
 s=s+a,
 j=j+1;
endo,
 if s >= .05;
    goto one;
 endif,

two.,
p=p - .001,
j=0;
s=0,
do while j <= rr;
 a=f[j+1,1]*p^j*(1-p)^(n-j);
 s=s+a;
 j=j+1;
endo;
 if s <= .05;
    goto two;
 endif;

three:;
x[i,4]=p+.001;
i=i+1;
ENDO;   {ENDS INITIAL DO LOOP}
```

```
{THE FOLLOWING STEPS ADD AN EXTRA LINE TO DATA MATRIX NEEDED IN
  CONFIDENCE BOUND ADJUSTMENT COMPUTED AT END OF ALGORITHM}
r=rows(x);
y=zeros(r,1);
x=x~y;
y=zeros(1,5);
y[1,2]=n;
y[1,3]=1;
y[1,4]=1;
x=x|y;

{THE FOLLOWING STEPS GENERATE LOWER CONFIDENCE BOUND}
r=rows(x);
i=1;
do while i <= r;    {BEGINS SECOND DO LOOP}
 rr=x[i,2],
 p=(trunc(100*x[i,3]))/100;
    if p==0;
       p=.001;
       goto six;
    endif;

four:;
p=p - .01;
 if p<=0;
    p=.001;
    goto six;
 endif;
j=rr,
s=0,
do while j <= n;
 a=f[j+1,1]*p^j*(1-p)^(n-j);
 s=s+a;
 j=j+1;
endo;
 if s >= .05;
    goto four;
 endif;

five:;
p=p+.001;
j=rr;
s=0;
do while j <= n;
 a=f[j+1,1]*p^j*(1-p)^(n-j);
 s=s+a;
 j=j+1;
endo;
 if s <= .05;
    goto five;
 endif;
```

```
six:;
x[i,5]=p − .001;
i=i+1;
ENDO;   {ENDS SECOND DO LOOP}


{ADJUST F̂a(y) and CONFIDENCE BOUND - AVERAGE SUCCESSIVE VALUES}
r=rows(x);
xx=x;
i=2;
do while i <= r-1;
 xx[i,3:5]=(x[i,3:5] + x[i-1,3:5])/2;
 i=i+1;
endo;

{OUTPUT FILE AND PRINT MATRIX x}
OUTPUT FILE=NAME;
OUTPUT ON;
"x"    "Sequence #"    "F(x)"    "F-upper(x)"    "F-lower(x)" ;
format /m1/rd 12,7;
print x;
OUTPUT OFF,

end;
```

Case 7– Estimation of $F_a(y)$: Discrete Resource, Equal Probabilities, N_a known. Confidence Bounds by Hypergeometric Algorithm.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is known.
3. There is an equal probability of selection of units from the subpopulation.

Outline for Algorithm

Under the given conditions, the confidence bounds can be based either on the binomial or on the hypergeometric distribution. The binomial algorithm presented in Example 6 is appropriate to use given the foregoing conditions. In addition, Example 9 provides the normal approximation approach, which is also applicable, given the foregoing conditions, to the confidence bound estimation.

To obtain confidence bounds for F(y) based on the hypergeometric distribution, refer to the algorithm provided for the confidence bound calculation for $N_a(y)$ in Example 1. Simply divide the lower and upper confidence bounds, and $\widehat{N}_a(y)$, by the known subpopulation size, N_a. No further changes are necessary to this algorithm to provide confidence bounds for $F_a(y)$ based on the hypergeometric distribution.

Variable Probability Selection

In this subsection, two cases are provided to demonstrate variable probability of selection. For both cases, the frame population size can be known or unknown. In Case 8, N_a can be unknown or known but not equal to $\widehat{N}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson ratio standard error and the normal approximation. For Case 9, N_a is known and equal to $\widehat{N}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson variance estimator and the normal approximation.

Case 8– Estimation of $F_a(y)$: Discrete Resource, Variable Probabilities, N_a unknown or known and not equal to $\widehat{N}_a$. Confidence Bounds by Horvitz-Thompson Ratio Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is known and not equal to $\widehat{N}_a$.
3. There is a variable probability of selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest.

Calculation of confidence bounds on $F_a(y)$ by the Horvitz-Thompson formulae

For each indicator, the following algorithm derives the distribution function and the confidence bound for $N_a(y)$ similar to that given in Example 4. In this section, however, the interest is in obtaining a distribution function for proportions. Therefore, the variance of a ratio estimator is used in this algorithm. The confidence bounds are computed based on a normal approximation.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, w_{1i}
   c. Tier 2 conditional weighting factor, $w_{2\,1i}$
   d. Indicator of interest (y)
   e. The subset of data corresponding to the subpopulation of interest, indexed by a.

2. Computation of weighting factors

   This step does not have to be made with each use of the datum, as the weights are permanent attributes of a sampling unit. The following details are

given for completeness.

The Tier 1 and Tier 2 weights are included for each record in the data set. These weights are used to compute the total weight of selecting the i^{th} unit in the Tier 2 sample. Compute the following weight for each record:

$$w_{2i} = w_{1i}w_{2.1i}$$

where w_{1i} is the weighting factor for the i^{th} unit in the Tier 1 sample (the inverse of its Tier 1 inclusion probability) and $w_{2.1i}$ is the inverse of the conditional Tier 2 inclusion probability. The pairwise inclusion weight is defined below. The sample size at Tier 2, n_2, is not subpopulation specific.

3 Algorithm for $\widehat{F}_a(y)$ and Confidence Intervals

a. Sorting of data. The data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of proportions, $\widehat{F}_a(y)$, denotes the proportion of units in the target population that have a value less than or equal to the y for a specific indicator. Conversely, if it is of interest to estimate the proportion of units in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable. The distribution function generated by the analysis in descending order is $[1 - \widehat{F}_a(y)]$.

b. First, compute $\widehat{N}_a = \sum_{S_a} w_{2i}$ (this sums over data matrix).

c. Define a matrix of q column vectors, which will be defined as the following. There is one row for each data record and five statistics for each row.
   q_1 = value of y variable for the record
   $q_2 = \widehat{F}_a(y)$
   $q_3 = \text{var}[\widehat{F}_a(y)]$
   q_4 = upper confidence bound for $\widehat{F}_a(y)$
   q_5 = lower confidence bound for $\widehat{F}_a(y)$

d. Index rows using i from 1 to n; the i^{th} row will contain q-values corresponding to the i^{th} record in the file, as analyzed.

e. Read first observation (first row of data matrix), following with the successive observations, one at a time. Accumulate and store the q -statistics, below, as each observation is read into file. Continue this loop until the end of file is reached.

   i. $q_1[i] = y[i]$

   ii. $q_2[i] = q_2[i-1] + \frac{w_{2i}}{\widehat{N}_a}$

Multiple observations with one y-value creates multiple records in the preceding analysis for one distinct value of y. The last record for that y

contains all the information needed for $\hat{F}_a(y)$. Therefore, at this stage of the analysis, eliminate from the q-file all but the last record for those y values that have multiple records.

f. Entries in the first column (q_1) of the q-matrix identifies the vector of y -values for the remainder of the calculations. For each such y-value, y_i, make the following calculations. Note that this part of the algorithm is not recursive; each calculation is made over the entire sample.

iii. $$q_3[i] = \{\sum_j d_j^2 w_{2j}(w_{2j}-1) + \sum_j \sum_{k \neq j} d_j d_k (w_{2j}w_{2k} - w_{2jk})\}/\hat{N}_a^2$$

where,

$$w_{2jk} = \frac{2n_2 w_{2j} w_{2k} - w_{2j} - w_{2k}}{2(n_2 - 1)}$$

and,

$$d_j = I(y_j \leq y_i) - \hat{F}_a(y_i)$$

Similarly for d_k

iv. $q_4[i] = q_2[i] + 1.645 * \sqrt{q_3[i]}$

v. $q_5[i] = q_2[i] - 1.645 * \sqrt{q_3[i]}$

g. Output of interest
From the q column vectors:

i. q_1 represents the ordered vector of distinct values of y.
ii. q_2 represents the estimated distribution function, $\hat{F}_a(y)$, corresponding to the values of y.
iii. q_4 represents the 95% one-sided upper confidence bound of the distribution function, $\hat{F}_a(y)$.
iv. q_5 represents the 95% one-sided lower confidence bound of the distribution function, $\hat{F}_a(y)$.

Case 9– Estimation of $F_a(y)$: Discrete Resource, Variable Probabilities, N_a known and equal to $\widehat{N}_a$. Confidence Bounds by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size, N, can be known or unknown.
2. The subpopulation size, N_a, is known and equal to $\widehat{N}_a$.
3. There is a variable probability of selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest.

Calculation of confidence bounds on $F_a(y)$ by the Horvitz-Thompson formulae

For each indicator, the following algorithm derives the distribution function and the confidence bounds for $N_a(y)$ exactly as given in Example 4. Because N_a is known and equal to $\widehat{N}_a$, it is not necessary to use the ratio estimator applied in Case 8. The distribution function of $F_a(y)$ is obtained by dividing the distribution function, $\widehat{N}_a(y)$, and the associated bounds, by N_a. (These additional steps are included in this algorithm.) The Horvitz-Thompson variance estimator, discussed in Section 2.1, is used to compute the variance in this algorithm. The confidence bounds are computed based on a normal approximation.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, w_{1i}
   c. Tier 2 conditional weighting factor, $w_{2\,1i}$
   d. Indicator of interest (y)

2. Sorting of data

The data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of proportions, $\widehat{F}_a(y)$, denotes the proportion of units in the target population that have a value less than or equal to the y for a specific indicator. Conversely, if it is of interest to estimate the proportion of units in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable. The distribution function generated by the analysis in descending order is $[1 - \widehat{F}_a(y)]$.

3. Computation of weighting factors

For this step, refer to the program steps given in Example 4 to derive the distribution function and the confidence bound for $N_a(y)$. Follow the steps labeled 3 and 4. Additional steps, shown here, are needed to obtain $\widehat{F}_a(y)$ and its corresponding confidence bounds. Proceed with the following steps after conducting steps 3 and 4 from Example 4:

e. The operations that follow generate the q vectors to compute the estimated distribution function and appropriate confidence bounds for $F_a(y)$. These are denoted by q_6 through q_8. Each element of q_6-q_8 is computed by performing the following operations on the corresponding elements of q_2, q_4, and q_5.

   i. q_6 = Divide each element of q_2 by the known subpopulation size
   ii. q_7 = Divide each element of q_4 by the known subpopulation size
   iii. q_8 = Divide each element of q_5 by the known subpopulation size

From the q vectors:

   i. q_6 represents the estimated distribution function, $\widehat{F}_a(y)$
   ii. q_7 represents the 95% one-sided upper confidence bound of the distribution function, $\widehat{F}_a(y)$.
   iii. q_8 represents the 95% one-sided lower confidence bound of the distribution function, $\widehat{F}_a(y)$.

3.1.3 Rationales for Approaches

Justification for the variance estimators used in the algorithms in Sections 3.1.1 and 3.1.2 was presented in Section 2 of this report. The different choices proposed for confidence bound estimation, under some conditions, were also discussed. For example, both the hypergeometric and binomial approaches to the confidence bound calculation for $F_a(y)$, when N_a is known, were provided in the above cases. Choice of one of the approaches presented to compute confidence bounds for $F_a(y)$, when the subpopulation size is known, depends in part on the available information and in part on the purpose of inference. The bounds based on the hypergeometric distribution provide for inferences directed to the finite population. For example, if data are available for every lake in a small population of lakes, there is no uncertainty relative to this attribute for this population (in the absence of measurement error). Bounds based on hypergeometric or on the normal approximation approach will reduce to zero width as $n \to N$, because of the finite population correction. These bounds are more relevant for management purposes In contrast, those bounds based on the binomial distribution provide for inferences directed to the superpopulation parameter. In this situation, the entire population is considered as a sample from the superpopulation. Statements about the set of high mountain lakes in New England are finite, but general statements about high mountain lakes, based on those found in New England, are relative to a hypothetical, infinite, superpopulation. Therefore, the confidence bounds obtained by the binomial distribution are broader than those provided by the hypergeometric distribution to account for this additional level of variability.

3.1.4 Estimation of Size-Weighted Statistics

A few algorithms are presented to compute the distribution functions for size-weighted totals and size-weighted proportions of totals. The following subsection describes algorithms to compute the distribution function for size-weighted totals. The next subsection presents algorithms to compute the distribution function for the proportions of size-weighted totals.

Estimation of Size-Weighted Totals

In this subsection, two examples are provided based on information that is known or unknown. For the first algorithm, the size-weight, Z_a, is unknown or known and equal to $\widehat{Z}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson standard error and the normal approximation. For the second algorithm, Z_a is known but not equal to $\widehat{Z}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson ratio standard error and the normal approximation.

Case 10– Estimation of $Z_a(y)$: Discrete Resource, Size-Weighted Estimate, Equal or Variable Probabilities. Z_a unknown or known and equal to $\hat{Z}_a$. Confidence Bounds by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size-weighted total, Z, can be known or unknown.
2. The subpopulation size-weighted total, Z_a, is unknown or known and equal to $\hat{Z}_a$.
3. There can be an equal or variable probability selection of units from the subpopulation.

Outline for Algorithm

General formulae for Tier 1 estimates were provided in Section 2.1.1. The general form of a size-weighted estimate in a subpopulation at Tier 1, denoted as $\hat{Z}_a$, is similar to Equation 2 The y_i in that equation refers to the size-weight value, now denoted as z_i:

$$\hat{Z}_a = \sum_{S_a} z_i w_{1i} \quad , \tag{53}$$

where z_i defines a size-weight, such as the area of a lake or the stream length in miles, and w is the inverse of the inclusion probability at Tier 1. Using these same definitions, the variance estimator for $\hat{Z}_a$ is similar to Equation 3a.

Estimation of $Z_a(y)$ by the Horvitz-Thompson formulae

For each indicator, the following algorithm derives the distribution function and the confidence bound for $Z_a(y)$. This algorithm is similar to the algorithm defined for the National Surface Water Surveys (Overton, 1987a,b). The Horvitz-Thompson variance estimator, discussed in Section 2.1, is used to compute the variance in this algorithm. The

confidence bounds are computed based on a normal approximation. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, $w_{1,}$
   c. Tier 2 conditional weighting factor, $w_{2\,1,}$
   d. Size-weighted value (z)
   e. Indicator of interest (y)

2. Sorting of data

The data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of size-weighted totals, $\hat{Z}_a(y)$, denotes the size-weights in the target population that have a value less than or equal to the y for a specific indicator. Conversely, if it is of interest to estimate the size-weight in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable. The distribution function generated by the analysis in descending order is $[\hat{Z}_a - \hat{Z}_a(y)]$.

3. Computation of additional weighting factors

The Tier 1 and Tier 2 weights are included for each observation in the data set. These weights are used to compute the total weight of selecting the i^{th} unit in the Tier 2 sample. First, compute this weight for each observation:

$$w_{2,} = w_{1,}w_{2.1,} \ ,$$

where $w_{1,}$ is the weighting factor for the i^{th} unit in the Tier 1 sample and $w_{2.1,}$ is the inverse of the conditional Tier 2 inclusion probability.

4. Algorithm for $\hat{Z}_a(y)$
   a. Define a matrix of q column vectors, which will be defined as the following. There is one row for each data record and four statistics for each row.
      q_1 = value of y variable for the record
      $q_2 = \hat{Z}_a(y)$
      $q_3 = \text{var}[\hat{Z}_a(y)]$
      q_4 = upper confidence bound for $\hat{Z}_a(y)$
      q_5 = lower confidence bound for $\hat{Z}_a(y)$

   b. Index rows using i from 1 to n; the i^{th} row will contain q-values corresponding to the i^{th} record in the file, as analyzed.

   c. Read first observation (first row of data matrix), following with the successive observations, one at a time. Accumulate the q-statistics as each observation is read into file. Continue this loop until the end of file is reached. At that time, store these vectors and go to d. It is necessary, as shown below for q_4, to identify the records for which $w_{2.1,}=1$.

i. $q_1[i] = y[i]$

ii. $q_2[i] = q_2[i-1] + w_{2,i} * z_i$

iii. $q_3[i] = q_3[i-1] + z_i^2 * w_{2,i} * (w_{2,i} - 1) + 2 \sum_{j<i} z_i z_j (w_{2,i} w_{2,j} - w_{2,ij})$

where, if neither $w_{2.1,i}$ or $w_{2.1,j}=1$:

$$w_{2,ij} = \frac{2n_2 w_{2,i} w_{2,j} - w_{2,i} - w_{2,j}}{2(n_2 - 1)}$$

where, if either $w_{2.1,i}$ or $w_{2.1,j}=1$:

$$w_{2,ij} = w_{1,ij} w_{2\ 1,i} w_{2.1,j}$$

and where:

$$w_{1,ij} = \frac{2n_1 w_{1,i} w_{1,j} - w_{1,i} - w_{1,j}}{2(n_1 - 1)}$$

iv. $q_4[i] = q_2[i] + 1.645 * \sqrt{q_3[i]}$

v. $q_5[i] = q_2[i] - 1.645 * \sqrt{q_3[i]}$

Multiple observations with one y value create multiple records in the preceding analysis for one distinct value of y. The last record for that y contains all the information needed for $\widehat{Z}_a(y)$. Therefore, at this stage of the analysis, eliminate all but the last record for those y values that have multiple records.

d. Output of interest

From the last entry of the row of q-vectors just computed:

i. q_1 = largest value of y (or smallest if analysis is descending).
ii. $q_2 = \widehat{Z}_a$
iii. $q_3 = \text{var}(\widehat{Z}_a)$
iv. Standard error of $\widehat{Z}_a = \sqrt{q_3}$

From the q column vectors:

i. q_1 represents the ordered vector of distinct values of y
ii. q_2 represents the estimated distribution function, $\widehat{Z}_a(y)$, corresponding to the values of y.
iii. q_4 represents the 95% one-sided upper confidence bound of the distribution function, $\widehat{Z}_a(y)$.
iv. q_5 represents the 95% one-sided lower confidence bound of the distribution function, $\widehat{Z}_a(y)$.

Case 11– Estimation of $Z_a(y)$: Discrete Resource, Size-Weighted Estimate, Equal or Variable Probabilities. Z_a known and not equal to $\widehat{Z}_a$. Confidence Bounds by Horvitz-Thompson Ratio Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size-weighted total, Z, can be known or unknown.
2. The subpopulation size-weighted total, Z_a, is known and not equal to $\widehat{Z}_a$.
3. There can be an equal or variable probability selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest. The algorithm for the distribution function for the proportion of numbers, $\widehat{G}_a(y)$, given exactly the same conditions listed here, is presented in Case 12. To compute the distribution function of size-weighted totals, $\widehat{Z}_a(y)$, first use the algorithm in Case 12 to compute the distribution function with the corresponding confidence bounds for the proportion of size-weighted totals. Then, compute the following:

$$\widehat{Z}_a(y) = \widehat{G}_a(y) * Z_a \,, \tag{54}$$

where Z_a is the known size-weighted total. To compute the confidence bounds for $Z_a(y)$, simply multiply the upper and lower confidence limits of $G_a(y)$ by Z_a.

Estimation of Proportion of Size-Weighted Totals

In this subsection, two examples are provided based on varying conditions. For the first algorithm, the size-weight, Z_a, is unknown or known and not equal to $\widehat{Z}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson ratio standard error and the normal approximation. For the second algorithm, Z_a is known and equal to $\widehat{Z}_a$; this algorithm produces confidence bounds based on the Horvitz-Thompson standard error and the normal approximation.

Case 12– Estimation of $G_a(y)$: Discrete Resource, Size-Weighted Estimate, Equal or Variable Probabilities. Z_a unknown or known and not equal to $\widehat{Z}_a$. Confidence Bounds by Horvitz-Thompson Ratio Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size-weighted total, Z, can be known or unknown.
2. The subpopulation size-weighted total, Z_a, is unknown or known and not equal to $\widehat{Z}_a$.
3. There can be an equal or variable probability selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest. Another discussion of the formulae are presented in the previous section, Estimation of Size-Weighted Totals.

Calculation of confidence bounds on $G_a(y)$ by the Horvitz-Thompson formulae

For each indicator, the following algorithm derives the distribution function and the confidence bound for $Z_a(y)$ similar to that given in Case 10. Because Z_a is unknown or known and not equal to $\widehat{Z}_a$ in this example, however, the variance of a ratio estimator is used in this algorithm. The confidence bounds are based on a normal approximation.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, w_1,
   c. Tier 2 conditional weighting factor, $w_{2.1}$,
   d. Size-weighted value (z)
   e. Indicator of interest (y)
   f. The subset of data corresponding to the subpopulation of interest, indexed by a.

2. Computation of additional weighting factors

This step does not have to be made with each use of the datum, as the weights are permanent attributes of a sampling unit. The following details are given for completeness.

The Tier 1 and Tier 2 weights are included for each observation in the data set. These weights are used to compute the total weight of selecting the i^{th} unit in the Tier 2 sample. First, compute this weight for each observation:

$$w_{2,i} = w_{1,i} w_{2.1,i} ,$$

where $w_{1,i}$ is the weighting factor for the i^{th} unit in the Tier 1 sample and $w_{2.1,i}$ is the inverse of the conditional Tier 2 inclusion probability. The pairwise inclusion weight is defined below. The sample size at Tier 2, n_2, is not subpopulation specific

3. Algorithm for $\hat{G}_a(y)$ and Confidence Intervals

a. Sorting of data. The data file needs to be sorted on the indicator, either in an ascending or descending order When the data file is sorted in ascending order on the indicator, the distribution function of size-weighted proportions, $\hat{G}_a(y)$, denotes the proportion of size-weights in the target population, such as stream miles, that have a value less than or equal to the y for a specific indicator. Conversely, if it is of interest to estimate the proportion of size-weights in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable The distribution function generated by the analysis in descending order is $[1 - \hat{G}_a(y)]$.

b. Compute $\hat{Z}_a = \sum_{S_a} w_{2,i} * z_i$

c. Define a matrix of q column vectors, which will be defined as the following. There is one row for each data record and five statistics for each row.
   q_1 = value of y variable for the record
   $q_2 = \hat{G}_a(y)$
   $q_3 = \text{var}[\hat{G}_a(y)]$
   q_4 = upper confidence bound for $\hat{G}_a(y)$
   q_5 = lower confidence bound for $\hat{G}_a(y)$

d. Index rows using i from 1 to n; the i^{th} row will contain q-values corresponding to the i^{th} record in the file, as analyzed.

e. Read first observation (first row of data matrix), following with the successive observations, one at a time. Accumulate and store the q -statistics as each observation is read into file. Continue this loop until the end of file is reached.

   i. $q_1[i] = y[i]$

   ii. $q_2[i] = q_2[i-1] + \frac{w_{2,i} * z_i}{\hat{Z}_a}$

Multiple observations with one y-value create multiple records in the preceding analysis for one distinct value of y. The last record for that y contains all the information needed for $\widehat{G}_a(y)$. Therefore, at this stage of the analysis, eliminate from the q-file all but the last record for those y values that have multiple records.

f. Entries in the first column (q_1) of the q-matrix identifies the vector of y -values for the remainder of the calculations. For each such y-value, y_i, make the following calculations. Note that this part of the algorithm is not recursive; each calculation is made over the entire sample.

iii. $$q_3[i] = [\sum_j d_j^2 w_{2j}(w_{2j} - 1) + \sum_j \sum_{k \neq j} d_j d_k (w_{2j} w_{2k} - w_{2jk})]/\widehat{Z}_a^2$$

where,

$$w_{2jk} = \frac{2n_2 w_{2j} w_{2k} - w_{2j} - w_{2k}}{2(n_2 - 1)}$$

and,

$$d_j = [I(y_j \leq y_i) - \widehat{G}_a(y_i)] * z_i$$

Similarly for d_k.

iv. $$q_4[i] = q_2[i] + 1.645 * \sqrt{q_3[i]}$$

v $$q_5[i] = q_2[i] - 1.645 * \sqrt{q_3[i]}$$

g. Output of interest
From the q column vectors:

i. q_1 represents the ordered vector of distinct values of y.
ii. q_2 represents the estimated distribution function, $\widehat{G}_a(y)$, corresponding to the values of y
iii. q_4 represents the 95% one-sided upper confidence bound of the distribution function, $\widehat{G}_a(y)$
iv. q_5 represents the 95% one-sided lower confidence bound of the distribution function, $\widehat{G}_a(y)$

Case 13– Estimation of $G_a(y)$: Discrete Resource, Size-Weighted Estimate, Equal or Variable Probabilities. Z_a known and equal to $\widehat{Z}_a$. Confidence Bounds by Horvitz-Thompson Standard Error and Normal Approximation.

Conditions for approach

1. The frame population size-weighted total, Z, can be known or unknown.
2. The subpopulation size-weighted total, Z_a, is known and equal to $\widehat{Z}_a$.
3. There can be an equal or variable probability selection of units from the subpopulation.

Outline for Algorithm

The algorithm supplied in this section is based on the Horvitz-Thompson formulae, which were discussed in Section 2. This algorithm is appropriate for a sample subset for any subpopulation a that is of interest.

Calculation of confidence bounds on $G_a(y)$ by the Horvitz-Thompson formulae

For each indicator, the following algorithm derives the distribution function and the confidence bound for $Z_a(y)$ exactly as given in Case 10. Because Z_a is known and equal to $\widehat{Z}_a$, it is not necessary to use the ratio estimator. The distribution function of $G_a(y)$ is obtained by dividing the distribution function, $\widehat{Z}_a(y)$, and the associated confidence bounds by Z_a. (These additional steps are included in this algorithm.) The Horvitz-Thompson variance estimator, discussed in Section 2.1, is used to compute the variance in this algorithm. The confidence bounds are computed based on a normal approximation.

1. Data set
   a. Unit identification code
   b. Tier 1 weighting factor, w_{1i}
   c. Tier 2 conditional weighting factor, $w_{2\,1i}$
   d. Size-weighted value (z)
   e. Indicator of interest (y)

2. Sorting of data

The data file needs to be sorted on the indicator, either in an ascending or descending order. When the data file is sorted in ascending order on the indicator, the distribution function of size-weighted proportions, $\hat{G}_a(y)$, denotes the proportion of size-weights in the target population, such as lake area, that have a value less than or equal to the y for a specific indicator. Conversely, if it is of interest to estimate the proportion of size-weights in the target population with indicator variables greater than or equal to y, the data file would be sorted in descending order on this variable. The distribution function generated by the analysis in descending order is $[1 - \hat{G}_a(y)]$.

3. Computation of weighting factors

For this step, refer to the program steps given in Case 10 to derive the distribution function and the confidence bound for $Z_a(y)$. Follow the steps labeled 3 and 4. Additional steps, shown here, are needed to obtain $\hat{G}_a(y)$ and its corresponding confidence bounds. Proceed with the following steps after conducting steps 3 and 4 from Case 10.

e. The operations that follow generate the q vectors to compute the estimated distribution function and appropriate confidence bounds for $G_a(y)$. These are denoted by q_6 through q_8. Each element of q_6-q_8 is computed by performing the following operations on the corresponding elements of q_2, q_4, and q_5.

   i. q_6 = Divide each element of q_2 by the known subpopulation size
   ii. q_7 = Divide each element of q_4 by the known subpopulation size
   iii. q_8 = Divide each element of q_5 by the known subpopulation size

From the q vectors:

   i. q_6 represents the estimated distribution function, $\hat{G}_a(y)$
   ii. q_7 represents the 95% one-sided upper confidence bound of the distribution function, $\hat{G}_a(y)$
   iii. q_8 represents the 95% one-sided lower confidence bound of the distribution function, $\hat{G}_a(y)$

3.2 Extensive Resources

A detailed discussion of the formulae for obtaining area and proportion of areal extent for continuous and extensive resources was presented in Section 2.1.2. Formulae were presented for both areal and point samples.

3.2.1 Estimation of Proportion of Areal Extent

As discussed in Section 2.1.2, the confidence bounds for the proportion of areal extent in continuous and extensive resources can be based on the binomial distribution. This algorithm was presented in Section 3.1.2, Case 6, for discrete resources. No changes in this algorithm are needed.

3.2.2 Estimation of Area

Formulae for the estimation of total areal extent of the surveyed resources was proposed in Section 2.1.2. Proposed methods to compute areal extent for point and areal samples are discussed in the following subsections.

Point Samples

Formulae for the estimation of areal extent based on point sample was presented in Section 2.1.2.2. To obtain confidence bounds for $A_a(y)$ based on the binomial distribution, refer to the algorithm provided for the confidence bound calculation for $F_a(y)$ in Section 3.1.2, Case 6. Simply multiply the lower and upper confidence bounds, and $\widehat{F}_a(y)$, by the known area or estimated area of the resource. No further changes are necessary to this algorithm to provide confidence bounds for $A_a(y)$ based on the binomial distribution.

Areal Samples

Formulae for the estimation of areal extent based on areal samples are still under development However, some preliminary formulae are proposed in Section 2.1.2.1. Work in this area is continuing and will be included in the next version of this report.

3.3 Estimation of Quantiles

Overton (1987a) defines the calculations for both the ascending and descending sorted indicators. For the algorithm used in this report, it is not necessary to employ a different definition for percentiles for an ascending or descending analysis, distributions are identical as generated either way. The general algorithm computes the linear interpolation of the distribution function for both types of analyses. In the following equation, let r represent the proportion of the desired percentile. The fraction in this equation can be interpreted as the slope of the line. The coefficient of this fraction interpolates to the value $[Q(r) - a]$. The lower bound, a, is added to this piece, $[Q(r) - a]$, to obtain the quantile of interest.

Assuming an ascending analysis and that the generated distribution function is F(y):

$$Q(r) = a + [r - F(a)] \left\{ \frac{(b-a)}{F(b) - F(a)} \right\} , \tag{55}$$

where F(a) is the greatest value of $F(y) \leq r$ and F(b) is the least value of $F(y) > r$.

For a descending analysis, the distribution function generated was $F^*(y)=[1 - F(y)]$. To obtain the percentiles, calculate $F(y)=1 - F^*(y)$; the foregoing formula is appropriate for the analysis.

SECTION 4

TABLES

Table 1. Reference to Distribution Function Algorithms

A. Distribution Functions for Numbers - Estimation of $N_a(y)$

Equal Probability Selection:

Population Size	Subpopulation Size	Algorithm	Case
Known/unknown	Known	Hypergeometric[1]	1
		HT-NA[2]	3
Known	Unknown	Hypergeometric	2
		HT-NA	3

Variable Probability Selection:

Population Size	Subpopulation Size	Algorithm	Case
Known/unknown	Unknown or known and = $\hat{N}_a$	HT-NA	4
Known/unknown	Known and $\neq \hat{N}_a$	HTR-NA[3]	5

1 Hypergeometric refers to the exact hypergeometric distribution algorithm used to obtain confidence bounds.
2 HT-NA refers to Horvitz-Thompson variance with normal approximation to obtain confidence bounds.
3 HTR-NA refers to Horvitz-Thompson ratio estimator of variance with normal approximation to obtain confidence bounds.
4 Binomial refers to the exact binomial distribution algorithm used to obtain confidence bounds.

Table 1 - Continued.

B. Distribution Functions for Proportions of Numbers - Estimation of $F_a(y)$

Equal Probability Selection:

Population Size	Subpopulation Size	Algorithm	Case
Known/unknown	Known/unknown	Binomial[4]	6
Known/unknown	Known	Hypergeometric	7

Variable Probability Selection:

Population Size	Subpopulation Size	Algorithm	Case
Known/unknown	Unknown or known and $\neq \hat{N}_a$	HTR-NA	8
Known/unknown	Known and $= \hat{N}_a$	HT-NA	9

1 Hypergeometric refers to the exact hypergeometric distribution algorithm used to obtain confidence bounds.
2 HT-NA refers to Horvitz-Thompson variance with normal approximation to obtain confidence bounds.
3 HTR-NA refers to Horvitz-Thompson ratio estimator of variance with normal approximation to obtain confidence bounds.
4 Binomial refers to the exact binomial distribution algorithm used to obtain confidence bounds.

Table 1 - Continued.

C. Distribution Functions for Size-Weighted Statistics for Both Equal and Variable Probability Selection

Population Size	Subpopulation Size	Algorithm	Section
Estimation of $Z_a(y)$			
Known/unknown	Unknown or known and = $\hat{Z}_a$	HT-NA	10
Known/unknown	Known and ≠ $\hat{Z}_a$	HTR-NA	11
Estimation of $G_a(y)$			
Known/unknown	Unknown or known and ≠ $\hat{Z}_a$	HTR-NA	12
Known/unknown	Known and = $\hat{Z}_a$	HT-NA	13

1 Hypergeometric refers to the exact hypergeometric distribution algorithm used to obtain confidence bounds.
2 HT-NA refers to Horvitz-Thompson variance with normal approximation to obtain confidence bounds.
3 HTR-NA refers to Horvitz-Thompson ratio estimator of variance with normal approximation to obtain confidence bounds.
4 Binomial refers to the exact binomial distribution algorithm used to obtain confidence bounds.

Table 2. Summary of Notation Used in Formulae and Algorithms

Symbol	Definition
Populations:	
N	Population size
N_a	Subpopulation size
Distribution Functions:	
Discrete Resources:	
$\hat{N}(y)$	Estimated distribution function for total number
$\hat{F}(y)$	Estimated distribution function for proportion of numbers
$\hat{Z}(y)$	Estimated distribution function of size-weighted totals
$\hat{G}(y)$	Estimated distribution function for a size-weighted proportion
Continuous and Extensive Resources:	
$\hat{A}(y)$	Estimated distribution function for areal extent
$\hat{F}(y)$	Estimated distribution function for proportion of areal extent
Inclusion Probabilities:	
π_i	Probability of inclusion of unit i
π_{ij}	Probability that unit i and j are simultaneously included
π_{1i}	Probability of inclusion of unit i at Tier 1
π_{2i}	Probability of inclusion of unit i at Tier 2
$\pi_{2.1i}$	Conditional Tier 2 inclusion probability
Weights:	
w	Inverse of the above inclusion probabilities (Same definitions apply with corresponding subscripts)
Sample Notation:	
n	General notation for sample size
n_1	Sample size at Tier 1
n_2	Sample size at Tier 2
S_1	Sample of units at Tier 1
S_2	Sample of units at Tier 2
(These may be made specific for subpopulations or resources by appending an a or r. For example:)	
n_a	Sample size for subpopulation a
n_{ri}	Sample size for a resource r at grid point i
S_{1r}	Sample of units at Tier 1 for resource r
S_{2r}	Sample of units at Tier 2 for resource r

SECTION 5

REFERENCES

Chambers, R.L., and R. Dunstan. 1986. Estimating distribution functions from survey data. Biometrika, 73, 597-604.

Cochran, W.G. 1977. Sampling Techniques, Third Edition. Wiley, New York.

Cordy, C.B. In press. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Accepted in Probability in Statistics Letters.

Cox, D.R., and D. Oakes. 1984. Analysis of Survival Data. Chapman and Hall, New York.

Hansen, M.H., W.G. Madow, and B.J. Tepping 1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Amer. Stat. Assoc. 78: 776-793.

Hartley, H.O., and J.N.K. Rao. 1962. Sampling with unequal probability and without replacement. The Annals of Mathematical Statistics, 33, 350-374.

Horvitz, D.G., and D.J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Stat. Assoc. 47: 663-685.

Hunsaker, C.T., and D.E. Carpenter, eds. 1990. Environmental Monitoring and Assessment Program: Ecological Indicators. EPA/600/3-90/060. U.S.EPA, Office of Research and Development, Washington, DC.

Kaufman, P.R., A.T. Herlihy, J.W. Elwood, M.E. Mitch, W.S. Overton, M.J. Sale, K.A. Cougan, D.V. Peck, K.H. Reckhow, A.J. Kinney, S.J. Christie, D.D. Brown, C.A. Hagley, and H.I. Jager. 1988 Chemical Characteristics of Streams in the Mid-Atlantic and Southeastern United States. Volume I: Population Descriptions and Physico-Chemical Relationships. EPA/600/3-88/021a. U.S. EPA, Washington, DC.

Landers, D.H., J.M. Eilers, D.F. Brakke, W.S. Overton, P.E. Kellar, M.E. Silverstein, R.D. Schonbrod, R.E. Crowe, R.A. Linthurst, J.M. Omernik, S.A. Teague, and E.P. Meier. 1987. Characteristics of Lakes in the Western United States. Volume I: Population Descriptions and Physico-Chemical Relationships. EPA-600/3-86/054a. U.S. EPA, Washington, DC.

Linthurst, R.A , D.H. Landers, J.M. Eilers, D.R Brakke, W.S.Overton, E.P. Meier, and R.E. Crowe. 1986 Characteristics of Lakes in the Eastern United States, Volume I: Population Descriptions and Physico-Chemical Relationships. EPA-600/4-86/007a. U.S. EPA, Washington, DC

Little, R.J.A., and D.B. Rubin. 1987. Statistical Analysis with Missing Data. Wiley, New York.

Messer, J.J., R.A. Linthurst, and W.S. Overton. 1991. An EPA Program for Monitoring Ecological Status and Trends. Environ. Monit. and Assess. 17, 67-78.

Messer, J.J., C.W. Ariss, R. Baker, S.K. Drouse, K.N. Eshelman, P.R. Kaufmann, R.A. Linthurst, J.M. Omernik, W.S. Overton, M.J. Sale, R.D. Schonbrod, S.M. Stambaugh, and J.R. Tutshall, Jr. 1986. National Surface Water Survey: National Stream Survey, Phase I - Pilot Survey. EPA/600/4-86/026. U.S. EPA, Washington, D.C.

Miller, R.G. 1981. Survival Analysis. Wiley, New York.

Särndal, C-E., B Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Overton, W.S. 1987a. Phase II Analysis Plan, National Lake Survey, Working Draft. Technical Report 115, Department of Statistics, Oregon State University.

Overton, W S. 1987b. A Sampling and Analysis Plan for Streams in the National Surface Water Survey. Technical Report 117, Department of Statistics, Oregon State University.

Overton, W.S. 1989a. Calibration Methodology for the Double Sample Structure of the National Lake Survey Phase II Sample. Technical Report 130, Department of Statistics, Oregon State University.

Overton, W.S. 1989b. Effects of Measurements and Other Extraneous Errors on Estimated Distribution Functions in the National Surface Water Surveys. Technical Report 129, Department of Statistics, Oregon State University.

Overton, W.S., and S.V. Stehman. 1987. An Empirical Investigation of Sampling and Other Errors in National Stream Survey: Analysis of a Replicated Sample of Streams. Technical Report 119, Department of Statistics, Oregon State University.

Overton, W.S., and S.V. Stehman. 1992. The Horvitz-Thompson theorem as a unifying perspective for sampling. Proceedings of the Section on Statistical Education of the American Statistical Association, pp. 182-187.

Overton, W.S , and S V. Stehman. 1993a Properties of designs for sampling continuous spatial resources from a triangular grid. Communications in Statistics - Theory and Methods, 22, 2641-2660.

Overton, W.S., and S.V. Stehman. 1993b. Improvement of Performance of Variable Probability Sampling Strategies Through Application of the Population Space and the Fascimile Population Bootstrap. Technical Report 148, Department of Statistics, Oregon State University

Overton, W.S., D. White, and D.L. Stevens. 1990 Design Report for EMAP, Environmental Monitoring and Assessment Program. EPA/600/3-91/053. U.S EPA, Washington, DC.

Overton, J M., T C. Young, and W.S. Overton. 1993. Using found data to augment a probability sample. procedure and case study. Environ. Monitoring and Assessment, 26, 65-83.

Porter, P.S., R C. Ward, and H F. Bell. 1988. The detection limit. Environ. Sci. Technol., 22, 856-861.

Rao, J.N.K., J.G. Kovar, and H.J. Mantel. 1990. On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, 77, 365-375.

Särndal, C.E., B. Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.

Smith, T M.F 1976. The foundations of survey sampling: a review. J. of Roy. Stat. Soc , A, 139, Part 2, 183-195

Stehman, S.V., and W.S. Overton. 1989. Pairwise Inclusion Probability Formulas in Random-order, Variable Probability, Systematic Sampling. Technical Report 131, Department of Statistics, Oregon State University.

Stehman, S.V., and W.S Overton. In press. Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling Jour. of Amer. Stat. Assoc

Stevens, D.L. In press. Implementation of a National Monitoring Program. Jour. of Envir. Management.

Thomas, Dave. Oregon State University, Statistics Department, Corvallis, OR.

Wolter, K.M. 1985. Introduction to Variance Estimation, New York: Springer-Verlag.

SECTION 6

GLOSSARY OF COMMONLY USED TERMS

Continuous attribute: an attribute that is represented as a continuous surface over some region. Examples are certain attributes of large bodies of water, such as chemical variables of estuaries or lakes.

Discrete resource: resources consisting of discrete resource units, such as lakes or stream reaches. Such a resource will be described as a finite population of such units.

Distribution function: a mathematical expression describing a random variable or a population. For real-world finite populations, these distributions are knowable attributes (parameters) of the population, and may be determined exactly by a census, or estimated from a sample. The general form will be the proportion (or other measure, like numbers, length, or area) of the resource having a value of an attribute equal to or less than a particular value Proportions may also be of the different possible measures, like number (frequency distributions), area (areal distributions), length, or volume.

Domain: a frame feature that includes the entire area within which a potential sample might encounter the resource. The domain of any one resource can include other resources.

Extensive resource: resources without natural units. Examples of extensive resources are grasslands or marshes.

40-hex: a term for the landscape description hexagon or areal sampling unit centered on each of the grid points in the EMAP sampling grid. The area of each hexagon is approximately 40 km^2.

Inclusion probability (π_i): the probability of including the i^{th} sampling unit within a sample.

Pairwise inclusion probability (π_{ij}): the probability that both element i and element j are included in the sample.

Population: often used interchangeably with the term universe to designate the total set of entities addressed in a sampling effort. The term population is defined in this report to designate any collection of units of a specific discrete resource, or any subset of a specific extensive resource, about which inferences are desired or made.

Randomized model: a model invoked in analysis, assuming the population units have been randomly arranged prior to sample selection In many cases, this is equivalent to assuming simple random sampling.

Resource: an ecological entity that is identified as a target of sampling, description, and analysis by EMAP. Such an entity will ordinarily be thought of and described as a population. Two resource types, discrete and extensive, recognized in EMAP pose different problems of sampling and representation. EMAP resources are ordinarily treated as strata at Tier 2.

Resource class: a subset of a resource, represented as a subpopulation. For example, two classes of substrate, sand and mud, can be defined in the Chesapeake Bay. Subpopulation estimates require only that the classification be known on the sample.

Stratum: a stratum is a sampling structure that restricts sample randomization/selection to a subset of the frame. Samples from different strata are independent. Inclusion probabilities may or may not differ among strata.

Tier1/Tier2: these terms represent different phases of the EMAP program Relative to the EMAP sample, they refer to the two phases (stages) of the EMAP double sample. The Tier 1 sample is common to all resources and provides for each a sample from which the Tier 2 sample is selected. The Tier 2 sample for any resource is a set of resource units or sites at which field data will be obtained.

Weights: in a probability sample, the sample weights are inverses of the inclusion probabilities; these are always known for a probability sample.