EPA/620/R-96/XXX
Revision 0
May 1996

EMAP Statistical Methods Manual

by

Susan Diaz-Ramos
Anteon Corportion
200 S.W. 35th Street
Corvallis, Oregon 97333

Don L. Stevens, Jr.
Dynamac Corporation
200 S.W. 35th Street
Corvallis, Oregon 97333

Anthony R. Olsen
NHEERL Western Ecology Division
U.S. Environmental Protection Agency
200 S.W. 35th Street
Corvallis, OR 97333

May 1996

Environmental Monitoring and Assessment Program
National Health and Environmental Effects Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency
Corvallis, OR 97333

ABSTRACT

The Statistical Methods Manual documents statistical analysis methods applicable to data collected by the Environmental Monitoring and Assessment Program (EMAP). The methods described give procedures to estimate the current status of ecological resources that are appropriate for survey designs implemented by EMAP. The methods apply to analyses of EMAP regional demonstration studies and R-EMAP studies. Sufficient information is given to enable a user to determine if the method is appropriate for the survey design used in these studies. Additional methods will be added as appropriate to include updated analyses procedures or to cover additional EMAP or R-EMAP studies. The audience for the manual are statisticians or scientists with a reasonable background in statistics. The calculations are detailed so that a scientific computer programmer can implement the methods.

Key Words: survey design, cumulative distribution estimation, status estimation, ecological monitoring, U.S. EPA-EMAP.

Preferred citation:

Diaz-Ramos, S., D.L. Stevens, Jr., and A.R. Olsen. 1996. EMAP Statistics Methods Manual. EPA/620/R-96/XXX. Corvallis, OR: U.S. Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory.

ACKNOWLEDGEMENTS

We could not have prepared this methods manual without the cooperation and aid of many individuals. Over the past five years, we have benefitted from technical discussions with all participants in the EMAP Design and Statistics research effort. Kathleen Purdy, a graduate student at Oregon State University, wrote two of the methods on deconvolution of measurement error. We thank Danny Kugler for his work on the method that presents simplified algorithms suitable for spreadsheet software. We recognize the work by Doug Heimbuch, Harold Wilson, and John Seibel of Coastal Environmental Services, Inc. and Steve Weisberg and Jon Volstad of Versar, Inc. who wrote two of the general overview papers included. Similarly, the report by Virginia Lesser and Scott Overton was central to our effort and is included for completeness.

Notice

The research described in this document has been funded by the U.S. Environmental Protection Agency. This document has been prepared at EPA National Health and Environmental Effects Research Laboratory, Western Ecology Division, in Corvallis, Oregon, through Contract #68-C4-0019. It has been subjected to the Agency's peer and administrative review and approved for publication. Mention of trade names or commercial products does not constitute endorsement for use.

CONTENTS

CONTENTS

INTRODUCTION

The Statistical Methods Manual documents statistical analysis methods applicable to data collected by the Environmental Monitoring and Assessment Program (EMAP). A primary use of the EMAP data is to estimate the current status of ecological resource characteristics using scientifically sound procedures. The methods described give procedures to estimate current status applicable to survey designs implemented by EMAP. A distinct feature of EMAP is the use of survey designs as the foundation for site selection and subsequent scientific inference to an ecological resource target population. Consequently, it is essential that the appropriate statistical analysis method be linked with the survey design used for the collection of the data.

The audience for the manual are statisticians or scientists with a reasonable background in statistics. The methods were written with sufficient detail so that a scientific computer programmer can implement the calculations; for this reason, the methods contain more simplified notation than that used in this introduction. The appendices A and B are intended for those with little statistical training who may become involved in the analysis of R-EMAP studies. See appendix C for more information on the general theoretical development upon which the algorithms in this manual are based.

The methods in the manual are appropriate to use for analyses of EMAP regional demonstration studies and R-EMAP studies. The methods give sufficient information to enable a user to determine if the method is appropriate for the survey design used in these studies. Most methods reference one or more EMAP or R-EMAP studies for which the method is appropriate.

Most of the methods in the document provide estimators for the cumulative distribution or its variance for a variety of survey designs and conditions. Method 13 provides simplified estimation algorithms for those using spreadsheet software. These estimates are to be used for internal research only and not intended for use in any internal or external documents. Methods 9 and 11 address the case when substantial measurement error is present in the data (observations). In this case, the estimator of the cumulative function is biased. The bias may be substantial and is most prevalent in the tails of the distribution. These two methods present techniques to adjust for this bias. The following table provides a quick summary of the methods.

STATISTIC	RESOURCE	ESTIMATOR	METHOD #
CDF for proportion	Discrete Extensive	Horvitz-Thompson	1
Variance of the CDF for proportion	Discrete	Horvitz-Thompson Yates-Grundy Simple Random Sample	5 12 13
	Extensive	Horvitz-Thompson Yates-Grundy Simple Random Sample	10 12 13
CDF for total number	Discrete	Horvitz-Thompson	2
Variance of the CDF for total number	Discrete	Horvitz-Thompson Simple Random Sample	6 13
Size-weighted CDF for proportion	Discrete	Horvitz-Thompson	3
Variance of the size-weighted CDF for proportion	Discrete	Horvitz-Thompson Simple Random Sample	7 13
Size-weighted CDF for total	Discrete	Horvitz-Thompson	4
Variance of the size-weighted CDF for total	Discrete	Horvitz-Thompson Simple Random Sample	8 13
CDF for proportion or total in the presence of measurement error; Variance	Discrete	Parametric Jackknife; Horvitz-Thompson	9
CDF for proportion or total in the presence of measurement error; Variance	Discrete	Simulation-Extrapolation (SIMEX); SIMEX Variance	11

We highly recommend that any analysis of EMAP regional demonstration study data or R-EMAP study data be preceded by a thorough reading of reports that document the

survey design, field measurement protocols, and indicator descriptions. This information is available in EMAP reports and should be reviewed.

BACKGROUND

The Environmental Monitoring and Assessment Program is an interagency, interdisciplinary program that will contribute to decisions on environmental protection and management by integrating research, monitoring, and assessment. EMAP's strategies use rigorous science while taking into account social values and policy-relevant questions. It was initiated by EPA's Office of Research and Development to monitor status and trends in the condition of ecological resources, to develop innovative methods for anticipating emerging environmental problems, and in general, to provide a greater capacity for assessing and monitoring the condition of the nation's ecological resources (Messer et al. 1991).

EMAP was designed to provide information that will enable policy-makers, decision-makers and the public to:

- Estimate the current status, trends, and changes in selected indicators of the Nation's ecological resources on a regional basis with known confidence.
- Estimate the geographic coverage and extent of the Nation's ecological resources with known confidence.
- Seek associations between selected indicators of natural and anthropogenic stresses and indicators of condition of ecological resources.
- Provide annual statistical summaries and periodic assessments of the Nation's ecological resources.

A general overview of EMAP in mostly non-technical language is in the *EMAP Program Guide* (Thornton et. al., 1993). Additional information on the assessment framework used by EMAP as a common approach for planning and conducting a wide variety of ecological assessments is given by EPA (1994). The statistical analysis of data from EMAP is best undertaken with an understanding of the measurement selection process. Barber (1994) describes the indicator development strategy used by EMAP in their regional demonstration studies. The statistical methods in this report were intended primarily for these demonstration studies and as well as studies conducted by EPA Regions in conjunction with EMAP.

Survey Design Approach

A distinctive feature of EMAP is strict reliance on probability sampling. Overton et al. (1990) describe the conceptual framework for the sampling-design approach for EMAP. Stevens (1994) gives a description of how the conceptual framework is used in

research-demonstration studies for particular ecological resources. The implementation of the conceptual framework required development of sampling designs directed at environmental resources distributed over space.

Probability sampling is fundamental to EMAP. Probability sampling provides the basis for estimating resource extent and condition, for characterizing trends in extent or condition, and for representing spatial pattern, all with known certainty. A probability sample has some inherent characteristics that distinguish it from other samples: first, the population being sampled is explicitly described; second, every element in the population has the opportunity to be sampled with known probability; and third, the selection is carried out by a process that includes an explicit random element. A probability sample from an explicitly defined resource population is a means to certify that the data collected are free from any selection bias, conscious or not. This probability sample is an essential requirement for a program such as EMAP that aims to describe the condition of our national ecological resources. Further, analytical methods that are as free as possible from the appearance of subjectivity are also required. These two requirements are satisfied in EMAP by adherence to probability-based sampling protocols and analytical methods that rely on the statistical design for their inferential soundness. Thus, EMAP relies on design-based inference procedures for basic estimates of population descriptors. See Hansen et al. (1983), Särndal (1978), or Smith (1976) for discussions of the issues involved in design-based versus model-based inference. These issues are also discussed in a spatial context by de Gruijter and Ter Braak (1990) and Brus and de Gruijter (1993).

Design-based inference relies on the methodology of statistical survey sampling (Cochran 1977) to extend the results from a sample to the population. This extension is valid only with a probability sample. The design specifies what information is to be collected at specified locations; there must also be protocols or methods that are coherent with the design, and that specify how the inference is drawn. The combination of a sample design and an inference protocol is called a sampling plan. This plan includes the prescription of not only what and where to sample, but also how to analyze the resulting data. In many instances in EMAP, the resource groups used novel sampling designs tailored to the resource. These designs are documented in Overton et al. (1990) and Stevens (1994, 1995) and in the various research plans for the particular resources. A general prescription for the analyses is given in Overton et al. (1990), and specific details of the analyses for some designs are in Lesser and Overton (Appendix C). However, these documents do not cover all of the designs that have been applied by the EMAP resource groups. This Methods Manual fulfills the second part of the prescription for a sampling plan by providing detailed descriptions of the methods for analyzing data collected using any of EMAP's sample designs along with computational algorithms where appropriate.

Sampling Ecological Resources

The property of a particular ecological resource that has the most impact on a statistical sampling plan is the dimension of the conceptual representation of the resource in two-dimensional space. An implementation of the sampling strategy may represent (or

model) the resource populations as points, lines, or areas. Resources that are represented as points for sampling purposes are labeled *discrete resources*. A discrete resource — such as small to medium-sized lakes—has distinct, natural units. Such a resource is represented in 2-dimensional space as a point because the objective of the sampling is to describe the resource unit as an entity, even though the resource unit may occupy appreciable area in the landscape. An attribute associated with a unit of a discrete resource, such as pH or an indicator of biodiversity, is viewed as a property of the entire unit. The ensemble of all units of a discrete resource is treated statistically as a finite population. Population inferences for a discrete resource are most appropriately based on numbers of units that possess some property. For example, a statement about lakes in good condition would pertain to numbers of lakes, not, for instance, surface area of lakes. An inference couched in terms of surface area might be possible, but neither the sampling plan nor the measurements taken on the units would be well suited for such an inference.

Resources such as streams, riparian wetlands, or forested shelter belts may be given a 1-dimensional representation in 2-dimensional space, and sampled as *linear resources*. In fact, such resources are 2-dimensional, but their area is very small in proportion to landscape area. These features are much longer than they are wide, and they do not have well-defined natural units. Inferences are appropriately stated in units of length, e.g., proportion of stream-miles in poor condition. Attributes are viewed as being defined at a point rather than being associated with a unit. Thus, a chemical concentration might change continuously along the length of a stream and be defined and measurable at every point on the stream.

Resources that extend over large regions in a more or less continuous and connected fashion are treated as 2-dimensional, or *extensive resources*. Like the linear resource, an extensive resource does not have distinct natural units. Instead, it covers relatively large sections of the landscape and lacks a high degree of functional integration. For example, forests, arid ecosystems, and large wetlands such as salt marshes or the Everglades fall into this category. The domain of an extensive resource has area; it does not consist of a collection of separable points. An attribute of an extensive resource is viewed as a definition of a surface in the sense that it is possible in principle to assign a value to the attribute at every point in the domain. Generally, the attribute surface is reasonably smooth, although there may well be step discontinuities. For example, the domain of a forest could include stands of 50-year-old timber and adjacent newly clear-cut areas. A parameter measuring biomass could show a discontinuity as the boundary is crossed. Population inferences are usually based on area of the resource with some property, e.g., acres of forest with a visual canopy rating indicative of degraded condition.

The distinctions between discrete, linear, and extensive are not always clear, and in some cases a resource may be viewed as both: a resource consisting of isolated fragments may be treated as extensive for sampling but as discrete for analysis, or vice versa. Greater efficiency, that is, lower variance for a fixed sampling effort, will usually result if the sampling and analysis are carried out from the same viewpoint. For example, streams could be sampled as a finite population of stream segments defined by confluences

(discrete), but analyzed in terms of miles of stream channel (linear). Thus, a simple random sample of a stream-segment population results in a variable probability sample of points on streams, and is not the most efficient sample to make an inference about miles of streams.

Linear and extensive resources are sampled somewhat differently but analyzed using similar methods. The important distinction in the analysis is between finite, discrete populations and infinite, continuous populations. Methods for both types of populations are provided in this document.

ESTIMATION AND ANALYSIS

Each resource to be sampled can be represented by a set, R, whose elements index the points where the resource exists. Thus, for a discrete resource, $R = \{s_1, s_2, ..., s_N\}$ where each s_i represents the location of one unit of the resource. If R is, for example, a set of lakes, then each s_i represents the location of one of the lakes in R. For an extensive resource, R is the set of points covered by the resource, for example, the area covered by forest or a linear stream network. If R represents a forest, then each $s \in R$ is a point in the forest; if R is a stream network, each $s \in R$ is a point on some stream in the network. Each attribute of interest of the resource R is a fixed but unknown function defined on R; that is, at each element $s \in R$ there is a fixed value of the attribute denoted as $z(s)$. The population parameter to be estimated is the total of the attribute over R, that is $z_T = \sum_{s_i \in R} z(s_i)$ in the discrete case or $z_T = \int_R z(s)\,ds$ in the continuous case. This is a quite general population parameter, because estimates of mean values, variances, proportions, and distribution functions can all be formulated as estimates of sums or integrals over R. For example, the distribution function $F_z(x)$ for $z(s)$ over R is the proportion of R with value of z less than or equal to x. For a discrete resource, this is

$$F_z(x) = \frac{\sum_{s_i \in R} I_{\{s \mid z(s) \le x\}}(s_i)}{\sum_{s_i \in U} I_R(s_i)} .$$

For an extensive resource, the distribution function is

$$F_z(x) = \frac{\int_R I_{\{s \mid z(s) \le x\}}(s)\,ds}{A_R} ,$$

where

$I_B(x)$ is the indicator function for B defined as $I_B(x) = \begin{cases} 1, & x \in B \\ 0, & \text{otherwise} \end{cases}$.

The methods from finite population sampling can be applied to make inferences about z_T for discrete resources. Finite population sampling methods are extensively developed and well-documented (Cassel *et al.* 1977, Cochran 1977, Kish 1965, Thompson 1992, Yates 1960). However, environmental populations are, in many instances, more appropriately conceptualized as continuous, infinite populations rather than discrete and finite.

Estimates of extent for a resource (e.g., wetlands, forests) or for a subset of a resource (e.g., salt marshes, deciduous forest) can be obtained from classification of a sample. Estimates of ecological condition for a resource class are generated from condition indicators. Cumulative distribution functions with confidence bounds are the fundamental method for describing regional (or national) condition in EMAP. The essential feature of this approach is the emphasis on estimating the cumulative total (or proportion) of a resource class with an indicator of condition (or area) less than or equal to a specified value (e.g., the proportion with indicator value less than or equal to some value of interest). Although distribution functions provide the estimates of condition, the information from them can be presented in several forms (bar graphs, tables, distribution function plots), with the choice of format related to the intended audience.

The primary theoretical justification for the estimation methods presented in this document is the Horvitz-Thompson Theorem (Horvitz and Thompson, 1952) or its continuous population analogue (Cordy, 1994; Stevens, 1995 submitted). The sampling background to Horvitz-Thompson estimation is summarized here very briefly to provide a context for the estimation methods presented in the body of this document. The theory and notation are very similar for discrete and continuous resources.

The inference paradigm is based on the inclusion probabilities and the pairwise inclusion probabilities of the sampled units under the following sampling model: A sample is selected from the universe U by picking the values of n random variables $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ from a joint probability distribution specified by $Pr(s_1, s_2, \ldots, s_N)$, which is defined by the sampling design. (In EMAP, the s_i can be thought of as points, as they will be actual points in an extensive resource or reference points that identify the location of a discrete resource.) The selected points are classified as being in or out of some target population R, and z need be determined only for those points in R. In general, this sampling method gives a fixed total sample size (n), but the size of the achieved sample in R is a random variable. Allowing a random sample size entails some technical complication but provides valuable flexibility. In particular, it provides the ability to make estimates for arbitrary subpopulations, that is, R could be defined after all the sampling has taken place. The only difference between the discrete and extensive case is the form of the probability distribution: in the discrete case, the probability distribution gives the numerical probability that a particular sample is selected; in the continuous case, the probability is replaced with a probability density function for the samples.

For the discrete case, the *inclusion probability* $\pi(k)$ for unit k is the probability that unit k is included in the sample, i.e., $Pr(s_1 = k \text{ or } s_2 = k \text{ or...or } s_n = k)$. For designs such that $Pr(s_i = s_j) = 0$ for all $i \neq j$, (e.g., sampling without replacement)

$$\pi(k) = \sum_{i=1}^{n} \Pr(s_i = k) .$$

The *joint inclusion probability* for units k and l is the probability that units k and l are simultaneously in the sample, and is given by

$$\pi(k, l) = \sum_{i=1}^{n} \sum_{j \neq i}^{n} \Pr(s_i = k, s_j = l) .$$

In cases where R is not finite, but rather an extensive resource, continuous probability distributions are used to specify the sampling design. As a result, the inclusion probability functions used in the discrete case are replaced with inclusion density functions. Let $f(s_1,s_2,...,s_n)$ be the joint probability density function (pdf) of the sample locations, $f_i(s)$ be the (marginal) pdf of s_i, and let $f_{ij}(s, t)$ be the joint pdf of s_i and s_j, $i \neq j$. The inclusion density function is defined by

$$\pi(s) = \sum_{i=1}^{n} f_i(s).$$

The pairwise inclusion density function for $s, t \in U$ is defined by

$$\pi(s, t) = \sum_{i=1}^{n} \sum_{j \neq i}^{n} f_{ij}(s, t).$$

Horvitz and Thompson (1952) provided an estimator of the population total for variable-probability, without-replacement, finite-population sampling design, along with an expression for the variance of the estimated total and a related variance estimator. Alternative expressions for the variance and its estimator were provided by Yates and Grundy (1953) and Sen (1953). (The variance estimators associated with Horvitz and Thompson are given in subsequent equations and are denoted by the subscript "HT"; the Yates-Grundy forms are denoted by the subscript "YG".) As was shown by Cordy (1993), a version of the Horvitz-Thompson theorem holds when sampling from U when the inclusion density and pairwise inclusion density function are defined as above.

The Horvitz-Thompson theorem provides estimators of the total (sum or integral) of z over R and its associated variance in terms of the quantities $z(s_i)$, $\pi(s_i)$, and $\pi(s_i, s_j)$. The form for the estimator of the total is the same for both the discrete and continuous

versions; the only difference between the two is the expression for the variance of the estimator. The (unbiased) estimator of z_T is given by

$$\hat{z}_T = \sum_{i=1}^{n} \frac{I_R(s_i) z(s_i)}{\pi(s_i)} .$$

The estimators of variance of $\hat{z}_T$ for the discrete case are

$$\hat{V}_{HT}(\hat{z}_T) = \sum_{i=1}^{n} \frac{I_R(s_i)(1 - \pi_i) z^2(s_i)}{\pi^2(s_i)} + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\frac{\pi(s_i, s_j) - \pi(s_i)\pi(s_j)}{\pi(s_i, s_j)\pi(s_i)\pi(s_j)} \right] I_R(s_i) I_R(s_j) z(s_i) z(s_j)$$

or

$$\hat{V}_{YG}(\hat{z}_T) = \sum_{i=1}^{n} \sum_{j>i}^{n} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i, s_j)}{\pi(s_i, s_j)} \right] \left[\frac{z(s_i) I_R(s_i)}{\pi(s_i)} - \frac{z(s_j) I_R(s_j)}{\pi(s_j)} \right]^2 .$$

and for the continuous case,

$$\hat{V}_{HT}(\hat{z}_T) = \sum_{i=1}^{n} \frac{I_R(s_i) z^2(s_i)}{\pi^2(s_i)} + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\frac{\pi(s_i, s_j) - \pi(s_i)\pi(s_j)}{\pi(s_i, s_j)\pi(s_i)\pi(s_j)} \right] I_R(s_i) I_R(s_j) z(s_i) z(s_j)$$

or

$$\hat{V}_{YG}(\hat{z}_T) = \sum_{i=1}^{n} \sum_{j>i}^{n} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i, s_j)}{\pi(s_i, s_j)} \right] \left[\frac{z(s_i) I_R(s_i)}{\pi(s_i)} - \frac{z(s_j) I_R(s_j)}{\pi(s_j)} \right]^2 .$$

All of the above estimators of variance are unbiased, provided $\pi(s, t) > 0$ in U.

An estimator of the mean of z can be obtained by dividing $\hat{z}_T$ by the size of R (the number of units in R, or the length or area of R) , i.e., $\hat{\mu}_z = \hat{z}_T / N_R$ or $\hat{\mu}_z = \hat{z}_T / A_R$. The estimator $\hat{z}_T$ will tend to have low variance if z and π are strongly positively correlated. Since many environmental surveys have multiple objectives and collect observations on multiple attributes at the same location, there will often be little or no correlation between

z and π. A ratio estimator (so-called because it is the ratio of two estimators) for μ_z of the form $\hat{\mu}_{z,r} = \hat{z}_T / \hat{A}_R$, where $\hat{A}_R = \sum_{i=1}^{n} [I_R(s_i) / \pi(s_i)]$ estimates A_R, may well be more precise than $\hat{\mu}_z$. The two estimators $\hat{z}_T$ and $\hat{A}_R$ are subject to the same sources of sampling variation, and hence are likely to be positively correlated. Thus, if there is substantial variability in $\hat{A}_R$, $\hat{\mu}_{z,r}$ will likely be more precise than $\hat{\mu}_z$. The ratio estimator of the total is then $\hat{z}_{T,r} = \hat{\mu}_{z,r} A_R$. An approximate variance estimator for $\hat{z}_{T,r}$ is obtained by applying either the Horvitz-Thompson or Yates-Grundy formulas with $d(s_i) = z(s_i) - \hat{\mu}_{z,r}$ in place of $z(s_i)$.

The distribution function $F_z(x)$ of the response z is estimated by applying the Horvitz-Thompson theorem to the indicator function $I_{\{t \in R | z(t) \le x\}}(s)$. An unbiased estimator of the size (number, length, or area) of the subset of R with indicator $z(t) \le x$ is given by

$$\hat{z}^*_T(x) = \sum_{i=1}^{n} \frac{I_{\{t \in R | z(t) \le x\}}(s_i)}{\pi(s_i)} ,$$

so that $\hat{F}_z(x) = \hat{z}^*_T(x) / A_R$ is an unbiased estimator of $F_z(x)$. The ratio estimator $\hat{F}_{z,r}(x) = \hat{z}^*_T(x) / \hat{A}_R$ avoids the possibility of obtaining estimates that exceed 1, and in many cases will be more precise than $\hat{F}_z(x)$, for the same reasons as given for $\hat{\mu}_{z,r}$ relative to $\hat{\mu}_z$. An approximate variance estimator for $\hat{F}_{z,r}$ is obtained by applying the Horvitz-Thompson or Yates-Grundy formulas with $d_i(x) = I_{\{t \in R | z(t) \le x\}}(s_i) - \hat{F}_{z,r}(x)$ in place of $z(s_i)$ and dividing by $\hat{A}^2_R$.

MISSING DATA

All surveys must address the issue of how to handle missing data in statistical estimation. Missing data should always be investigated for patterns, including why it is missing. Two types of missing data are possible in EMAP or R-EMAP surveys. One type is a missing sample unit, such as a missing lake, stream location, or forest site. Sample units may be missing due to inaccessibility, land owner refusal, or other reasons. Finding detectable patterns in missing data could lead to alterations in survey management, including obtaining access permission and identifying situations where the population inference needs to be qualified.

The other type of missing data occurs within a sampling unit, such as a missing observation for an indicator such as a chemical concentration or habitat structure variable. Observations may be missing due to field collection problems, lost samples, laboratory analysis problems, or other reasons. Although it is possible to use different statistical methods to address the two types of missing data, for the purposes of this manual the two types will be treated the same. We associate all missing data as being a missing sample unit.

Two views may be taken. For each view, the missing sample units unavailable for measurement can be considered to be a subset of the target population of interest. One view is to remove this subset from the target population by redefining the target population as the original target population excluding the missing subset. The statistics methods may then be applied as given without adjustments. Another view is to assume the data are missing at random and retain the original definition of the target population. In this case, status estimators of the cumulative distribution expressed as a proportion or fraction of the total remain unbiased estimators. Estimators for population totals or cumulative distributions expressed as amounts (number, length, area) are biased.

REFERENCES

Barber, M. C. ed. 1994. *Environmental Monitoring and Assessment Program: Indicator development strategy*. EPA/620/R-94/022. Athens, GA: U.S. Environmental Protection Agency, Office of Research and Development.

Brus, D. J., and J. J. de Gruijter. 1993. Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science. *Environmetrics* 4:123–152.

Cassel, C. M., C. E. Särndal, and J. H. Wretman. 1977. *Foundations of inference in survey sampling*. New York: John Wiley.

Cochran, W. G. 1977. *Sampling techniques*. 3rd Edition. New York: John Wiley & Sons.

Cordy, C. B. 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters* 18:353–362.

de Gruijter, J. J., and C. J. F. Ter Braak. 1990. Model free estimation from survey samples: A reappraisal of classical sampling theory. *Mathematical Geology* 22:407–415.

Hansen, M. H., W. G. Madow, and B. J. Tepping. 1983. An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association* 78:776–760.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663–685.

Kish, L. 1965. *Survey sampling*. New York: John Wiley & Sons.

Lesser, V. M., and W. S. Overton. 1994. *EMAP status estimation: Statistical procedures and algorithms*. EPA 620/R-94/008. Corvallis, OR: U.S. Environmental Protection Agency, Environmental Research Laboratory.

Messer, J. J., R. A. Linthurst, and W. S. Overton. 1991. An EPA program for monitoring ecological status and trends. *Environmental Monitoring and Assessment* 17:67–78.

Overton, W. S., D. White, and D. L. Stevens Jr. 1990. *Design report for EMAP, Environmental Monitoring and Assessment Program*. EPA 600/3-91/053. Corvallis, OR: U.S. Environmental Protection Agency, Environmental Research Laboratory.

Särndal, C. 1978. Design-based and model-based inference for survey sampling. *Scandinavian Journal of Statistics* 5:27–52.

Sen, A. R. 1953. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 7:119–127.

Smith, T. H. 1976. The foundations of survey sampling: A review. *Journal of the Royal Statistics Society* A.

Stevens, Jr., D. L. 1994. Implementation of a national environmental monitoring program. *Journal of Environmental Management* 42:1–29.

Stevens, Jr., D. L. 1995. A family of designs for sampling continuous spatial populations. *Environmetrics.* Submitted.

Thompson, S. K. 1992. *Sampling*. New York: Wiley.

Thornton, K. W., D. E. Hyatt, and C. B. Chapman, eds. 1993. *Environmental Monitoring and Assessment Program guide*. EPA/620/R-93/012. Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Research and Development.

U.S. Environmental Protection Agency (EPA). 1994. *Environmental Monitoring and Assessment Program assessment framework.* EPA/620/R-94/016. Research Triangle Park, NC: U.S. Environmental Protection Agency, Office of Research and Development.

White, D., A. J. Kimerling, and W. S. Overton. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartography and Geographic Information Systems* 191:5–22.

Yates, F. 1960. *Sampling methods for censuses and surveys*. London: Charles Griffin & Co.

Yates, F., and P. M. Grundy. 1953. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* B15:253–261.