Appendix B. Guidelines for Development of Sampling

B.1 Introduction. Two commonly stated objectives of parametric statistics are to: 1) estimate population parameters; and 2) test hypotheses concerning these parameters (Freese 1974). In HEP, the primary concern is to estimate habitat parameters. How well the estimate approximates the true value depends on the accuracy of the sample data, the sample size, and the natural variability of the sampled parameter. Clearly then, a statistically valid sampling design can increase both accuracy and efficiency in estimating HSI model variables. Carefully designed sampling also permits inferences to be drawn from specific ecological measurements with a specified degree of confidence.

The purposes of this Appendix are to emphasize the need for sound statistical considerations and to discuss basic statistical concepts that apply to sampling design and to estimating HSI model variables. It is beyond the scope of this discussion to present detailed explanations of experimental design or sophisticated statistical procedures. A professional statistician should be consulted early and frequently during a HEP study to obtain assistance with site specific problems associated with sampling HSI model variables.

B.2 Basic statistical concepts. The term population is used in statistics to denote the aggregate from which a sample is chosen. It is the collection, or set, of individuals, objects, or measurements whose properties are to be analyzed. The concept of population is the most fundamental in statistics. The population of interest, or target population, is considered well defined only when its membership is specified. Biologists typically think of a population as a collection of individuals of a particular species. However, a statistical population can be a collection of measurements, such as a set of tree heights or estimates of canopy closure. The population usually referred to in a HEP analysis is the set of all available habitat for each evaluation species in the study area.

Sample refers to a subset of the target population and consists of individuals, objects, or measurements selected from the population by some specified procedure. The purpose of a sample is to accurately estimate population parameters without having to measure the entire population. A parameter is a measurable characteristic that describes the entire population while a statistic is a measurable characteristic of a sample. A statistic is used to estimate the population parameter of interest. Two statistics, the mean and the standard deviation, play an important role in procedures for parameter estimation in HEP. In ecological studies, there are no "typical" objects. Assume, for example, that some characteristic is measured n times from a population. The n measures can be denoted by: $x_1, x_2, x_3, \ldots, x_n$. An average describes the condition about which these measurements fluctuate. The sample mean ($\bar{x}$) is used to mark that central location. The mean is computed by:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \qquad (B1)$$

where x = the observed value of each unit in the sample

n = the number of units in the sample

Variation around, or dispersion of sample units about the mean is characterized by the standard deviation. The standard deviation (s) is calculated by the following formula:

$$s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

where $(x-\bar{x})$ = the deviation of each observed unit value from the sample mean.

The standard deviation is more easily computed by:

$$s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} \qquad (B2)$$

There are three types of variability pertinent to sampling problems in a HEP analysis:

A. Temporal variability. Ecological units are often time-related. Parameters assume different values according to diurnal, seasonal, or annual fluctuations and, thus, vary cyclically over time. Sampling must coincide with the time periods identified in the HSI models. Literature reviews or pilot studies may be required to determine the status of temporal variability and its effects in the study area.

B. Spatial variability. Ecological units can also vary spatially. Parameters assume different values according to changes over area. Cover types are used to reduce spatial variability by identifying areas of land and water which exhibit some degree of homogeneity of physical, chemical, and biological characteristics.

C. Measurement variability. The possibility of error is present whenever measurement occurs. It is useful to distinguish between two different types of errors that may be present in statistical measurements: 1) random errors; and 2) systematic errors. These are strongly related to the concepts of precision and accuracy, respectively.

(1) Random errors. Random errors arise from a large number of uncontrollable factors and can be conveniently described by the term "chance." For example, if repeated samples are drawn from a statistical population, the mean will differ somewhat from sample to sample. These sample means tend to distribute themselves around the "true" population mean. The difference between the sample mean and the "true" parameter mean is referred to as a random error because the complete collection of factors that could explain the difference is unknown. The size of the differences between the sample value and the corresponding population value are indications of reliability or precision. The magnitude of a random error usually decreases as sample size is increased. Larger samples are preferred over smaller ones because the results are generally more reliable (i.e., more precise).

(2) Systematic error or bias. Systematic error is due to problems in the measurement process. These errors may also occur in the planning stage or before or after the collection of data. Causes of systematic errors include faulty sampling design, incorrect use of measurement instruments, and deficiencies in the measurement techniques being employed. The potential presence of systematic error or bias in HEP applications may be illustrated by a simple example. Suppose several stream velocity measurements were made at the same spot in a stream with a current meter. If the meter was not calibrated correctly, there would be a systematic error present in each of the measurements. Note that the observations may have been very precise in the sense that all measurements were very close to that of every other measurement. However, no measurement would have been accurate; accuracy refers to how closely the sample statistic estimates the true parameter value.

(3) Summary. Each individual measurement may be viewed as the sum of three components: 1) the true parameter value; 2) systematic error; and 3) random error. This relationship is stated in equation form below:

$$\text{Individual measurement} = \text{True parameter value} + \text{Systematic error} + \text{Random error} \qquad \text{(B3)}$$

Accuracy of individual measurements is increased when systematic and random errors are minimized. Systematic error is avoided by proper instrument calibration, proper statistical sampling design, and use of proper measurement techniques. Random error can be minimized by increasing the sample size. The remaining sections in this Appendix discuss how statistical sampling procedures aid in increasing accuracy for parameter estimation.

B.3 Considerations in population sampling. The field sampling plan in a HEP analysis should be designed to describe the condition of the habitat for each evaluation species. The objective of sampling is to estimate the important ecological parameters as defined by each HSI model. The following discussion is pertinent to HSI models that require the measurement of parameters from several sample sites within each cover type. Other types of HSI models may require other sampling designs (Chapter 4).

Sampling design commonly has two objectives: 1) definition of the rules or operations used to sample members of the population; and 2) estimation of sample statistics (Kish 1965). The following discussion is limited to the first objective. The second objective is dictated by the specific HSI model used and the assorted parameter estimation techniques and statistics to be used.

Statistical sampling theory plays a minor role in some of the steps in accomplishing this first objective. These steps include the following: 1) the determination of the population of available habitat for each evaluation species; 2) the selection of the types of variables to be measured for each HSI model; 3) the selection of the measurement methods and techniques, including size and type of sample site needed to collect the data; and 4) the planning and organization of field work. Although these topics are not discussed in detail here, their importance should not be ignored. Sampling demands attention to all phases and poor work in one phase may ruin a sample in which everything else is done well.

Application of sampling theory increases efficiency in estimating target population statistics. Sampling theory aids in the development of sample selection and of estimation methods that provide, at the lowest possible cost, estimates that are precise enough for the given purpose.

A. Defining the population of interest. The population of all available habitat should be defined for each evaluation species in the study area. Habitat use can vary greatly between evaluation species. Ideally, a separate sampling strategy should be designed and applied for each evaluation species. Realistically, however, this usually is not feasible. Therefore, the sampling design should be based on the evaluation species identified as most important to the study.

The sample population should be the same as the population of interest. Sometimes, however, the sample population is more restricted than the population of interest. Restricted sampling occurs when all of an area is not available for sampling, and sample sites are selected only from the areas that are available for sampling. Restricted sampling is used, for example, when difficulty of access or ownership excludes the sampling of some portion of the study area. Inferences drawn from a restricted sample only apply to the sampled population and cannot be readily applied to the population of interest.

B. Distribution of sample sites. The variability of ecological phenomena makes it necessary for sampling to be conducted in such a way that each sample statistic approximates its true parameter value. Three sampling designs often employed in HEP applications are: 1) simple random sampling; 2) stratified random sampling; and 3) systematic sampling. Restricted sampling, described earlier, may be applied to any of these three sampling designs.

(1) Simple random sampling. Simple random sampling occurs when all potential sample sites have an equal chance of being sampled. This technique is best employed in homogeneous habitats. The random location of a sample site in an area can be accomplished by several methods, including: 1) randomly selecting points defined by latitude and longitude and designating each point as either a fixed location or the starting location of a transect or plot; 2) superimposing a grid structure over the area and randomly selecting grid cells as sampling locations; or 3) randomly selecting an access point and using a predetermined direction and distance from the access point to locate a plot or transect. A table of random numbers, available in most statistical textbooks, should be used to locate random sites.

(2) Stratified random sampling. Stratified random sampling can be used to divide heterogeneous habitats into subareas, each of which is internally homogeneous. Sample sites are then selected randomly in each subarea. Stratification usually occurs by cover types although further stratification within cover types is often useful.

The principal reasons for stratification are:

(a) Different sets of parameters may need to be sampled in different strata;

(b) Data of a different precision may be desired for certain subdivisions; and

(c) Stratification may produce an increased precision using fewer sample points for certain parameters.

(3) Systematic sampling. Systematic sampling involves the selection of sample sites at fixed intervals throughout an area. This technique is often easier to use without error than the other sample designs, and ensures that samples are more evenly spread across the study area. Systematic sampling is also effectively utilized in multiple sampling of periodic parameters (e.g., stream flow and temperature over time) when using line transects for measurements or when collecting spatial measurements from aerial photographs or cover type maps.

Systematic samples are located so that a sample site occurs at every kth unit, where k represents some interval. For example, transects could be located every fifth mile along a river or a seasonal variable could be sampled weekly.

Systematic sampling can also be employed to distribute sample sites within each subdivision after stratification has occurred.

C. Determination of sample size. The determination of sample size involves three tasks: 1) establishing reliability standards; 2) estimating the mean and standard deviation of habitat parameters important in an HSI determination; and 3) applying the appropriate formula dictated by the selected sampling design.

(1) Reliability standards. When only portions of the population are sampled, reliability standards provide a means of describing the degree to which the sample mean represents the true mean. Two parameters, relative precision (D) and confidence level (C), are used to define the reliability standards. The relative precision determines the magnitude of the difference that can be tolerated between the sample parameter mean and the true parameter mean. A relative precision of 10%, for example, indicates that enough sites must be sampled so that the sample site mean $\bar{x}$ ± 10% of $\bar{x}$ will include the true mean within a specified confidence level. The confidence level determines the probability (expressed as a percent) that the interval determined by D will include the true parameter mean. For example, if just enough sample sites are selected to meet a relative precision of 10% with a confidence level of 80% (0.80), then 8 times out of 10 the sample site mean $\bar{x}$ ± 10% of $\bar{x}$ will bound the true mean. Reasonable reliability standards for most HEP analyses are 25% relative precision and 90% confidence level.

Reliability standards can be set for a parameter, a set of parameters functioning as model inputs, or for the HSI value of individual evaluation species. It is usually more appropriate to base reliability standards on a set of model parameters. However, increasing the accuracy of estimation of various model inputs does not necessarily increase the accuracy of the HSI value derived from the model. Close inspection of the model should alert the user to important parameters whose accurate estimation will significantly increase HSI accuracy. The user then can make decisions about basing reliability standards on the HSI, or on certain parameters.

(2) Estimating the mean and standard deviation. There are three ways of estimating population means and standard deviations for sample size determinations;

(a) Analyze the results of a pilot survey (or presample);

(b) Analyze the results of previous sampling of the same or similar habitat types (usually within close proximity); and

(c) Assume certain facts about the structure of the population, assisted by some mathematical results. For example, sample size for parameters measured in percentages or proportions (such as HSI) can be determined by use of the binomial distribution, which requires only that the mean (x) be estimated.

(3) Application of the appropriate formula. The final task in sample size determination is the application of a formula suitable to the chosen sample design. These formulas are presented below:

(a) Simple random sampling. The formula for determining sample size (n) in sampling for percentages of units in the population which posses some characteristic (e.g., % dbh greater than 12 in) or proportions (e.g., HSI) is:

$$n = \frac{Z_c^2 \cdot p \cdot q}{D^2} \qquad \text{(B4)}$$

where n = the recommended sample size

Z_c = the value obtained from a standardized normal table (Table B-1). C is the specified confidence level. Use the corresponding Z-value of C found in the left column of Table B-1

p = the estimate of the parameter mean expressed in decimal form

q = 1 - p

D = the relative precision

Table B-1. Standard normal table (two-tailed).

Z_c value	Confidence level C
2.576	0.99
1.960	0.95
1.645	0.90
1.440	0.85
1.282	0.80
1.150	0.75
1.036	0.70

The formula for determining sample size (n) in sampling parameters which are not proportions or percentages (e.g., tree height, stream width) is:

$$n = \left(\frac{Z_C \cdot s}{D \cdot \bar{x}}\right)^2 \qquad (B5)$$

where s = the estimated standard deviation

$\bar{x}$ = the estimated mean

(b) Stratified random sampling. Determination of sample size for stratified random sampling designs can be complex. The total sample size n is equal to the sum of all samples in each stratum or

$$n = \sum_{i=1}^{m} n_i \qquad \text{(B6)}$$

where m = number of strata

n_i = sample size for each stratum

Initially, the sample size for each stratum can be determined by the rules and formulas presented for simple random sampling. However, the total sample size (sum of stratum sample sizes) may not fall within budget constraints. In that case, a decision will need to be made about which stratum sample sizes are to be reduced. Some suggested considerations for the final determination of the sample size for each stratum include:

(1) Stratum importance. An attempt should be made to meet the reliability standards in the most important stratum. Importance may be determined by the number or value, or both, of the evaluation species to be sampled in the stratum.

(2) Area size. Sample sites are often allocated in proportion to area size; i.e., larger strata are allocated more sample sites.

(3) Equity. Reductions in sample sites can be shared equally by every stratum. For example, if a particular stratum was originally allocated 10 of 100 total sample sites, but only 80 sites can be used, then the sample sites in that stratum would be reduced to 8, or one-tenth of the adjusted total sample size.

(c) Systematic sampling. The rules for determining sample size using systematic sampling techniques can be very complex. However, a good general rule to follow is to use the simple random sample formula and then distribute the sample sites systematically.

(4) Finite population correction factor. Sampling theory assumes that the population from which the sample is taken is infinitely large; this may not always be the case. In a situation where area plots are used for sample sites (instead of line transects or points), the target population may be finite. Each sample plot should be distinct and nonoverlapping. For example, in a study area of 10 acres, only 100 sample sites with a size equal to 0.1 acre could possibly be sampled. When the population size is finite, the factor $\sqrt{1 - n/N}$ should be used to adjust the estimate of the standard deviation. Here N refers to total population size and n to sample size. The factor $\sqrt{1 - n/N}$ for the standard deviation is called the finite population correction (fpc). This factor remains close to one when the sampling fraction n/N remains low. In practice, the fpc can be ignored whenever the sampling fraction n/N does not exceed 10%. The effect of ignoring the correction is to overestimate the standard deviation. The fpc can be used to reduce the overestimated standard deviation by use of the following formula:

$$s_f = s_o \sqrt{1 - n/N} \qquad \text{(B7)}$$

where s_f = corrected standard deviation

s_o = standard deviation as estimated from the sample

n = sample size used

N = total population size

For example, data on stream width might have been collected from 52 out of 110 possible sites. Suppose the mean width was found to be 20 feet with a standard deviation of 1.0. The estimate of the standard deviation could be adjusted downward by:

$$s_f = 1.0 \sqrt{1 - 52/110}$$

$$s_f = 0.726$$

Thus, the dispersion around the mean, as measured by the standard deviation, can be significantly reduced by the fpc.

The fpc for the standard deviation is primarily used in HEP to estimate the sample size needed in a planned application. In determining sample size, the correction is used to adjust downward the suggested sample size (n); n is first computed by one of the sample size formulas presented in Section C above. Then, the correction is made with the following formula:

$$n_f = \frac{n}{1 + n/N} \qquad \text{(B8)}$$

where n_f = corrected sample size

n = original estimate of sample size

N = total population size

For example, if the calculation of sample size indicated that 100 sites (1 acre plots) were necessary to meet the reliability standards and there were only 110 possible sites (total area = 110 acres), the revised number of sites would be:

$$n_f = \frac{100}{1 + \frac{100}{110}}$$

$$= 52 .$$