maximum-likelihood estimators), is often much more straightforward and flexible. Furthermore, the statistical properties of the maximum-likelihood estimation-style estimators can be superior.

Although maximum-likelihood estimation allows for greater model flexibility, it requires a search algorithm to find a global maximum (overall maximum), unlike generalized least-squares models. For very complex models, only a local maximum may be found, or there may be no convergence. Many statistical packages have built-in procedures for mixed-linear or nonlinear models, allowing for easier application of these relatively new procedures.

Overall Considerations in Designing and Analyzing Forestry Experiments

In order to obtain results that can be interpreted with little or no confounding, experimental units should be carefully selected to remove factors that are not of interest to the experimenter, but would affect the variables of interest. Random allocation of treatments is also needed to equalize the impacts of any remaining factors that were not removed through careful selection. Identifying factors as fixed versus random and using the appropriate design is essential to correct interpretation of results. Also, the correct analysis of hierarchical designs should be stressed; incorrect analyses sometimes appear in literature. For least-squares analysis, expected mean-squares should be calculated to determine appropriate *F*-tests. Power analysis is strongly recommended, during the design of the experiment, to ensure that statistically significant results indicate differences of practical importance.

Because of the large time and spatial scale of many forest processes, experimental units often are large and long-term, in order to have meaningful results. This leads to problems with traditional designs, in that experimental units are large and very heterogeneous, and some are lost over time. Also, there is low power as there are few experimental units. Assumptions of least-squares analysis are commonly not met, resulting in difficulties in analysis and interpretation.

New technologies using maximum-likelihood methods allow greater variability in the analysis of data. These methods have improved our ability to conduct analyses when the assumptions of least-squares analysis are not met, and have increased the flexibility in the design of forestry experiments.

See also: **Afforestation**: Species Choice; Stand Establishment, Treatment and Promotion - European Experience. **Ecology**: Human Influences on Tropical Forest Wildlife. **Environment**: Environmental Impacts. **Experimental Methods and Analysis**: Biometric Research; Statistical Methods (Mathematics and Computers). **Health and Protection**: Diagnosis, Monitoring and Evaluation. **Inventory**: Forest Inventory and Monitoring; Modeling. **Landscape and Planning**: Spatial Information. **Mensuration**: Yield Tables, Forecasting, Modeling and Simulation. **Recreation**: Inventory, Monitoring and Management. **Soil Development and Properties**: Soil Contamination and Amelioration. **Tree Breeding, Practices**: Biological Improvement of Wood Properties. **Wood Formation and Properties**: Wood Quality.

Further Reading

Box GEP and Cox DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society Series B* 26: 211–252.

Cochran WG and Cox GM (1957) *Experimental Designs*. New York: John Wiley.

Cressie NAC (1993) *Statistics for Spatial Data*. revd. edn. Toronto, Canada: John Wiley.

Hurlbert SH (1984) Pseudoreplication and the design of ecological experiments. *Ecological Monographs* 54(2): 187–211.

John JA and Williams ER (1995) *Cyclic and Computer Generated Designs*. London, UK: Chapman & Hall.

Kirk RE (1982) *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, CA: Brooks/Cole.

McCullagh P and Nelder JA (1991) *Generalized Linear Models*. New York: Chapman & Hall.

Meredith MP and Stehman SV (1991) Repeated measures experiments in forestry: focus on analysis of response curves. *Canadian Journal of Forestry Research* 21: 957–965.

Neter J, Kutner MH, Nachtsheim CJ, and Wasserman W (1996) *Applied Linear Statistical Models*, 4th edn. San Francisco, CA: McGraw-Hill.

Schabenberger O and Pierce FJ (2002) *Contemporary Statistical Models for the Plant and Soil Sciences*. New York: CRC Press.

Scheffé H (1959) *The Analysis of Variance*. Toronto, Canada: John Wiley.

Sheskin DJ (1997) *Handbook of Parametric and Nonparametric Statistical Procedures*. New York: CRC Press.

Statistical Methods (Mathematics and Computers)

H T Schreuder, USDA Forest Service, Fort Collins, CO, USA

Inference

Scientific inference becomes statistical inference when the connection between the unknown 'state

of nature' and the observed information is expressed in probabilistic terms.

Statistical inference from sample surveys can be model-based, in which inference relies on a statistical model to describe how the probability structure of the observed data depends on uncontrollable chance variables and frequently on other unknown nuisance variables. Inference can also be design-based, in which reliance is placed on probabilistic sampling. The following is a brief summary of both approaches.

In nonprobabilistic or model-based sampling, inference is made by specifying an underlying superpopulation model ξ for the values of the variable in the actual population being sampled where the actual values are considered random variables from the superpopulation. Then the actual population or a sample from it is considered a sample from this superpopulation of interest. Sample elements do not have to be chosen at random or with known probability as long as they are not selected based on their values of interest $y_i, i = 1, \ldots, N$. Inferences and conclusions rely heavily on the model assumed, which can be a serious liability if the model is not correctly specified. But if correctly specified, an increase in precision can be expected over the design-based approach.

The design-based approach to inference relies heavily on probabilistic sampling, in which each unit and pairs of units in the population have a positive probability of being selected and the probability of each sample can be calculated. The statistical behavior of estimators of a population characteristic is based on these probabilities and the probability-weighted distribution of all possible sample estimates. A weakness of this approach is that samples that were not drawn are considered heavily in evaluating the properties of the inference procedure, yet should not inference about a population parameter be based solely on the actual sample drawn? But the approach is objective and the only assumption made is that observational units are selected at random so the validity of the inference only requires that the targeted and sampled populations are the same. And careful attention to sample selection within the framework of probabilistic sampling will eliminate the least desirable samples from consideration. The idea behind probabilistic sampling is to make the sample representative of the population being sampled.

A crucial difference between design-based and model-based inference is that design-based inferences are made about the finite, usually large, population sampled, whereas model-based sampling inference, although initially restricted to the usually small sample being taken, is generalized to superpopulations by the use of models. Note that there is a distinction between enumerative (or descriptive) and analytical (or comparative) surveys. In enumerative surveys a 100% sample of the existing population provides the complete answer to the questions posed, but is still inconclusive in analytical surveys (see **Figure 1** for informative distinctions between analytical and enumerative surveys). Design-based sampling is widely accepted now and we limit our discussion to it.

Basic Concepts

Why Sample?

Most decisions are made with incomplete knowledge. Your physician may diagnose disease from a single drop of blood, for example. We hope that the drop represents the nonsampled portions of the body. A complete census is rare – a sample is commonplace. A ranger advertises timber sales with estimated volume. Bidders take the truth and reliability of this information at their own risk and judgment.

Sampling will frequently provide the essential information more timely at a far lower cost and can be more reliable than a complete enumeration. There are several reasons why this might be true. With fewer observations to be made and more time available, crews will get less tired and remain more committed to careful measurement of the units in the sample. In addition, a portion of the saving resulting from sampling could be used to buy better instruments and to employ or train higher-caliber personnel. But it is critical that the sample represents the population well!

Populations, Parameters, Estimators, and Estimates

The central notion in any sampling problem is the existence of a population, a collection of units with values of variables of interest attached. The units are selected and the values of interest obtained from the selected elements, either by measurement or observation. Whenever possible, matters will be simplified if the units of the population are the same as those that can be selected for the sample. If we wish to estimate the total weight of earthworms in the top 15 cm of soil for some area, it would be best to think of a population made up of blocks of soil of some specified dimension with the weight of earthworms in the block being the unit value. Such units are easily selected for inclusion in the sample and projection of sample data to the entire population is relatively simple. If we think of individual earthworms as the

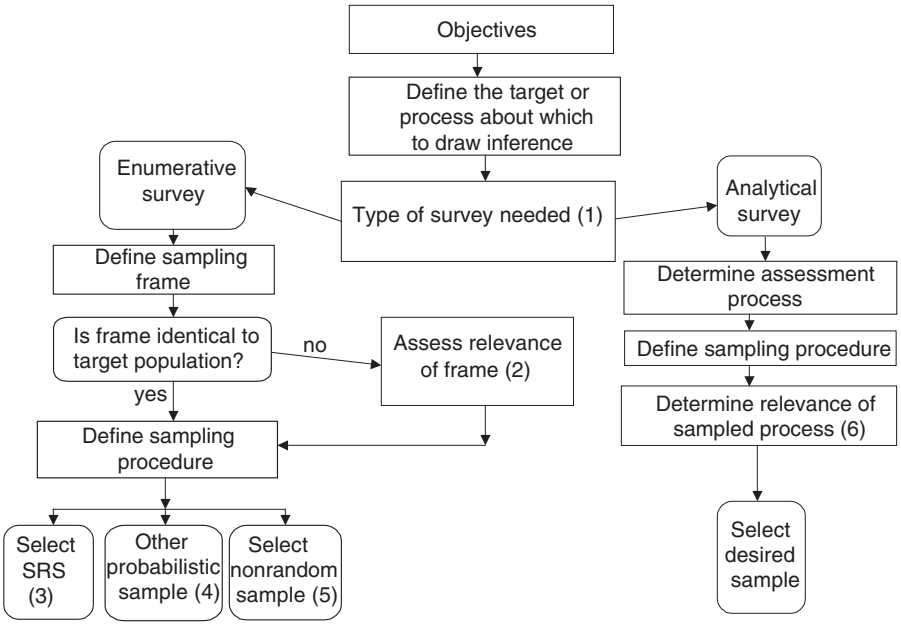

Figure 1 A comparison of enumerative and analytical surveys. The numbers refer to the following:

(1) Are the objectives to draw conclusions about an existing finite population (enumerative survey) or to act on or predict the performance of a (frequently future) process (analytical survey)?
(2) Statistical intervals apply to the frame from which the sample was drawn. Inferences could be biased if the target population is different from the population used for the frame.
(3) Often simple random sampling (SRS) is assumed in constructing confidence intervals.
(4) Confidence intervals can also be constructed for other probabilistic procedures; for example, bootstrapping intervals for the most complex ones.
(5) Statistical confidence intervals are not meaningful here.
(6) Statistical confidence intervals apply to the sampled process and not necessarily to the process or population of interest.

Adapted from Hahn GJ and Meeker WO (1993). Assumptions for statistical inference. *American Statistics* 47: 1–11. Reprinted with permission from *The American Statistician*. Copyright 1993 by the American Statistical Association. All rights reserved.

units, selection of the sample and expansion from the sample to the population may both be very difficult.

To characterize the population as a whole, we often use certain constants of interest called parameters. The proportion or the number of living seedlings in a pine plantation are parameters. Usually the parameters estimated are the population mean or total of one or more variables or change therein over time but we are now often also interested in potential explanations of why interesting changes in parameters happen. Parameter estimates are generated from samples using mathematical formulas called estimators.

Bias, Accuracy, and Precision

A good estimate of a population trait or parameter is one that is close to the true value and obtained from a sample at a reasonable cost. But what happens if the person selecting the sample is prejudiced in some manner in terms of either selecting the sample or making measurements? Either one of these would introduce bias into our estimate.

Statisticians have well-defined expressions for bias, accuracy, and precision. Bias is a systematic distortion. A distinction is made between bias in measurement, in method of selecting the sample, or in estimation of the parameter.

Measurement bias can result, for example, when an observer counts trees on plots and systematically excludes or includes border trees.

Bias due to sampling selection arises when certain units are given a greater or lesser representation in the sample than in the population and this is not compensated for in estimation. If we only sample recreation preferences of visitors to a park on weekends, bias would occur because weekday users had no opportunity to appear in the sample.

The technique of estimating the parameter after the sample has been taken is also a possible source of bias. If the most common recreation preference of users on two national forests is estimated by taking a

simple arithmetic average of the preferences from the two forests, the resulting average may be seriously biased if there is a considerable difference in their size and use.

Selection and measurement biases are rarely acceptable. Estimation bias may be acceptable when some biased estimator is more precise with only slight bias relative to unbiased ones.

A biased estimate may be precise but it is not accurate. Accuracy refers to the success of estimating the true value of the parameter; precision refers to the clustering of sample values about their own average, which, if biased, cannot be the true value. Accuracy, or closeness to the true value, may be absent because of bias, lack of precision, or both (**Figure 2**).

Variables, Continuous and Discrete

Variation is a fact of life. Coping with some of the sampling problems created by variation is an important part of making valid inferences. For example, tree height is a variable.

Continuous variables are those expressed in a numerical scale of measurement, any interval of which may, if desired, be subdivided into an infinite number of values, say amount of time spent recreating. Discrete variables are qualitative or those represented by integral values or ratios of integral values, either attributes such as the proportion of trees having a specific attribute or counts such as number of people in a recreation group.

Continuous and discrete data may require different statistical procedures. Most of the sampling methods and computational procedures discussed are for use with continuous variables and we focus on those. The procedures for discrete variables are generally more complex. Often count variables can be treated as continuous variables, especially for larger sample sizes.

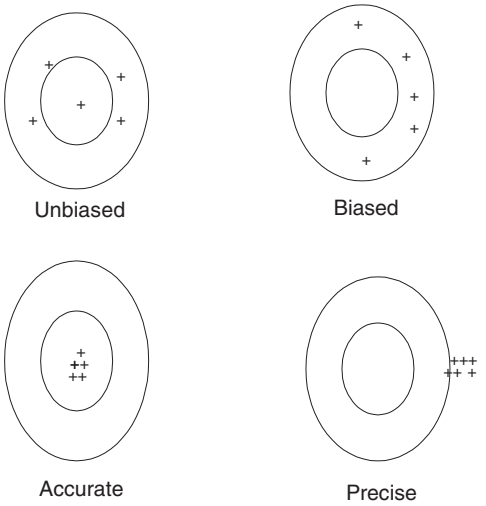

Figure 2 An example of bias, precision, and accuracy if average distance to plot center is used in estimating distance to center of target for five shots.

Distribution Functions

A distribution function shows, for a population, the relative frequency with which different values of a variable occur so that the proportion of units within certain size limits can be determined. Each population has its own distinct distribution function that can often be approximated by certain general types of function, such as the normal, binomial, Poisson, and negative binomial. The bell-shaped normal distribution is often encountered in dealing with continuous variables such as volume per hectare in old-growth stands of timber. The binomial is associated with data where a fixed number of individuals are observed on each unit, characterized by the number of individuals having some particular attribute such as number of seed germinating on a dish. The Poisson distribution may arise where individual units are characterized by a count having no fixed upper limit, particularly if zero or very low counts tend to predominate, such as number of dead trees per hectare. For such data the negative binomial may be useful if low counts do not dominate.

The form of the distribution function dictates the appropriate statistical treatment of a set of data. The exact form of the distribution will seldom be known, but some indications may be obtained from the sample data or from a general familiarity with the population. The methods of dealing with normally distributed data are simplest and fortunately the distribution of means of large samples may be approximated well by this distribution.

Sample estimates are subject to variation just like the individual units in a population. The mean diameter of a stand as estimated from a sample of three trees will frequently be different from the mean estimated from other samples of three trees, and a sample of size 6 would usually produce a more precise estimate than a sample of size 3.

The measure of variation most commonly used is the variance, a measure of the dispersion of individual attribute values about their mean estimated from a sample. Large and small variances indicate wide and little dispersion respectively. The variance of an attribute is a parameter.

The estimator of the variance from a simple random sample is given by:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

where s^2 = sample estimate of the population variance, y_i = the value of the i^{th} unit in the sample, $\bar{y}$ = the arithmetic mean of the sample, i.e., $\bar{y} = \sum_{i=1}^{n} y_i/n$, n = the number of units observed in the sample, and s is the standard deviation, the square root of the variance. Measures of the same form, called the variance of the estimate (s^2/n for simple random sampling (SRS)) and the standard error of estimate (standard error of estimate = square root of the variance of the estimate) are merely the variance and standard deviation among estimates rather than among individual units. Repeated sampling is unnecessary; the variance and the standard error can be obtained from a single set of sample units where the variability of an estimate depends on the sampling method, the sample size, and the variability among the individual units in the population. A sample estimate should be presented with an indication of its reliability as measured by the standard error.

With the standard error, confidence limits can be estimated suggesting how close we might be to the parameter being estimated. For large samples (usually more than 30) the parameter estimated will be on average roughly within 2 standard errors of the estimated parameter (based on an approximation to the normal distribution) unless a 1 in 20 chance occurred (95% confidence limits).

Design

Objectives can be to:

1. Generate current status estimates such as area or volume in a forest, where the forest is, and how it is distributed, and monitor change in such parameters.
2. Identify possible cause-and-effect relationships such as a growth decline in pine forests that could be due to drought or pollution.

Sampling Frame and Representative Sampling

Sampling Frame

Each unit in the population should have a positive, known probability of being selected for the sample so a list of units in the population, called a sampling frame, is required. This frame gives for all N units in the population:

1. The known positive probability of selection, π_i, $i = 1, \ldots, N$ for each unit.
2. The joint positive probability of selection, π_{ij}, i, $j = 1, \ldots, N$, $i \neq j$ for all pairs of units.

In the following we only discuss without replacement sampling of units since it is more efficient than with replacement sampling. Potential sample units can have equal or unequal probabilities and joint probabilities of selection. One of the big advantages of unequal probability sampling is that all single-phase probabilistic procedures used are special cases. Understanding the concept of unequal probability sampling will facilitate comprehension of the other procedures and why and when it is advantageous to use them. This flexibility leads us to the designs discussed: SRS, stratified sampling, cluster sampling, sampling with probability proportional to size (PPS), and systematic sampling with a random start. We then discuss estimation so that we have a sampling strategy consisting of both the sampling design and the estimation procedure used.

Sample Designs

Unequal probability sampling If π_i is the probability of selecting unit i and π_{ij} is the joint probability of selecting units i and j, then the unbiased Horvitz–Thompson estimator of the population total Y is:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} y_i/\pi_i \tag{1}$$

with variance:

$$V(\hat{Y}_{HT}) = 1/2 \sum_{i \neq j}^{N} w_{ij}(y_i/\pi_i - y_j/\pi_j)^2 \tag{2}$$

with $w_{ij} = \pi_i\pi_j - \pi_{ij}$.

Unbiased variance estimators are:

$$v_1(\hat{Y}_{HT}) = 1/2 \left\{ \sum_{i \neq j}^{n} [(\pi_i\pi_j - \pi_{ij})/\pi_{ij}](y_i/\pi_i - y_j/\pi_j)^2 \right\} \tag{3}$$

and

$$v_2(\hat{Y}_{HT}) = \sum_{i=1}^{n} [(1 - \pi_i)/\pi_i^2] y_i^2 + \sum_{i \neq j}^{n} [(\pi_{ij} - \pi_i\pi_j)/\pi_{ij}](y_i y_j/\pi_i\pi_j) \tag{4}$$

Examination of the above equations for understanding If $\pi_i = ky_i$, with k a constant, then $\hat{Y}_{HT}$ in eqn [1] is a constant, and Y and V in eqn [2] would be 0, the ideal situation. This won't happen in practice but we can approximate it. For example, we can practically select trees proportional to their diameter breast height squared if we are interested in tree volume and then the ratios y_i = volume for tree i/x_i = basal area for tree i are essentially

constant, so that $(y_i/\pi_i - y_j/\pi_j)^2$ in eqn [2] is close to 0. Similarly, if we are interested in tree counts, then giving each tree an equal weight in selection is efficient. The efficiency of the methods discussed depends on the strength of the relationship between the variable of interest y and the covariate x used for probability of selection, how the covariate is used in selection, and joint probabilities of units. With this background on the ideas behind the sampling designs, we now list them specifically.

Simple random sampling This is the simplest probabilistic approach. All samples of size n have the same probability of selection from the N units in the population. SRS sampling has the advantages that since all units have the same probabilities of selection, applicable analysis techniques are easy to implement and estimation is straightforward and understandable, for example when estimating the mean or total of a population. The estimator of the mean $\bar{y}$ is:

$$\bar{y} = \sum_{i=1}^{n} y_i/n$$

with sample size n and y_i the value for variable of interest on sample unit i.

An unbiased estimator of the population variance of the mean is:

$$\begin{aligned} v(\bar{y}) &= [(N-n)/(Nn)]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)\right] \\ &= [(N-n)/(Nn)]s^2 \end{aligned}$$

where N = number of elements in population, s^2 is the sample variance and $(N-n)/N$ is called the finite population correction. An estimator of the total Y, $\hat{Y}$, would be obtained by multiplying $\bar{y}$ by N, so $\hat{Y} = N\bar{y}$ and its variance would be $v(\hat{Y}) = N^2 v(\bar{y})$.

In various circumstances we may have complete knowledge on a covariate associated with the variable of interest for which we know all the values in the population or we can get those with relative ease. Usually this information is combined with the information on the variable of interest measured on a subsample of the units in the population. This information can be used in various manners in sample selection and in estimation.

Denoting by y = variable of interest and x = covariate, numerous sample selection schemes and estimators are possible.

Stratified sampling This is a simple but powerful extension of SRS where the population of interest is divided into subpopulations or strata of interest. The idea behind stratification is as follows:

1. We are interested in those subpopulations (strata) too.
2. The subpopulations are internally more homogeneous than the population so we can gain efficiency in estimation by distributing the sample in a good manner over them.
3. We are thus able to apply different sampling procedures in the different subpopulations for convenience.

Estimator of the population mean is:

$$\bar{y}_{st} = \sum_{h=1}^{k} N_h \bar{y}_h / N$$

with estimated variance of the mean:

$$v(\bar{y}_{st}) = \sum_{h=1}^{k} (N_h^2/N^2)[(N_h - n_h)/N_h]s_h^2/n_h$$

where: $\bar{y}_h$ = sample mean for stratum h, k = number of strata, $s_h^2 = \sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2/(n_h - 1)$ and N_h and n_h are number of elements in the population and sample respectively in stratum h.

Cluster sampling In this extension of SRS, clusters of (say) trees are sampled by SRS. The idea behind cluster sampling is twofold:

1. It is useful when no list of sample units is available, as is often true with trees, but lists of clusters are available or easily constructed (e.g., stands or plots respectively).
2. It is usually cheaper to visit clusters of trees than individual trees as in SRS because travel expense is often the biggest item in sampling forests.

Ideally, clusters are very heterogeneous, in contrast to strata, because it is more efficient that way. Usually, reduced cost is the reason behind cluster sampling.

If we select n out of N clusters at random and each cluster sampled is measured completely for the variable of interest, then a biased estimator, $\bar{y}_{cl}$, of the mean per unit is:

$$\bar{y}_{cl} = \sum_{i=1}^{n} M_i \bar{y}_{i.} \Big/ \sum_{i=1}^{n} M_i$$

where M_i is the number of units in cluster i, with an estimator of the variance:

$$v(\bar{y}_{cl}) = [(N-n)/Nn]\sum_{i=1}^{n}(M_i^2/M_n^2)(\bar{y}_{i.} - \bar{y}_{cl})^2/(n-1)$$

with N = number of clusters in the population, n = number of clusters selected by SRS, $M_n = \sum_{i=1}^{M} m_i/n$

is the average number of units per cluster in the sample, and $\bar{y}_{i.} = Y_{i.}/M_i$, where $Y_{i.}$ is the total for all observations in cluster i.

PPS sampling In PPS sampling it is assumed that there is a covariate (or independent variable) available which is positively correlated with the variable of interest and units are selected proportional to the value of the covariate. The information collected on the covariate and on the variable of interest are combined into an estimator such as the Horvitz–Thompson estimator in eqn [1].

PPS sampling is useful when individual selection probabilities are nearly proportional to the variable of interest.

Estimator of the population mean is:

$$\bar{y}_{HT} = \sum_{i=1}^{n} y_i/(N\pi_i) \text{ with estimated variance :}$$

$$v(\bar{y}_{HT}) = (1/2)\sum_{i\neq j}^{n}[(\pi_i\pi_j - \pi_{ij})/(N^2\pi_{ij})](y_i/\pi_i - y_j/\pi_j)^2$$

with n and $N =$ number of elements in the sample and population respectively and all π_i and π_{ij} are assumed to be larger than 0.

Systematic sampling with a random start In systematic sampling with a random start, a random starting unit is selected and then every kth unit is selected. Systematic sampling assumes that the population can be arrayed in some order, which may be natural, say, days of the week in recreation sampling, or artificial, such as numbered plot locations on a map. The ordering may be haphazard in the latter case but needs to be carefully considered in the earlier one. For example, in sampling use of a recreation area we probably do not want to sample every seventh day, say every Sunday. Systematic sampling has not in the past been generally endorsed by theoretical statisticians because it is not a strictly probabilistic procedure in that several units have joint probabilities of selection of 0. But practitioners and applied statisticians have prevailed in getting it used widely because it is a very practical way of collecting information in the field and avoids the problem of poorly distributed samples in the field, as can happen with some of the earlier procedures discussed. Generally, systematic sampling (with a random start) is treated as SRS, the assumption being that the variance estimate for SRS should usually give an overestimate of the variance achieved with systematic sampling.

Estimator of the population mean is:

$$\bar{y}_{syst} = \sum_{i=1}^{n} y_i/n$$

with variance estimator:

$$v(\bar{y}_{syst}) = [(N-n)/N]s^2/n$$

Note that the formulas are the same as for SRS.

Another estimator Although the Horvitz–Thompson estimator is quite efficient in many situations, it can be quite unreliable in some cases. A specific example involves populations where some of the covariate values, x, are quite small relative to the values of the variable of interest, y. It is clear that if some of the sample units contain y and x values where x is quite small, for those ratios in the estimator, y/x can be quite large yielding large estimates. For example, if $x = 0$ for one or more units, the ratio would be ∞. Units with $x = 0$ would not be selected by PPS sampling (causing bias in the estimation) but would be with SRS. Having extreme values is a general problem with such mean of ratio estimators which are generally not recommended to be used at all with SRS.

In general, more complex – but also more robust – estimators such as the very general, efficient generalized regression estimator developed by C. E. Sarndal should be used if possible:

$$\hat{Y}_{gr} = \sum_{i=1}^{n} y_i/\pi_i + a_{gr}\left(N - \sum_{i=1}^{n} 1/\pi_i\right) + b_{gr}\left(X - \sum_{i=1}^{n} x_i/\pi_i\right)$$
$$= \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} e_i/\pi_i$$

where:

$$\hat{y}_i = a_{gr} + b_{gr}x_i, e_i = y_i - \hat{y}_i$$

$$a_{gr} = \left[\sum_{i=1}^{n} y_i/(\pi_i v_i) - b_{gr}\sum_{i=1}^{n} x_i/(\pi_i v_i)\right] \Big/ \sum_{i=1}^{n} 1/(\pi_i v_i)$$

$$V(\hat{Y}_{gr}) = (1/2)\sum_{i\neq j}^{N}(\pi_i\pi_j - \pi_{ij})(e_i/\pi_i - e_j/\pi_j)^2$$

and a variance estimator:

$$v(\hat{Y}_{gr}) = (1/2)\sum_{i\neq j}^{n}[(\pi_i\pi_j - \pi_{ij})/\pi_{ij}](e_i'/\pi_i - e_j'/\pi_j)^2$$

where:

$$e_i = y_i - \tilde{y}_s - b_{gr}(x_i - \tilde{x}_s)$$

and:

$$e_i' = e_i - e_i \left\langle \left\{ \left[(\hat{N} - N) \sum_{l=1}^{n} x_l^2/(\nu_l \pi_l) - (\hat{X} - X) \sum_{l=1}^{n} x_l/(\pi_l \nu_l) \right] \Big/ \nu_i \right\} + \left\{ \left[-(\hat{N} - N) \sum_{l=1}^{n} x_l^2/(\nu_l \pi_l) + (\hat{X} - X) \sum_{l=1}^{n} 1/(\pi_l \nu_l) \right] \right\} (x_i/\nu_i) \Bigg\rangle \times \left\langle 1 \Big/ \left\{ \sum_{l=1}^{n} x_l^2/(\pi_l \nu_l) \sum_{l=1}^{n} 1/(\pi_l \nu_l) - \left[\sum_{l=1}^{n} x_l/(\pi_l \nu_l) \right]^2 \right\} \right\rangle$$

where:

$$\hat{N} = \sum_{l=1}^{n} 1/\pi_l, \tilde{N}_s = \sum_{l=1}^{n} 1/(\pi_l \nu_l),$$
$$\hat{X} = \sum_{l=1}^{n} x_l/\pi_l, \tilde{x}_s = \left\{ \sum_{l=1}^{n} x_l/(\pi_l \nu_l) \right\} \Big/ \tilde{N}_s$$

and:

$$\tilde{y}_s = \sum_{l=1}^{n} y_l/(\pi_l \nu_l)/\tilde{N}_s$$

The generalized regression is biased but consistent in the sense that as $n \to N$, the bias goes to 0.

Variance estimation in general Classical variance estimators discussed above are typically derivable and usually give unbiased or at least consistent estimates of the actual variance. In many cases the actual sampling strategy used is quite complex and such variance estimators cannot be derived. For such situations and even in cases where the actual variances can be derived and computed, other methods can be used, the best-known one being bootstrapping.

Bootstrapping takes full advantage of the computing power now available. It is a computer-based method that allows us to assign measures of precision to statistical estimates. Confidence intervals can be constructed without having to make normal theory assumptions. To illustrate for SRS, if we have a sample of n units of y, with sample mean $\bar{y}$ and variance $v(\bar{y})$, then in bootstrapping we take B samples of n units with replacement from the n sample units. Then, for each of the B samples we compute means $\bar{y}_b$, $b = 1, \ldots, B$ with overall mean $\tilde{\bar{y}}_B = \sum_{b=1}^{B} \bar{y}_b/B$. The variance between these bootstrap estimates is: $v(\bar{y}_B) = \sum_{b=1}^{B} (\bar{y}_b - \tilde{\bar{y}}_B)^2/(B-1)$, which can also be used for $\bar{y}$. In addition, the B sample estimates generate a distribution of estimates that can be used for easy confidence interval construction. There are various ways of bootstrapping.

Multi-Information Sources and Sampling over Time

Often covariates are available or information on them can be more easily and cheaply obtained than for the variable(s) of interest, but not for all units in the population, so more than one sampling phase is required. A voluminous literature is available on this topic and on sampling over time too (see Schreuder *et al.* (1993) in Further Reading, below).

See also: **Experimental Methods and Analysis**: Biometric Research; Design, Performance and Evaluation of Experiments. **Mensuration**: Yield Tables, Forecasting, Modeling and Simulation.

Further Reading

Dawid AP (1983) Inference, statistical: I. In: Kotz S and Johnson NL (eds) *Encyclopedia of Statistical Science*, vol. 4, pp. 89–105. New York: John Wiley.

Deming WE (1975) On probability as a basis for action. *American Statistics* 29: 146–152.

Duncan GJ and Kalton G (1987) Issues of design and analysis of surveys across time. *International Statistical Review* 55: 97–117.

Fraser DAS (1983) Inference, statistical: II. In: Kotz S and Johnson NL (eds) *Encyclopedia of Statistical Science*, vol. 4, pp. 105–114. New York: John Wiley.

Hahn GJ and Meeker WO (1993) Assumptions for statistical inference. *American Statistics* 47: 1–11.

Kruskal WH and Mosteller F (1988) Representative sampling. In: Kotz S and Johnson NL (eds) *Encyclopedia of Statistical Science*, vol. 8, pp. 77–81. New York: John Wiley.

Koch GG and Gillings DB (1983) Inference, design based vs model based. In: Kotz S and Johnson NL (eds) *Encyclopedia of Statistical Science*, vol. 4, pp. 84–88. New York: John Wiley.

Schreuder HT and Gregoire TG (2001) For what applications can probability and non-probability sampling be used? *Environmental Monitoring and Assessment* 66: 281–291.

Schreuder HT and Thomas CE (1991) Establishing cause–effect relationships using forest survey data. *Forestry Science* 37: 1497–1525. (includes discussion).

Schreuder HT, Gregoire TG, and Wood GB (1993) *Sampling Methods for Multiresource Forest Inventory*. New York: John Wiley.

Schwarz CJ and Seber GAF (1999) Estimating animal abundance. Review III. *Statistical Science* 14: 427–456.

Smith TMF (1994) Sample surveys: 1975–1990; an age of reconciliation? *International Statistics Review* 62: 5–34.