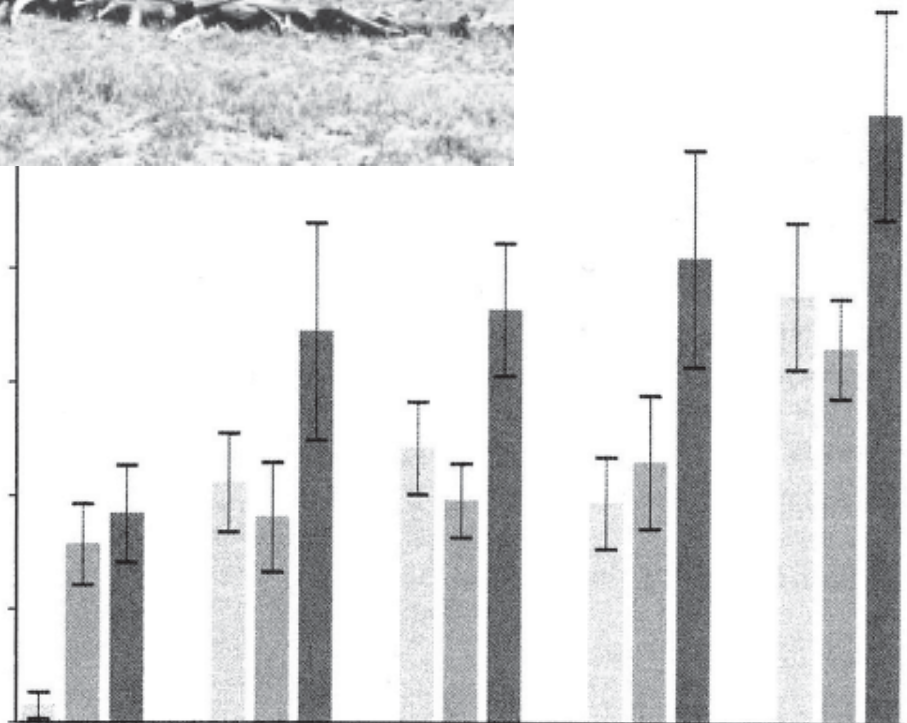# Double Sampling for Stratification: a Forest Inventory Application in the Interior West

David C. Chojnacky

The Author

David C. Chojnacky is a Research Forester in the Bark Beetle Disturbance Ecology Research Work Unit at Rocky Mountain Research Station in Logan, UT. He holds a Ph.D. degree in forest biometry from Colorado State University. His research interests include forest measurements, sampling, wildlife habitat assessment, bark beetle hazard rating, and dryland forests.

Research Summary

The Forest Inventory and Analysis (FIA) group within the Forest Service, U.S. Department of Agriculture, conducts forest resource inventories nationwide with various sample designs. This paper documents the use of double sampling for FIA inventories in the Interior West. Results show that 18 equations describe the entire inventory summarization process for estimating population totals and means, and respective variances. Most equations are for standard use of double sampling, but equations are also given for a ratio method and a special method for data defined by county and land ownership. Inventory data from 800,000 ha of forest land in southern Idaho are used to illustrate application of several equations.

Acknowledgments

Contents

Double Sampling for Stratification: a Forest Inventory Application in the Interior West

David C. Chojnacky

Introduction

Since 1928, the Forest Inventory and Analysis (FIA) group within the Forest Service, U.S. Department of Agriculture (USDA 1992) has inventoried the nation's forests. Fundamental to current FIA inventories is a two-phase design using double sampling for stratification (Cochran 1977).

In FIA's use of double sampling for stratification, a large first-phase sample from aerial photographs (or satellite images) is followed by a smaller second-phase ground sample. Neyman (1938) originally devised the theory to sample human populations. About 25 years later FIA adopted it based on Bickford's (1952, 1959) efforts in the Northeastern United States.

Even though FIA has used this method for more than 30 years, the statistical details have not been published in an accessible reference document. Born and Barnard (1983) documented the computer program FINSYS-2 that uses double sampling to compile FIA inventories, but their notation for the sampling formulas is oriented toward computer programming, unlike notation commonly found in sampling textbooks. Jeyaratnam and others (1984) documented double sampling and extended it to four phases for application to Forest Service inventories, but they did not include all the variations of the double design used by FIA.

This paper (1) briefly overviews the theoretical aspects of double sampling for stratification, and (2) documents the equations used in the FINSYS-2 Fortran program to compile an inventory conducted by the Interior West FIA Program. Eighteen equations are presented to estimate totals, means, and variances for resource attributes. These equations were selected from analysis of a southern Idaho inventory (Chojnacky 1995) by using SAS (1990) software to duplicate FINSYS-2 output.

Notation

L = number of strata in population

h = strata subscript, h = 1,2,...,L

N = total area of population

N_h = total area of stratum h

n' = total number of sample units (points) in first-phase sample

n'_h = number of sample units (points) in first-phase sample for stratum h

n = total number of sample units (ground plots) in second-phase sample

n_h = number of sample units (ground plots) in second-phase sample for stratum h

j = subscript for second-phase sample units, j = 1,2,...,n_h

$W_h = N_h/N$, proportion of population in stratum h

$w_h = n'_h/n'$, proportion of first-phase sample in stratum h

$v_h = n_h/n'_h$, sampling fraction for sample units selected in second-phase sample

Y = population total for attribute of interest, such as volume or area

Y_{hj} = random variable for attribute on ground plot j in stratum h

$\hat{Y}$ = estimate of attribute total

$\hat{R} = \hat{Y}_1/\hat{Y}_2$, ratio of two attribute totals

k = subunit for observing or measuring attribute within plot j, usually a tree or subplot area

y_{hjk} = observed attribute on subunit k within ground plot j in stratum h

π_{hjk} = probability of selection for y_{hjk}

y_{hj} = observed attribute expressed on a per-unit-area basis for ground plot j in stratum h (usually for some subpopulation of interest)

p_{hj} = proportion of plot j in mapped condition of interest, a substitute for y_{hj} when estimating area with the "mapped design"

$\bar{y}_h$ = sample mean of attributes observed in stratum h (estimated stratum mean)

$\bar{y}$ = sample mean of observed attributes

$\bar{y}''$ = sample mean for special case where attribute y_{hj} is for county and/or ownership categories

m'_h = first-phase sample size used in place of n'_h when attribute is observed for county and/or ownership categories

m_h = second-phase sample size used in place of n_h when attribute is observed for county and/or ownership categories

c_h = number of county and/or ownership categories in stratum h

Background Theory

Classical double sampling for stratification (Cochran 1977) efficiently estimates a mean $(\bar{Y})$ from a population of N units by grouping them into L strata with N_h units per stratum $(h = 1,...,L)$. In the first phase, a large simple random sample of size n' is selected from the population of N units. The number of first phase sample units falling in stratum h is n'_h. This phase estimates stratum weights, which are respectively defined for population and sample as:

$W_h = N_h/N$, the proportion of population units falling in stratum h, and

$w_h = n'_h/n'$, the proportion of first-phase sample units falling in stratum h.

Because the first-phase sample is a random selection, w_h is an unbiased estimate of the unknown population stratum weight, W_h.

In the second phase, a much smaller stratified random sample of size n is selected. For each stratum h, a subsample of size n_h is drawn from n'_h first-phase sample units. Here $n_h = v_h n'_h$, where $0 < v_h \le 1$, and v_h is chosen in advance. The purpose of this phase is observation (or measurement) of random variables (Y_{hj}) made on sample units (j) within strata (h). The observations are used to estimate stratum means, $\bar{y}_h$.

In putting both phases together, the mean for the entire sample $\left(\bar{y} = \sum\limits_{h=1}^{L} w_h \bar{y}_h\right)$ is an unbiased estimate of the population mean $\left(\overline{Y} = \sum\limits_{h=1}^{L} W_h \overline{Y}_h\right)$.

In summary, classical double sampling for stratification is a method that uses two random samples, where the second sample is a stratified subsample of the first sample (Cochran 1977). The first-phase sample is for estimating proportions of the population found in various strata in order to calculate stratum weights. The second-phase sample is for making observations of random variables in the sample units.

Double Sampling in FIA Inventories

In the Interior West, FIA first used stratified sampling (Cochran 1977, p. 89) by relying on forest type maps to calculate known stratum weights (W_h) (Born, personal communication). Because of the limited availability of usable resource maps and newly available double sampling theory for forestry, a shift was made, around 1965, to aerial-photo sample points for estimating stratum weights. This initiated FIA's application of double sampling for stratification that is still in use today.

Currently, the inventory population is theoretically defined by FIA in the Interior West as a finite grid of N points, where each point represents a 0.4 ha (1 acre) sample unit. When the inventory uses English units, N is also equal to the total area (acres) of the population. For metric units, it is still convenient to view N as the area (hectares) of the population, but there are some complications in computing finite population corrections for variances that will be discussed later. This definition of N sidesteps a precise description of a fixed-area sample unit by focusing on points instead of sample units.

The points are sampled by observing or measuring data attributes within a "support region" around each point (Campbell 1994; Moisen and others 1995). Historically, FIA in the Interior West has used 0.4 ha (1 acre) as the size of the support region for most observations.

Actual sampling uses a 1 km grid laid over the population of interest to select sample points in a systematic fashion. If such a grid starts in a random fashion, the application of double sampling theory to the systematic data still yields unbiased estimates. FIA in the Interior West uses the 1 km grid that is conveniently labeled on all 7 ½ minute U.S. Geologic Survey maps, so a random start is not possible. However, it seems reasonable to consider the 1 km grid random with respect to the vast landscapes sampled. Variance estimation for a grid sample may also use the "random sample" formulas if there is no strong correlation between the grid and the population (Husch and others 1982, p. 182; Schreuder and others 1993, p. 54).

The first-phase sample n' is a 1 km grid of points corresponding to 1 km UTM (easting and northing) coordinate intersections. First the grid is located on maps at regular 1 km intervals and then is visually transferred to aerial photographs. Observations taken at sample points include identification of county, land ownership, and four land type categories—timberland, woodland, nonforest, and water (USDA Forest Service 1989). These observations provide data to estimate strata weights ($w_h = n'_h/n'$).

Each sample point in the first-phase sample represents 100 ha, so it is easy to compare the total number of phase-one points to known areas for some ownerships within a State. Therefore, small adjustments are made to the stratum weights (w_h) when total areas estimated from double sampling do not match those available from the U.S. Bureau of Census and other public agencies.

The second-phase sample of n_h units is subset from the first phase by using a sampling fraction (v_h) generally equal to 1/25, which is done with a 5 km grid. Variables of interest are observed on second-phase ground plots (sample locations) in a "support region" of about 0.4 ha (1 acre) around each 5 km grid point. A detailed discussion of the plot design is given in Moisen and others (1995).

Two types of variables are observed in the second-phase sample: (1) resource attributes (Y_{hj}) and (2) other variables used to define subpopulations of interest or "domains of study" (Chojnacky 1996; Cochran 1977, p. 34). Although estimation of population totals is important, it is often more important to estimate subpopulation totals (Williams 1995). Therefore, FIA in the Interior West collects a number of variables for defining subpopulations including forest type, ownership, site class, habitat type, species, crown descriptions, disease ratings, and many other resource classifications. Generally, variables defining subpopulation are categorical and variables defining attributes are continuous.

Attribute variables are the focus of the estimation process; the subpopulation variables merely subset the attribute totals into meaningful categories. An attribute is a random variable (Y_{hj}) observed on ground plot j within stratum h (where y_{hj} is the observed value of Y_{hj}). Examples of attributes include volume, growth, forest area, numbers of trees, and canopy cover. Attributes can be direct measurements as tree basal area or require complex models based upon functions of other attributes.

With current notation, it is not beneficial to expressly define "subpopulation variables" with an additional subscript or by other means. Instead it is more useful to view a "subpopulation variable" as a combination of an attribute and an indicator variable:

$$Y_{hj} = \begin{cases} Y_{hj}, & \text{if attribute is in subpopulation of interest} \\ 0, & \text{otherwise.} \end{cases}$$

Estimation of forest area for a subpopulation is a special binomial case:

$$Y_{hj} = \begin{cases} 1, & \text{if plot } j \text{ represents the subpopulation of interest} \\ 0, & \text{otherwise.} \end{cases}$$

Estimation Equations

The first step in the double sampling process is expression of observed attributes (y_{hj}) on a per-unit-area basis by using selection probabilities to summarize data within plots:

$$y_{hj} = \sum_k \frac{y_{hjk}}{\pi_{hjk}} \tag{1}$$

where

y_{hjk} = attribute observed for some k subdivision of second-phase plot j in stratum h

k = subunit observed or measured for attributes within a plot j, usually a tree or subplot area

π_{hjk} = probability of selection for attribute y_{hjk}.

The probability of selection is a constant that depends upon the plot design used in the second-phase sample. For example, it is 0.1 for tree attributes (such as volume or basal area) when using 0.1 ha fixed-area plots. Also, the probability of selection (π_{hjk}) can be adjusted to obtain attribute data for a portion of a plot. Suppose, for example, a 0.1 ha field plot straddled a subpopulation boundary and the attribute of interest (y_{hjk}) was desired for only half of the plot, then π_{hjk} would be 0.05 instead of 0.1 for the portion of interest. In other words, numerous plot-layout designs can be accommodated by equation 1 as long as the data of interest are expressed on a per-unit-area basis.

For area estimation, $y_{hj} = 1$ if the plot represents the subpopulation of interest, and is 0 otherwise. If area is desired for a portion of plot that straddles a boundary, then y_{hj} is modified to a proportion (p_{hj}) of the plot. This is discussed in more detail later (see eq. 18).

With attribute observations y_{hj} appropriately expressed, a stratum mean is:

$$\bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj} \tag{2}$$

where

$\bar{y}_h =$ sample mean of y_{hj} in stratum h (stratum mean)

$n_h =$ number of second-phase plots in stratum h.

In classical sampling, stratum means are often estimators of interest, but in double sampling for stratification they must be summed over all strata and generally expanded to totals before being interpreted (Jeyaratnam and others 1984).

Averaging stratum means ($\bar{y}_h$) over all strata with weights (w_h) from the phase-one sample gives a sample mean:

$$\bar{y} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{3}$$

where

$w_h = n'_h/n'$

$n' =$ total number of first-phase points in the sample

$n'_h =$ number of first-phase points sampled in stratum h

$L =$ number of strata in population.

For some situations in the Interior West, FIA estimates the sample mean ($\bar{y}_h$) with known stratum weights (W_h) instead of using photo points. This practice simplifies the double sampling design to stratified random sampling (Cochran 1977, p. 89). Some call this situation "double sampling with known stratum weights." Examples include mapping strata as polygons of known area on National Forest lands, and applications where classified satellite imagery define strata. Here the sample mean is:

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{4}$$

where

$W_h = N_h/N$

$N_h =$ known area of stratum h

$N =$ known area of population.

Even though $\bar{y}$ estimates the classical population mean, it generally has little biological meaning for most subpopulations of interest because of the way it is calculated. In equation 2, y_{hj} is zero for plots (j) that do not have observed attributes for the subpopulation of interest while n_h is constant regardless of how many y_{hj} equal zero. The resulting $\bar{y}$ is a per-unit-area mean for the entire population sampled. It does not correspond simply to the area for the subpopulation of interest. However, a useful attribute total for the subpopulation results when $\bar{y}$ is multiplied by the total area (N):

$$\hat{Y} = N\bar{y} \tag{5}$$

where

$\hat{Y} =$ estimate of attribute total (usually for subpopulation of interest)

$N =$ known total area of population.

The estimation of useful means for the double sample design requires a ratio estimator (Cochran 1977, p. 153):

$$\hat{R} = \frac{\hat{Y}_1}{\hat{Y}_2} \tag{6}$$

For example, total volume ($\hat{Y}_1$) for a subpopulation divided by total area of that subpopulation ($\hat{Y}_2$) yields mean volume per unit area for the subpopulation.

In situations where subpopulations are defined by land ownership or county classifications (the main stratification criteria), FIA in the Interior West utilizes known (a priori) county and ownership area information within a given stratum h. Unlike all other subpopulations, estimation of sample means for county and ownership subpopulations ($\bar{y}''$) use a stratum weight different from w_h:

$$\bar{y}'' = \sum_{h=1}^{L} w_h'' y_h \tag{7}$$

where

$w_h'' = m_h'/n'$

$m_h' =$ first-phase sample size in stratum h for county and/or owner subpopulation.

This equation simply partitions the first-phase points (n_h') within a stratum into c_h groups to proportionately weight attribute data for county and ownership categories $\left(\text{note: } n_h' = \sum_{c_h} m_h'\right)$. Ideally, it would be desirable to place all county and ownership categories into separate strata, but budget constraints prohibit the necessary larger sample sizes.

Attribute totals for county and owner categories are obtained from equation 5 by replacing $\bar{y}$ with $\bar{y}''$.

Variance Equations

Variance computations start with attributes (y_{hj}) summed on a per-unit-area basis for each plot (j). Traditionally in extensive forest inventories, variation is not considered for subplot or tree components within plots nor is model error considered for attributes calculated from auxiliary models.

Computation of the sample variance for attribute means ($\bar{y}$) requires two steps. The first step is calculation of the sample variance for a stratum mean ($\bar{y}_h$) (Cochran 1977, eq. 5.11):

$$v(\bar{y}_h) = \sum_{j=1}^{n_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h(n_h - 1)} \tag{8}$$

In the second step, combining $v(\bar{y}_h)$ and first-phase sample information gives a sample variance for the attribute mean (Bickford and others 1963, formula 2):

$$v(\bar{y}) = \frac{1}{n'(n'-1)} \sum_{h=1}^{L} \left[n_h'(n_h' - 1)v(\bar{y}_h) + n_h'\bar{y}_h^2 - n_h'\bar{y}^2 \right] \tag{9}$$

This equation differs slightly in format from previous documentation (Born and Barnard 1983, p. 127) by writing it for more convenient summing of all terms over L at the same time $\left(\text{note that } n'\bar{y}^2 = \sum\limits_{h=1}^{L} n_h'\bar{y}^2\right)$.

In situations where the stratum weights (W_h) are known (as in eq. 4), Interior West FIA uses the sample variance from stratified random sampling (Cochran 1977, eq. 5.13) in place of equation 9:

$$v(\bar{y}) = \sum_{i=1}^{L} W_h^2 v(\bar{y}_h) \tag{10}$$

Whether stratum weights are known (eq. 10) or estimated with first-phase points (eq. 9), the sample variance of the attribute total $\hat{Y}$ (eq. 5) for the subpopulation of interest is:

$$v(\hat{Y}) = N^2\, v(\bar{y}) \tag{11}$$

This equation applies for either English or metric units because N corresponds to total area of the population.

Because attribute means for county and ownership subpopulations are estimated differently ($\bar{y}''$ in eq. 7), the sample variance for these means differs from equation 8. To account for a smaller sample size, FIA in the Interior West replaces n_h with smaller m_h to estimate the sample variance for the strata means:

$$v(\bar{y}_h'') = \sum_{j=1}^{n_h} \frac{(y_{hj} - \bar{y}_h)^2}{m_h(n_h - 1)} \tag{12}$$

This procedure increases the variance for strata means in proportion to the actual number of second-phase field plots (m_h) in county and owner subpopulations. In case there are no field plots for a given county and/or owner category, m_h is set to one to avoid division by zero.

The sample variance for $\bar{y}''$ is a copy of equation 9 plus a substitution of $v(\bar{y}_h'')$ for $v(\bar{y}_h)$. The variance for the attribute total is equation 11.

In the past, Interior West FIA has not computed a variance for the ratio of two population totals (eq. 6), but a variance for this ratio ($\hat{R}$) can be approximated by assuming the two totals ($\hat{Y}_1, \hat{Y}_2$) are random variables (Mood and others 1974, p. 181; Scott and Bechtold 1995, eq. 24):

$$v(\hat{R}) \approx \hat{R}^2 \left[\frac{v(\hat{Y}_1)}{\hat{Y}_1^2} + \frac{v(\hat{Y}_2)}{\hat{Y}_2^2} - \frac{2\,cov(\hat{Y}_1, \hat{Y}_2)}{\hat{Y}_1 \hat{Y}_2} \right] \tag{13}$$

The covariance of $\hat{Y}_1$ and $\hat{Y}_2$ is obtained by substituting covariance terms in place of variance terms into equation 9:

$$cov(\hat{Y}_1, \hat{Y}_2) = \frac{N^2}{n'(n'-1)} \sum_{h=1}^{L} \left[n_h'(n_h' - 1)cov(\bar{y}_{1h}, \bar{y}_{2h}) + n_h' \bar{y}_{1h} \bar{y}_{2h} - n_h' \bar{y}_1 \bar{y}_2 \right] \tag{14}$$

From appropriate substitution into equation 8, the covariance of $\bar{y}_{1h}$ and $\bar{y}_{2h}$ is:

$$cov(\bar{y}_{1h}, \bar{y}_{2h}) = \sum_{j=1}^{n_h} \frac{(y_{1hj} - \bar{y}_{1h})(y_{2hj} - \bar{y}_{2h})}{n_h(n_h - 1)} \tag{15}$$

There are other variance approximations (besides equation 13) for ratio estimators (Schreuder and others 1993, p. 88), but these do not seem applicable for large scale FIA inventories at this time. Jackknife estimation looks most promising, but an efficient jackknife computer program is needed to test this.

Often it is convenient to express variation in terms of confidence intervals. For most surveys, the sample size is large enough so that a normal distribution can be assumed for computing confidence intervals (Cochran 1977, p. 20):

$$CI = \hat{X} \pm z_\alpha \sqrt{v(\hat{X})} \tag{16}$$

where

CI = confidence interval

$\hat{X} = \hat{Y}$ or $\hat{R}$

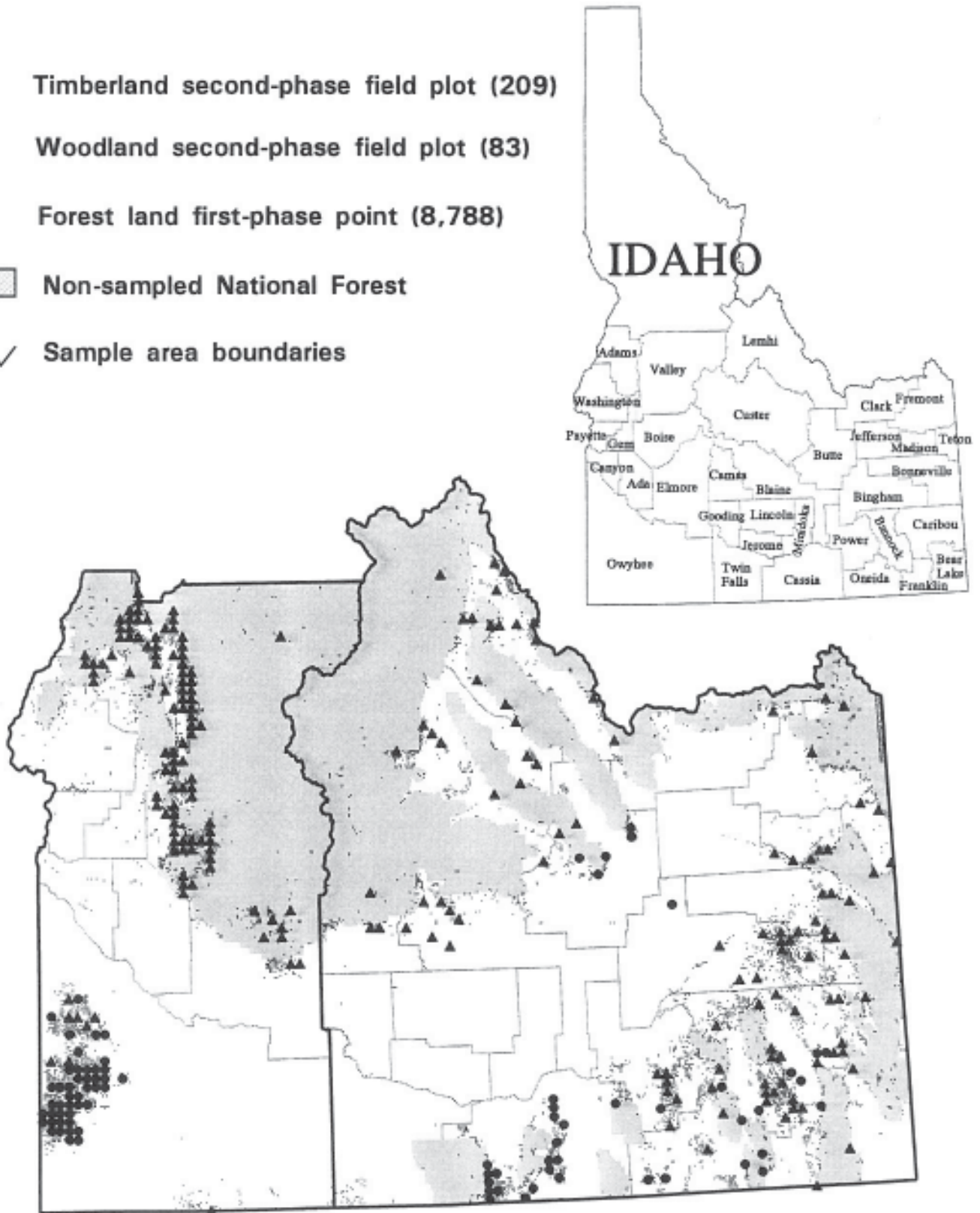z_α = 1.96 for a 95 percent CI, which corresponds to 0.975 probability for the normal distribution.

Discussion

In a general sense, application of the double sampling design by FIA in the Interior West follows three different strategies depending upon the type of attribute or subpopulation that is estimated. There are (1) standard use of double sampling, (2) a ratio method, and (3) a special method for county and owner subpopulations. There is also a small adjustment to the phase-one sample size n' that impacts all estimation methods.

To facilitate discussion, I use data from southern Idaho's recent inventory (Chojnacky 1995) to illustrate some equation results. These data include 4,251 second-phase plots that were subset from 108,029 first-phase points (5 km grid subset from a 1 km grid). Forest land was classified on 8,788 first-phase photo points, and 292 second-phase ground plots were identified as either timberland or woodland (fig. 1). Data were grouped into 27 strata identified by county, ownership, and land type (table 1). The inventory estimated more than 800,000 ha of forest land in southern Idaho.

Figure 1—Points selected in southern Idaho by using double sampling for stratification for forest land outside National Forests. A 1 km grid was used for the first phase (photo points) and a 5 km grid was used for the second phase (field plots).

Table 1—Description and sample sizes of 27 strata used to sample lands outside National Forests in southern Idaho, 1991.

		Sample sizes			
		First phase (n'_h)		Second phase (n_h)	
County[a]	Strata description	Adjusted[b]	Actual	Forested	Total
Adm,W	Public timberland	123.96	132	8	8
Adm,W	Private timberland	192.36	192	5	8
Adm	Industry timberland	206.75	201	6	6
Ge,V	Public timberland	300.77	247	8	8
Ge,V	Private timberland	192.41	209	6	8
V	Industry timberland	364.93	355	22	22
Ada,Boi,Can,Pa	Public timberland	294.80	307	10	11
Ada,Boi,Can,Pa	Private timberland	296.20	307	10	11
Boi	Industry timberland	122.74	115	7	7
E,Ow	Public timberland	331.70	342	14	15
E,Ow	Private timberland	145.63	131	5	8
Ada,Adm,Boi,Can,Pa,V,W	Woodland	232.29	235	3	8
Ow	Woodland	1,088.18	1,090	39	40
All SA 2 counties	Nonforest	34,235.23	34,406	5	1,361
All SA 2 counties	Noncensus water	34.66	34	0	3
Le	Public timberland	301.44	304	9	11
Le	Private timberland	73.71	72	2	5
Bl,Bu,Cam,Cas,Cu,Go,Jer,Li,Mi,TF	Public timberland	465.31	454	14	18
Bl,Bu,Cam,Cas,Cu,Go,Jer,Li,Mi,TF	Private timberland	158.48	164	5	5
Bi,Bon,Cl,Fre,Jef,Ma,Te	Public timberland	302.74	305	10	14
Bi,Bon,Cl,Fre,Jef,Ma,Te	Private timberland	549.39	563	11	15
Ba,BL,Car,Fra,On,Po	Public timberland	573.85	590	17	20
Ba,BL,Car,Fra,On,Po	Private timberland	565.52	562	15	17
All SA 3 counties	Public woodland	1,181.04	1,141	21	26
All SA 3 counties	Private woodland	736.43	770	17	22
All SA 3 counties	Nonforest	64,620.74	64,725	23	2,567
All SA 3 counties	Noncensus water	75.57	76	0	7
Total		107,766.82	108,029	292	4,251

[a]Counties were sampled in two sample areas (SA). SA 2 included Ada, Adams (Adm), Boise (Boi), Canyon (Can), Elmore (E), Gem (G), Owyhee (Ow), Payette (Pa), Valley (V), and Washington (W). SA 3 included Bannock (Ba), Bear Lake (BL), Bingham (Bi), Blaine (Bl), Bonneville (Bon), Butte (Bu), Camas (Cam), Caribou (Car), Cassia (Cas), Clark (Cl), Franklin (Fra), Fremont (Fre), Gooding (Go), Jefferson (Jef), Jerome (Jer), Lemhi (Le), Lincoln (Li), Madison (Ma), Minidoka (Mi), Oneida (On), Power (Po), Teton (Te), and Twin Falls (TF).
[b]Interior West FIA adjusts (weights) first-phase sample sizes within strata (n'_h) to coincide with known land areas.

Standard Double Sampling

Typically an attribute (y_{hj}), for a subpopulation of interest, is calculated for each second-phase plot on a per-unit-area basis (eq. 1). Equations 2, 3, and 5 estimate a subpopulation total $(\hat{Y})$ for the attribute. Because double sampling for stratification is primarily used to estimate attribute totals (Jeyaratnam and others 1984), combination of equations 2, 3, and 5 into a single equation may be desirable to avoid confusion over interpretation of stratum means:

$$\hat{Y} = \frac{N}{n'} \sum_{h=1}^{L} \frac{n'_h}{n_h} \sum_{j=1}^{n_h} y_{hj}$$

(17)

For example, Li and others (1992) suggest an improved double sampling estimator for FIA inventories that replaces equations 2 and 3, but this "improvement" has no effect when summed over all strata to estimate attribute totals (eq. 5).

Variance estimation for the double sampling estimator is well known (Cochran 1977, eq. 12.25), but different approximation formulas exist. The differences depend on whether the second-phase sample is a subset or independent sample of the first phase, and if the equations include a finite population correction ($fpc = 1 - \frac{n}{N}$). For example, equation 9 assumes large sample sizes (completely ignoring all *fpc* terms), and it assumes the second phase sample (n_h) is not necessarily a subset of the primary sample for n_h' (Born and Barnard 1983, p. 126). On the other hand, Rao (1973, eq. 4) (also in Jeyaratnam and others 1984, eq. 1.8) and Cochran (1977, eq. 12.24') give variance formulas where n_h is subset sample from n_h'.

Direct analytical comparison of the all variance approximations is difficult to interpret, but numerical comparison with realistic data offers some insight. For example, the southern Idaho data (table 2) shows little difference among alternative variance equations because over 97 percent of the sampling variation is due to a term common to all formulas—that is the summation of the variance for strata means $v(\bar{y}_h)$ (which is roughly equation 10). Cochran's (1977) equation 12.25 differs by less than 0.1 percent, and the alternative approximations differ by less than 1 percent. Because the differences are so small there seems little value in suggesting replacement of equation 9 even though it differs from recent work.

The numerical comparison also illustrates a small inconsistency between metric and English units for the definition of N, but it seems of little practical significance. All variance equations, except equation 9, require N defined as the "number of sample units" for computing the finite population correction factor ($fpc = 1 - \frac{n}{N}$). Because FIA in the Interior West does not divide its populations into contiguous sample units of known area, definition of N in terms of "sample units" is ambiguous. In general, this ambiguity has no effect on double sampling formulas except for *fpc* calculation, where N is about 2.5 times larger for English units than for metric units. However, the impact of this *fpc* discrepancy was less than 0.02 percent for a test with southern Idaho data (table 2). Because equation 9 includes no *fpc* term, it is consistent for both metric and English units (table 2).

Confidence intervals (eq. 16) for attribute totals ($\hat{Y}$) require equations 8, 9, and 11 for appropriate variance estimates. These are helpful for interpreting data in terms of sampling variation. For example, gross volumes of 9 to 26 million m^3 for several forest types (subpopulations) show 95 percent confidence intervals ranging from 25 to 40 percent of the totals (fig. 2).

Standard double sampling is quite flexible for different types of field data collection procedures. For example, a recent modification in FIA's field data collection procedures called the "mapped design" (Hahn and others 1995) fits nicely into FIA's use of double sampling. The mapped design physically divides ground plots into areas of similar condition such as nonforest, forest condition 1, forest condition 2, and so forth. The mapping is done to accommodate random plot location without having to "move" plots that straddle boundaries. The mapped design fits into double sampling because of the way attributes (y_{hj}) are initially calculated on a per-unit-area basis (eq. 1). For example, tree attributes require no modification other than considering "mapped condition" of interest to be a subpopulation throughout

Table 2—Comparing equation 9 variance to other alternatives[a] in terms of the coefficient of variation (CV) for gross volume outside National Forests in southern Idaho, 1991.

Forest type	No. of plots (n_h)	Eq. 9 CV for volume	Difference[b] as percent of eq. 9 CV				Difference[c] in CV (due to *fpc*) between English and metric units			
			eq. 10[d]	Rao eq. 4	Cochran eq. 12.24'	Cochran eq. 12.25	Eq. 9	Rao eq. 4	Cochran eq. 12.24'	Cochran eq. 12.25
			\- *Percent* \-							
Aspen	41	19	−0.58	−0.04	0.21	−0.04	0	−0.011	−0.012	−0.011
Douglas-fir	86	12	−1.05	−0.04	0.16	−0.09	0	−0.013	−0.014	−0.013
Other timberland	47	17	−0.64	−0.04	0.22	−0.09	0	−0.012	−0.013	−0.012
Ponderosa pine	35	20	−0.62	−0.04	0.37	−0.05	0	−0.013	−0.014	−0.013
Woodland	83	12	−2.59	−0.06	0.04	−0.07	0	−0.018	−0.018	−0.018

[a]The alternative variance equations for total volume ($\hat{Y}$) include:
 (1) Rao 1973 eq. 4 (also in Jeyaratnam and others 1984, eq. 1.8)

$$v(\hat{Y}) = \frac{N(N-1)}{n'(n'-1)} \sum_{h=1}^{L} n_h'^2 \left(\frac{1}{n_h} - \frac{1}{n_h'} \right) n_h v(\bar{y}_h) + \frac{N(N-n')}{n'(n'-1)} \left[\sum_{h=1}^{L} \frac{n_h'}{n_h} \sum_{j=1}^{n_h} y_{hj}^2 - \sum_{h=1}^{L} n_h \bar{y}^2 \right]$$

(2) Cochran (1977) eq. 12.24'

$$v(\hat{Y}) = N^2 \left[\sum_{h=1}^{L} w_h^2 v(\bar{y}_h) - \sum_{h=1}^{L} \frac{w_h n_h v(\bar{y}_h)}{N} + \frac{(N-n')}{n'(N-1)} \sum_{h=1}^{L} w_h (\bar{y}_h - \bar{y})^2 \right]$$

(3) Cochran (1977) eq. 12.25

$$v(\hat{Y}) = \frac{n'(N^2-N)}{(n'-1)} \left[\sum_{h=1}^{L} w_h n_h v(\bar{y}) \left(\frac{n_h'}{n' n_h} - \frac{1}{N} \right) + \frac{(N-n')}{n'(N-1)} \sum_{h=1}^{L} n_h v(\bar{y}) \left(\frac{w_h}{N} - \frac{n_h'}{n' n_h} \right) + \frac{(N-n')}{n'(N-1)} \sum_{h=1}^{L} w_h (\bar{y}_h - \bar{y})^2 \right]$$

[b]Difference defined as: $100(CV_i - CV_{eq.9})/CV_{eq.9}$, where i equals alternative variance equations.
[c]Difference defined as: $100(CV_m - CV_E)/CV_E$, where m equals metric units and E equals English units. The *fpc* is finite population correction factor.
[d]Variance is computed with equation 10 (which almost corresponds to the first term in equation 9):

$$v(\hat{Y}) = N^2 \sum_{h=1}^{L} w_h^2 v(\bar{y}^2)$$
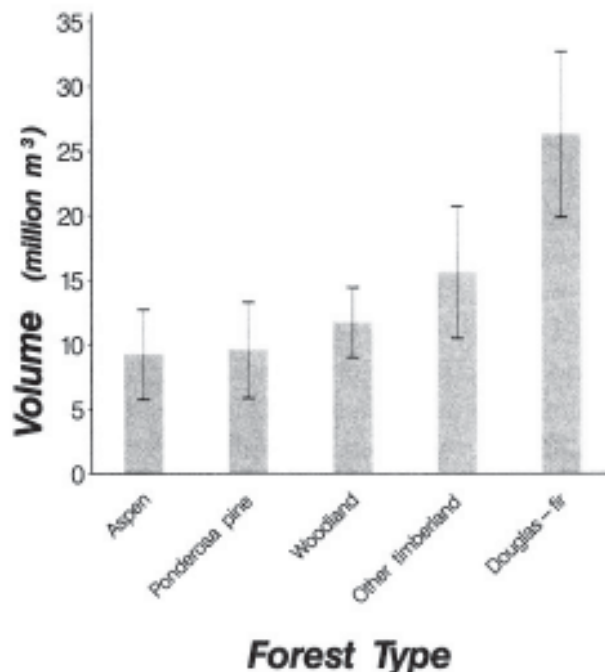
Figure 2—Gross volume estimated for forest types on land outside National Forests in southern Idaho, 1991. Uncertainty of each estimate is shown as a 95 percent confidence interval (black bar).

data summary (Scott and Bechtold 1995, eq. 22). For an area attribute (y_{hj} = 0 or 1), a proportional modification to equation 2 is needed to compute stratum means:

$$\bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} p_{hj} \tag{18}$$

where

p_{hj} = proportion of plot j area in a mapped condition of interest.

Likewise, variance calculations for area in the "mapped design" use p_{hj} substituted for y_{hj} in equations 8, 12, and 15 (Scott and Bechtold 1995, p. 56).

Ratio Method

The main purpose of the ratio method is to estimate subpopulation means by dividing attribute totals by corresponding area totals, but it has other uses. For example, FIA often reports attribute totals for subpopulations as percentages (ratios) of the population total (Chojnacky 1995, fig. 6-9). The ratio estimator uses equation 5 results in equation 6. Variance computations use equations 13, 14, and 15, in addition to equations 8, 9, and 11.

Another useful application of the ratio estimator is a situation where an attribute (y_{hj}) is a relative proportion of a ground plot, such as percentage canopy cover. For example, ratio calculations for southern Idaho data show understory cover ranging from 20 to 50 percent for plant subpopulations (forb, grass, and shrub) within forest types (fig. 3). In this example, the mean percentage of understory canopy cover (O'Brien and Van Hooser 1983) is obtained from the "cumulative" cover percentage ($\hat{Y}_1$) for a life-form forest-type subpopulation divided by the subpopulation's corresponding area ($\hat{Y}_2$).
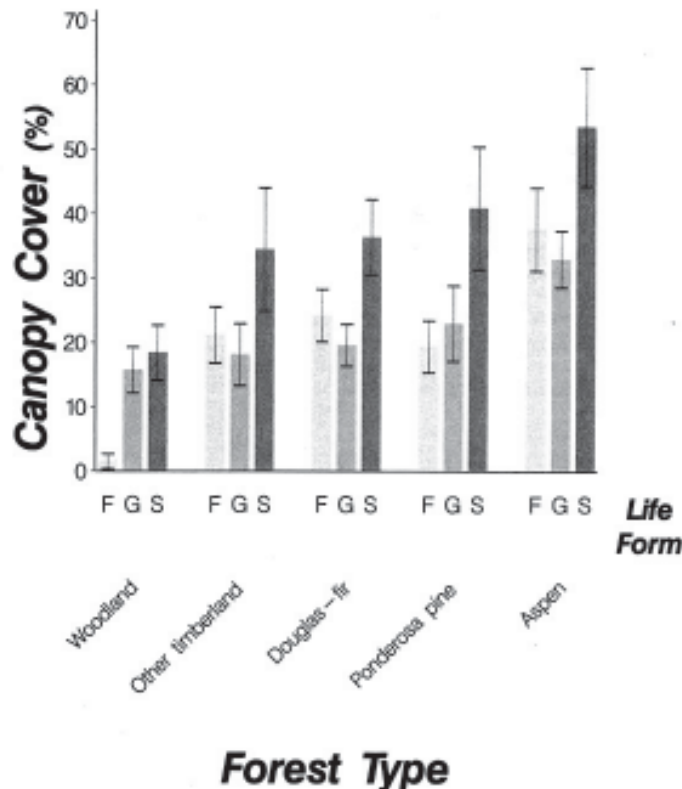
Figure 3—Understory canopy cover estimated for life forms (F = forb, G = grass, and S = shrub) within forest types on land outside National Forests in southern Idaho, 1991. Uncertainty of each estimate is shown as a 95 percent confidence interval (black bar).

County and Owner Estimation Method

All attributes for county and/or ownership subpopulations use equation 7 in place of equation 3 for estimating sample means. This estimator (eq. 7) utilizes the county and ownership information from the first-phase sample to categorize attributes more precisely than can be done from the smaller second-phase sample. This estimator is particularly appealing for estimating attributes for small ownership parcels that are missed for some counties in the coarse second-phase sample. However, the county and owner estimator (eq. 7) deviates from classical double sampling theory, and the statistical consequences of this deviation are unknown. Variance approximation for this estimator uses equation 12 in place of equation 8.

Phase-One Sample Size Modification

Interior West FIA adjusts (weights) the number of points sampled in each stratum (n_h') to correspond to known area data available for some land ownerships (table 1). The adjustment is done to make the inventory results match known land areas to the nearest 0.4 ha (1 acre) to eliminate inconsistencies with other published results. This adjustment has political significance, but a data summary of total forest land for several forest types in southern Idaho showed little difference between using "adjusted" and "unadjusted" first-phase points.

Conclusions _____

Strict application of a theoretical sampling design to actual field conditions is often a difficult task. Usually, additional modifications and assumptions are made to facilitate field sampling. The deviations from classical double sampling theory found in the FINSYS-2 Fortran program seem to have little practical significance for parameter estimation.

Overall, double sampling for stratification is a relatively simple yet powerful method used to summarize all attributes for FIA inventories in the Interior West. Eighteen equations are needed to describe the entire compilation process used to summarize attribute totals, means, and variances. Also the currently used double sample design accommodates FIA's new "mapped design" for field data collection.

References _____

Bickford, C. A. 1952. The sampling design used in the forest survey of the Northeast. Journal of Forestry. 50(4): 290-293.

Bickford, C. A. 1959. A test of continuous inventory for National Forest management based upon aerial photographs, double sampling, and remeasured plots. Proceedings, Society of American Forestry 1959: 143-148.

Bickford, C. A.; Mayer, C. E.; Ware, K. D. 1963. An efficient sampling design for forest inventory: the Northeastern forest survey. Journal of Forestry. 61(11): 826-833.

Born, J. D. 1996. [Personal communication]. March 11. Retired from U.S. Department of Agriculture, Forest Service, Intermountain Research Station, Forestry Sciences Laboratory, Ogden, UT.

Born, J. D.; Barnard, J. E. 1983. FINSYS-2: Subsystem TABLE-2 and OUTPUT-2. Gen.Tech. Rep. NE-84. Broomall, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 133 p.

Campbell, K. 1994. Sampling the continuum: the spatial support of environmental data. In 1993 proceedings of American Statistical Association, Statistical Computing Section. Alexandria, VA: American Statistical Association: 11-19.

Chojnacky, D. C. 1995. Southern Idaho's forest land outside National Forests, 1991. Resour. Bull. INT-RB-82. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 58 p.

Chojnacky, D. C. 1996. Forest inventory and analysis (FIA) variables: indicators of ecological integrity. In: Proceedings, North American workshop on monitoring for ecological assessment of terrestrial and aquatic ecosystems; 1995 September 18-22; Mexico City. Gen. Tech. Rep. RM-GTR-284. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Research Station: 247-258.

Cochran, W. G. 1977. Sampling Techniques. 3d ed. New York: John Wiley and Sons. 428 p.

Hahn, J. T.; MacLean, C. D.; Arner S. L.; Bechtold, W. A. 1995. Procedures to handle inventory cluster plots that straddle two or more conditions. Forest Science Monograph. 31: 12-25.

Husch, B.; Miller. C. I.; Beers, T. W. 1982. Forest Mensuration. 3d ed. New York: John Wiley and Sons. 402 p.

Jeyaratnam, S.; Bowden, D. C.; Graybill, F. A.; Frayer, W. E. 1984. Estimation in multiphase designs for stratification. Forest Science. 30(2): 484-491.

Li, H. G., Schreuder, H. T., Van Hooser, D. D., Brink, G. E. 1992. Estimating strata means in double sampling with corrections based on second-phase sampling. Biometrics. 48: 189-199.

Moisen, G. G.; Stage, A. R.; Born, J. D. 1995. Point sampling with disjunct support near population boundaries. Forest Science Monograph. 31: 62-82.

Mood, M. A.; Graybill, F. A.; Boes, D. C. 1974. Introduction to the theory of statistics. 3d ed. New York: McGraw-Hill. 564 p.

Neyman, J. 1938. Contribution to the theory of sampling human populations. Journal of American Statistical Association. 33: 101-116.

O'Brien, R. A.; Van Hooser, D. D. 1983. Understory vegetation inventory: an efficient procedure. Res. Pap. INT-323. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station. 6 p.

Rao, J. N. K. 1973. On double sampling for stratification and analytical surveys. Biometrika. 60(1): 125-133.

SAS Institute Inc. 1990. SAS procedures guide, version 6, 3d ed. Cary, NC: SAS Institute Inc. 705 p.

Schreuder, H. T.; Gregoire, T. G.; Wood, G. B. 1993. Sampling methods for multiresource forest inventory. New York: John Wiley and Sons. 446 p.

Scott, C. T.; Bechtold, W. A. 1995. Techniques and computations for mapping plot clusters that straddle stand boundaries. Forest Science Monograph. 31: 46-61.

U.S. Department of Agriculture, Forest Service. 1989. Idaho photopoint classification procedures, 1988-1989. Unpublished document on file at: U.S. Department of Agriculture, Forest Service, Intermountain Research Station, Forestry Sciences Laboratory, Interior West Resource Inventory, Monitoring, and Evaluation Program, Ogden, UT.

U.S. Department of Agriculture, Forest Service. 1992. Forest Service resource inventories: an overview. Forest Inventory, Economics, and Recreation Research Staff. Washington, DC: U.S. Department of Agriculture, Forest Service. 39 p.

Williams, M. S. 1995. Improved estimation of domain totals in forest surveys. Canadian Journal of Forest Research. 25: 1203-1207.

This paper documents the use of double sampling for Forest Inventory and Analysis (Forest Service, U.S. Department of Agriculture) inventories in the Interior West. Results show 18 equations describe the entire inventory summarization process for estimating population totals and means, and respective variances. Most equations are for standard use of double sampling, but equations are also given for a ratio method and a special method for data defined by county and land ownership. Inventory data from 800,000 ha of forest land in southern Idaho are used to illustrate application of several equations.

RMRS
ROCKY MOUNTAIN RESEARCH STATION

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona	Reno, Nevada
Fort Collins, Colorado*	Albuquerque, New Mexico
Boise, Idaho	Rapid City, South Dakota
Moscow, Idaho	Logan, Utah
Bozeman, Montana	Ogden, Utah
Missoula, Montana	Provo, Utah
Lincoln, Nebraska	Laramie, Wyoming

*Station Headquarters, 240 West Prospect Road, Fort Collins, CO 80526