Estimating Two-Way Tables Based on Forest Surveys

Charles T. Scott

Abstract.—Forest survey analysts usually are interested in tables of values rather than single point estimates. A common error is to include only plots on which nonzero values of the attribute were observed when computing the variance of a mean. Similarly, analysts often exclude nonforest plots from the analysis. The development of the correct estimates of forest area, attribute totals, and their means over the area of interest is described. Program TabGen was written to perform these calculations correctly assuming simple random sampling, stratified random sampling, or double sampling for stratification.

The results of forest surveys generally are produced in the form of one- and two-way tables. For example, forest managers and planners may make decisions based on tables of the area by forest type and stand size or volumes by species and diameter class. The statistical reports of the Forest Inventory and Analysis (FIA) units of the USDA Forest Service are compilations of such tables. Although we regularly use such tables, the forest survey and sampling literature describes sampling designs and alternative estimators for a single attribute of interest rather than for tables of them.

There is confusion over how to construct tables involving a measured variable, e.g., volume, which is then divided into rows and columns by categorical variables, such as stage of development and site class. Typically, the table total is divided into rows that are further broken down into columns. These values are not unrelated estimates because they must add across and down. The literature does not address this issue directly. Sometimes additivity constraints are imposed after estimating each cell independently (Li and Schreuder 1985). For the sampling designs and estimators described here, the tables are constructed to be additive (except for ratio estimators). Because cells are estimated independently, there often is confusion over the sample size of each table cell, which then affects the estimate of precision for that cell.

In many forest surveys, especially in the United States, there is interest in estimating not only various forest characteristics but also forest land area. This requires sampling the entire land area and then estimating the portion that is forested. This creates confusion in sample sizes and in the expression of the values as totals and on a per unit area basis.

In this paper, I present estimators for construction tables of values and their variances. The estimation steps presented are designed to avoid some of the confusion and mistakes described previously. These procedures have been incorporated into a software package called TabGen.

Project Leader, USDA Forest Service, Northeastern Research Station, 359 Main RD, Delaware, OH, USA.

METHODS

The first step in clarifying the estimation of tables is to present the estimators in the context of estimating tables. This is done at some length in this section to set the stage for addressing sources of confusion such as sample size, estimation of forest area, and expressing values on a per unit area basis.

Estimation

Typically, estimators are derived for a single attribute of interest, such as, total biomass of a forest. When estimating tables, one may be interested in estimating biomass by species group and diameter class. The estimation process is the same for each cell—only the attribute of interest changes. Alternatively, the estimation process and the attribute of interest are the same in each cell, but different conditions are placed on the attribute of interest in each table cell. The latter approach is taken here.

Each row and column category can be thought of as a condition to be placed on the attribute of interest. Row and column variables must be categorical or must be continuous variables that have been divided into classes, e.g., diameter classes. The attribute of interest is "summed" into a cell only when it meets the row and column conditions. Each categorical variable has an associated indicator variable—either the attribute is in the category of interest (1) or it is not (0). For example, when estimating the volume in Site Class 3, if the Site Class for the sampling unit (plot or cluster of plots) is 3, the indicator variable for the sampling unit is assigned as 1, otherwise it is 0. So for each cell, indicator variables can be assigned to an observation for each categorical variable (row and column variables)—either the observation

belongs in that cell or it does not. This method is described in Cochran (1977, p. 142-144) for estimating domain (cell) means.

These indicator variables can then be used in the estimation process. Because values of 0 and 1 were chosen, the indicator variables can be multiplied times the attribute of interest for each observation. If either the row or column indicator is 0, the product of the two indicator variables and the attribute of interest is zero. This product is then averaged across all sampling units within each stratum. In the simple random sampling case, this is the final estimate because simple random sampling can be thought of as stratified random sampling with a single stratum. However, stratification often is used to reduce the variance of the estimates by subdividing the population into relatively homogenous strata. In this case, the strata means are averaged using the stratum areas as weights (stratified random sampling estimator) or using estimated stratum weights (double sampling for stratification). The variances also are computed using the product of the attribute of interest and the indicator variables.

The formulas that follow assume stratified random sampling where strata are defined on maps or satellite imagery. Examples include map-based forest type, stand-density classes, and land-use classes. The simple random sampling formulas are the same except that there is only one stratum. Double sampling for stratification (Cochran 1977) estimators are the same except that additional terms are added to the variance because stratum areas also must be estimated (see also Chojnacky 1998). Examples include photointerpreted land-cover classes, or volume classes.

The population mean (per unit area) for the attribute of interest in row, r, and column, c, is estimated as:

$$\bar{Y}_{rc} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{Y}_{hrc} = \sum_{h=1}^{H} \frac{N_h}{N n_h} \left(\sum_{i=1}^{n_h} Y_{hi} I_{hir} I_{hic} \right) \quad (1)$$

where:

N_h = area in stratum h, where h=1, 2, ..., H

N = total area in population (across all H strata)

n_h = number of sampling units observed in stratum h

Y_{hi} = attribute of interest (expressed on a per unit area basis) on sampling unit i in stratum h. For example, the sum of tree observations divided by their plot area.

I_{hir} = indicator variable for row r of the first categorical attribute of interest for sampling unit i in stratum h. Equals 1 if the value of the attribute matches row r, but is 0 otherwise.

I_{hic} = indicator variable for column c of the second categorical attribute of interest for sampling unit i in stratum h. Equals 1 if the value of the attribute matches column c, but is 0 otherwise.

In other words, the products of the attribute of interest, Y_{hi}, and the two indicator variables are summed over all sampling units in a stratum and divided by n_h to form the strata means. These means are combined using the stratum weights to form an estimate of the population mean for each cell (r, c) in turn. Whatever the value of Y_{hi}, the contribution of a sampling unit is 0 if either of the indicator variables I_{hir} or I_{hic} is zero.

To estimate the proportion of the area in each table cell (combination of r and c), drop out the attribute of interest from equation (1):

$$\bar{A}_{rc} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{A}_{hrc} = \sum_{h=1}^{H} \frac{N_h}{N n_h} \left(\sum_{i=1}^{n_h} I_{hir} I_{hic} \right) \quad (2)$$

Note that the sum in parentheses gives the number of sampling units that fall in row r and column c within stratum h.

The estimated variances are computed from Cochran (1977, eq. 5.12) as:

$$v(\bar{Y}_{rc}) = \sum_{h}^{H} \frac{N_h}{N} \frac{N_h - n_h}{N_h} v\left(\bar{Y}_{hrc}\right) = \quad (3)$$

$$\sum_{h}^{H} \frac{N_h}{N} \frac{N_h - n_h}{N_h} \frac{\left(\sum_{i=1}^{n_h} Y_{hirc}^2 \right) - n_h \bar{Y}_{hrc}^2}{n_h (n_h - 1)}$$

and,

$$v(\bar{A}_{rc}) = \sum_{h}^{H} \frac{N_h}{N} \frac{N_h - n_h}{N_h} v\left(\bar{A}_{hrc}\right) = \quad (4)$$

$$\sum_{h}^{H} \frac{N_h}{N} \frac{N_h - n_h}{N_h} \frac{\left(\sum_{i=1}^{n_h} A_{hirc}^2 \right) - n_h \bar{A}_{hrc}^2}{n_h (n_h - 1)}$$

where:

Y_{hirc} = product of attribute of interest and the indicator variables
= $Y_{hi} I_{hir} I_{hic}$

A_{hirc} = product of the indicator variables
= $I_{hir} I_{hic}$

To estimate population totals such as total volume or area of Site Class 3, multiply equations (1) and (2) by the known total area in the population, A_T. Their estimated variances are then equations (3) and (4) multiplied by A_T^2.

If simple random sampling is used, then H = 1 and $N_h = N$. If double sampling for stratification is used, Cochran (1977, eq. 12.32) replaces equations (3) and (4). Note that equations (5) and (6) simplify to (3) and (4) if the total number of first-phase samples (usually interpreted on aerial photographs), N, equals the total area, A_T.

$$v(\bar{Y}_{rc}) = \left(1 - \frac{1}{A_T}\right) \sum_{h=1}^{H} \left(\frac{N_h - 1}{N - 1} - \frac{n_h - 1}{A_T - 1}\right) \frac{N_h}{N} \frac{\left(\sum_{i=1}^{n_h} Y_{hirc}^2\right) - n_h \bar{Y}_{hrc}^2}{n_h(n_h - 1)} + \left(1 - \frac{N}{A_T}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{\left(\bar{Y}_{hrc} - \bar{Y}_{rc}\right)^2}{N - 1} \quad (5)$$

and,

$$v(\bar{A}_{rc}) = \left(1 - \frac{1}{A_T}\right) \sum_{h=1}^{H} \left(\frac{N_h - 1}{N - 1} - \frac{n_h - 1}{A_T - 1}\right) \frac{N_h}{N} \frac{\left(\sum_{i=1}^{n_h} A_{hirc}^2\right) - n_h \bar{A}_{hrc}^2}{n_h(n_h - 1)} + \left(1 - \frac{N}{A_T}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{\left(\bar{A}_{hrc} - \bar{A}_{rc}\right)^2}{N - 1} \quad (6)$$

where:

N_h = number of first-phase samples that fell in stratum h

N = total number of first-phase samples taken to determine strata areas

These equations are all that is needed to construct tables of means for an attribute of interest and for area proportions, and to create tables of their variances under simple random sampling, stratified random sampling, or double sampling for stratification designs.

Sample Size

Note that for estimates of all cells, the sample sizes, n_h, have remained the same. Often, survey analysts assume that the sample size is the number of sampling units that "fell" into a cell, that is, the number of sampling units for which both indicators were 1. Mathematically, this is the sum of A_{hirc}. This sum is a random variable—it was not known (fixed) in advance of sampling (as was the true sample size, n_h). In fact, when the sum is divided by the sample size, it forms an estimate of the proportion of sampling units falling in the cell of interest, which is a random variate (we can estimate its variance in (4)). Another way to understand this is that all attributes of interest are observed on all sampling units but that often the observation is zero.

Often, interest is in the mean of an attribute of sampling units falling in a cell. Unfortunately, the mean estimated in equation (1) is the mean based on all sampling units. Often, these means are difficult to interpret, such as the average volume per acre that includes both forest and nonforest acres. Thus, estimates using equation (1) generally are transformed by multiplying by the total area, A_T, to yield totals across the population. Further, these totals can be divided by the estimated area in each cell resulting in an estimate of the mean of an attribute for sampling units falling in a cell, as shown in the following section.

Expressing Values on a Per Unit Area Basis

Often, interest is in the mean of observations falling within a particular cell, expressed on a per unit area basis, such as the volume per hectare of pine in pine plantations versus the volume of pine on an "average" hectare across the whole forest, as is given by equation (1). Rather than using only sampling units that fall in the cell of interest, all sampling units are used—first to estimate the total volume of pine in pine plantations and then to estimate the total area of pine plantations. Dividing the total by the area then gives a biased (of order 1/n) estimate of the attribute of interest on a per unit basis for the area of interest. This is equivalent to dividing the mean across the whole forest (1) by the proportion of the area in the class of interest (2):

$$R_{rc} = \frac{\bar{Y}_{rc}}{\bar{A}_{rc}} \quad (7)$$

The approximate variance of this combined ratio-of-means estimator (Cochran 1977, eq. 6.51) is:

$$v(R_{hrc}) = v(\bar{Y}_{hrc}) + R_{rc}^2\, v(\bar{A}_{hrc}) - 2\, R_{rc}\, cov(\bar{Y}_{hrc}, \bar{A}_{hrc}) \tag{8}$$

where:

$$v(R_{hrc}) = v(\bar{Y}_{hrc}) + R_{rc}^2\, v(\bar{A}_{hrc}) - 2\, R_{rc}\, cov(\bar{Y}_{hrc}, \bar{A}_{hrc})$$

$$v(R_{hrc}) = v(\bar{Y}_{hrc}) + R_{rc}^2\, v(\bar{A}_{hrc}) - 2\, R_{rc}\, cov(\bar{Y}_{hrc}, \bar{A}_{hrc})$$

Note that a single estimate of the ratio is used rather than one for each stratum. The covariance between Y_{hrc} and A_{hrc} is estimated as:

$$cov(\bar{Y}_{hrc}, \bar{A}_{hrc}) = \frac{\left(\sum_{i=1}^{n_h} Y_{hirc}\, A_{hirc}\right) - n_h\, \bar{Y}_{hrc}\, \bar{A}_{hrc}}{n_h\,(n_h - 1)} \tag{9}$$

Analogously, an approximation to the estimated variance in the double sampling for stratification case is:

$$v(R_{rc}) = \frac{1}{R_{rc}^2}\left(1 - \frac{1}{A_T}\right)\sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1} - \frac{n_h - 1}{A_T - 1}\right)\frac{N_h}{N}\, v(R_{hrc}) + \frac{1}{R_{rc}^2}\left(1 - \frac{N}{A_T}\right)\sum_{h=1}^{H}\frac{N_h}{N\,(N-1)}\left(\frac{\bar{Y}_{hrc}}{\bar{A}_{hrc}} - R_{rc}\right)^2 \tag{10}$$

This process is repeated for each combination of the categorical variables, plus the total across all classes (row and column margins). This provides all of the formulas needed to construct a variety of tables for attributes of interest, area, and the ratios between the two. Note that if estimates are to be placed on a per observation basis, e.g., per tree, A_{hi} values should be replaced with the number of observations per acre.

Estimation of Forest Area

Forest inventories for an ownership typically are map based, and only forest areas are sampled. In regional forest inventories such as FIA, no maps of all forested areas are available, so forest area must be estimated from the sample. The forest area is estimated easily using equation (2). The categories of interest are forest and nonforest. The proportion of sampling units falling in forested conditions times the total land area sampled, A_T, yields an estimate of the total forest area. Because remaining tables would focus only on the forested sampling units, it is tempting for the analyst to use only the forested sampling units in further analyses. However, the number of forested sampling units is a random variable, so all sampling units must be included in all analyses. This is done using the methods described earlier. FIA often presents the results as totals, A_T times equation (1), to simplify the estimation. Means across forest and nonforest areas, as given in equation (1), are rarely useful; means generally would have to be estimated using equation (7) with total forest area or some subset of it as the denominator. This is more complicated because it uses equations (1)-(4) and (7)-(9), rather than only (1)-(2).

Program TabGen

To avoid these complexities and confusion, program TabGen (Table Generator) was developed. Written in Visual Basic, TabGen allows the user to create tables in a point-and-click environment. It uses all the formulas presented here and avoids the pitfalls mentioned.

Currently, the program reads a flat file for each of the following: plot (sampling unit) data, site index trees, regeneration, overstory trees, and fields common to both regeneration and overstory trees. Although other files can be substituted, a hierarchy of one plot file and one or

more files with multiple observations per plot is assumed. TabGen also reads a control file that contains the stratum weights, N_h, the total area, A_T, and a list of plots that will be included in the analysis. This list allows the analyst to include only the subset of plots of interest. We have a script in Arc/Info (or ArcView) that creates this file based on the areas that have been selected using the GIS.

When the user selects the control file, TabGen reads a variable library (dictionary) that describes the variables that are read from each file, whether they are continuous or categorical, and, if categorical, the labels for each category. The program also allows the user to create categorical variables from continuous ones by assigning ranges to categories. The user then selects the row and column categorical variables and the attribute of interest (continuous variable). The results can be viewed as:

1. The percent of the area or total area in each cell—equation (2)
2. The mean or total of the attribute of interest in each cell—equation (1)
3. The ratio of the mean to the area estimate for each cell—equation (7)
4. The mean of individual observations (generally on a per tree basis)
5. The number of plots "falling" in each cell

The 95 percent sampling errors (confidence limits) are given for each cell:

$$SE\% = 100\, t\, \frac{\sqrt{V}}{\bar{X}} \approx 200\, \frac{\sqrt{V}}{\bar{X}} \qquad (11)$$

where:

t = Student's t-value with = 0.05 and n-H degrees of freedom. As n-H goes to infinity, t goes to 1.96; thus, 2 is substituted.

V = variance estimate of the mean

The estimates and their variances are computed using simple random sampling, stratified random sampling, or double sampling for stratification estimators.

TabGen generates one- or two-way tables, although filters can be used to create a series of two-way tables to create three-way tables. A filter is a rule created for a variable that determines which plots or observations will be included/excluded from the analysis. Multiple filters can be created for each variable. For example, a tree-size filter can be created from dbh. A table of volume by species and site class can be created first for poletimber trees and then for sawtimber trees. Filters also can be used in combination, for example, adding a filter only for pine plantations.

TabGen was written for a specific study, so it is not yet as general as it might be. However, it is public domain software and is available for modification for other purposes. Copies of TabGen are available from the author (cscott/ne_de@fs.fed.us).

CONCLUSIONS AND RECOMMENDATIONS

The literature provides little guidance on the estimation of tables, so survey analysts often encounter problems when summarizing forest surveys. A common problem results when using sampling units that fall into a particular category of interest to compute means. When simple random sampling is used, the means and perhaps the variances are correct. However, when stratified random sampling, double sampling for stratification, or unequal probability sampling is used, this "shortcut" yields inaccurate means and variances. The methods shown in equations (7)-(9) yield the correct results in all cases.

Survey analysts may wish to ignore nonforest sampling units, but they are part of the sample and must be included in the analysis. TabGen uses the methods described above and gives the survey analyst considerable flexibility to generate a variety of tables from forest survey data. However, the analyst must understand the estimation process to avoid drawing erroneous conclusions.

ACKNOWLEDGMENTS

I gratefully acknowledge the support of Mead Paper Corporation and thank Scott Klopfer for his part in programming TabGen. The following people reviewed this manuscript and contributed markedly to it: Stan Arner, USDA Forest Service, Radnor, PA, USA; Bill Bechtold and Stan Zarnoch, USDA Forest Service, Asheville, NC, USA; and Dave Chojnacky, USDA Forest Service, Ogden, UT, USA.

LITERATURE CITED

Chojnacky, D.C. 1998. Double sampling for stratification: a forest inventory application in the Interior West. Res. Pap. RMRS-RP-7. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 15 p.

Cochran, W.G. 1977. Sampling techniques. New York: John Wiley & Sons. 428 p.

Li, H.G.; Schreuder, H.T. 1985. Adjusting estimates in large two-way tables in surveys. Forest Science. 31: 366-372.