## 8.2 Stratified Random Sampling for Forest Assessment

Two principal reasons to use stratified sampling:

1. Desire estimates of total volume, or growth, or biomass, or ?, by management units, site productivity classes, forest types, value classes, age or size classes, etc., as well as for the whole forest

2. Precision in estimate is often higher for the whole forest for a given sampling effort (intensity), as long as our strata definitions produce rather homogenous units, compared to population as a whole

Consider a forest with 3 stands (strata), defined by criteria such as those in item 1 above
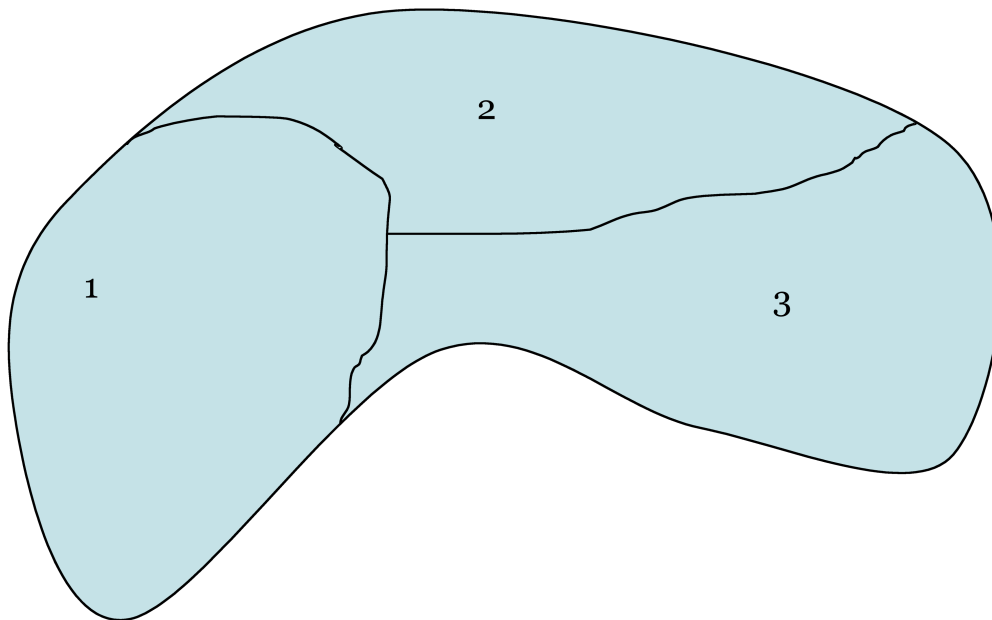
In general, we will have a number, M, of strata with sizes N_1, N_2, …, N_M , where $N_1 + N_2 + … + N_M = N$, the total size of the entire population.

Then, the stratified estimate of the total is given by

$$T_{st} = T_1 + T_2 + \cdots + T_M$$

Further, since $T_j = N_j \cdot \bar{x}_j$, an alternate way to express the total is in terms of the strata means,

$$T_{st} = N_1 \cdot \bar{x}_1 + N_2 \cdot \bar{x}_2 + \cdots + N_M \cdot \bar{x}_M$$

Dividing both sides by the total population size, N,

$$\frac{T_{st}}{N} = \bar{x}_{st} = \frac{N_1 \bullet \bar{x}_1}{N} + \frac{N_2 \bullet \bar{x}_2}{N} + \cdots + \frac{N_M \bullet \bar{x}_M}{N}. \text{ Letting } P_j = \left(\frac{N_j}{N}\right),$$

$$\bar{x}_{st} = P_1 \bullet \bar{x}_1 + P_2 \bullet \bar{x}_2 + \cdots + P_M \bullet \bar{x}_M$$

We see that the stratified estimate of the population mean (based on the sample) is just a weighted mean of the stratum means (weighted by stratum sizes)

This estimator of the mean for the entire population is built up from the descriptive statistics from each stratum. For each individual stratum, variance of the mean is given by

$$S_{\bar{x}_j}^2 = \frac{S_j^2}{n_j}\left(\frac{N_j - n_j}{N_j}\right),$$

which is used in the formula to estimate the variance of the stratified sample mean:

$$S_{\bar{x}_{st}}^2 = P_1^2 S_{\bar{x}_1}^2 + P_2^2 S_{\bar{x}_2}^2 + \cdots + P_M^2 S_{\bar{x}_M}^2$$

Example. We are interested in knowing the number of cords per acre on our 100-ac forest. We used aerial photography to divide the forest into three forest types: pine, mixed pine/hardwood, and bottomland hardwood.

Stratum	Type	Size
1	Pine	30 acres
2	Mxd P/H	50 acres
3	Bot. H	20 acres

We've got a budget for 20, 1-acre fixed area plots. We figure we ought to sample more in the types with large area, less in the types with small area to keep sampling intensity about the same in each type.

	Pine (cds)	Mxd P/H (cds)		Bot. H (cds)
	10	41	29	57
	17	39	33	54
	19	37	34	47
	15	41	40	39
	9	27		
	22	36		
n_j	6	10		4
$\displaystyle\sum_{i=1}^{n_j} x_{ij}$	92	357		197
$\bar{x}_j$	15.333	35.700		49.250
$\displaystyle\sum_{i=1}^{n_j} x_{ij}^2$	1,540	12,963		9,895
S_j^2	25.8667	24.2333		64.2500
$S_{\bar{x}_j}^2$	3.4489	1.9387		12.8500
N_j	30	50		20

$$\bar{x}_{st} = \frac{N_1 \cdot \bar{x}_1}{N} + \frac{N_2 \cdot \bar{x}_2}{N} + \frac{N_3 \cdot \bar{x}_3}{N}$$

$$\bar{x}_{st} = \frac{30 \cdot 15.33}{100} + \frac{50 \cdot 35.70}{100} + \frac{20 \cdot 49.25}{100} = 32.30$$

$$S_{\bar{x}_{st}}^2 = \left(\frac{N_1}{N}\right)^2 S_{\bar{x}_1}^2 + \left(\frac{N_2}{N}\right)^2 S_{\bar{x}_2}^2 + \left(\frac{N_3}{N}\right)^2 S_{\bar{x}_3}^2$$

$$S_{\bar{x}_{st}}^2 = \left(\frac{30}{100}\right)^2 3.4489 + \left(\frac{50}{100}\right)^2 1.9387 + \left(\frac{20}{100}\right)^2 12.8500 = 1.3091$$

$$S_{\bar{x}_{st}} = \sqrt{S_{\bar{x}_{st}}^2} = 1.14$$

Confidence Interval

$$\bar{x}_{st} \pm t_{\alpha,df} \cdot s_{\bar{x}_{st}},$$

where df = n – M, and n = n_1 + n_2 + … + n_M

So for 95% CI, in this case df = 20 – 3 = 17, and $t_{0.05,17}$ = 2.110

$$32.30 \pm 2.11 \cdot 1.14 \quad \Rightarrow \quad (29.89,\ 34.70) \text{ cds/acre}$$

Or on the entire 100-acre forest:

$$100 \cdot (29.89,\ 34.70) \quad \Rightarrow \quad (2989,\ 3470) \text{ cds}$$

What if Simple Random Sampling (SRS) or Systematic had been used?

$$S_{\bar{x}_{srs}} = \sqrt{S_{\bar{x}_{srs}}^2} = \sqrt{7.436} = 2.727$$

For 95% CI, df = n – 1 = 20 – 1 = 19, and $t_{0.05,19}$ = 2.093

$$32.30 \pm 2.093 \cdot 2.727 \quad \Rightarrow \quad (26.59,\ 38.01) \text{ cds/acre}$$

Or on the entire 100-acre forest:

$$100 \cdot (25.916,\ 38.684) \quad \Rightarrow \quad (2659,\ 3801) \text{ cds}$$

Relative Efficiency

Relative Efficiency (RE) is the statistical way to compare the performance of two sampling techniques. To compare sampling method B to A, calculate RE as

$$RE = \frac{S_{\bar{x}_A}^2}{S_{\bar{x}_B}^2} \qquad \text{(Note: "Standard" method goes in numerator, i.e.,}$$

numerator is the method you wish to compare another TO)

To compare Stratified Sampling to Simple Random, calculate

$$RE = \frac{S_{\bar{x}_{srs}}^2}{S_{\bar{x}_{st}}^2} = \frac{7.436}{1.309} = 5.68, \text{ i.e., stratified sampling is } \sim 6 \text{ times more efficient !}$$

Sample Allocation and Sample Size Calculation

Proportional allocation

- Simplest way to allocate samples and keeps sampling intensity the same in each stratum

  o $n_j = \left(\dfrac{N_j}{N} \right) n$

- Disadvantage is that large areas will receive more samples irrespective of variance

- When reliable strata can be recognized, prop. alloc. will be superior to sampling with same intensity and no strata

- Sample size calculation

  o $n = \dfrac{N \sum N_j S_j^2}{\dfrac{N^2 E^2}{t^2} + \sum N_j S_j^2}$ $\left(\text{or, relatively,} \quad n = \dfrac{N \sum N_j CV_j^2}{\dfrac{N^2 E\%^2}{t^2} + \sum N_j CV_j^2} \right)$

Optimal allocation

- Assures smallest possible standard error for a fixed cost

- Cost model:

  o $C = c_0 + \sum\limits_{j=1}^{M} c_j n_j$,

    where C is total cost of survey, c_0 is fixed overhead cost, c_j is cost of one sample unit in stratum j.

  o $n_j = \left(\dfrac{N_j S_j / \sqrt{c_j}}{\sum N_j S_j / \sqrt{c_j}} \right) n$

- Sample Size calculation

  o $n = \dfrac{\left(\sum N_j S_j \sqrt{c_j} \right)\left(\sum N_j S_j / \sqrt{c_j} \right)}{\dfrac{N^2 E^2}{t^2} + \sum N_j S_j^2}$