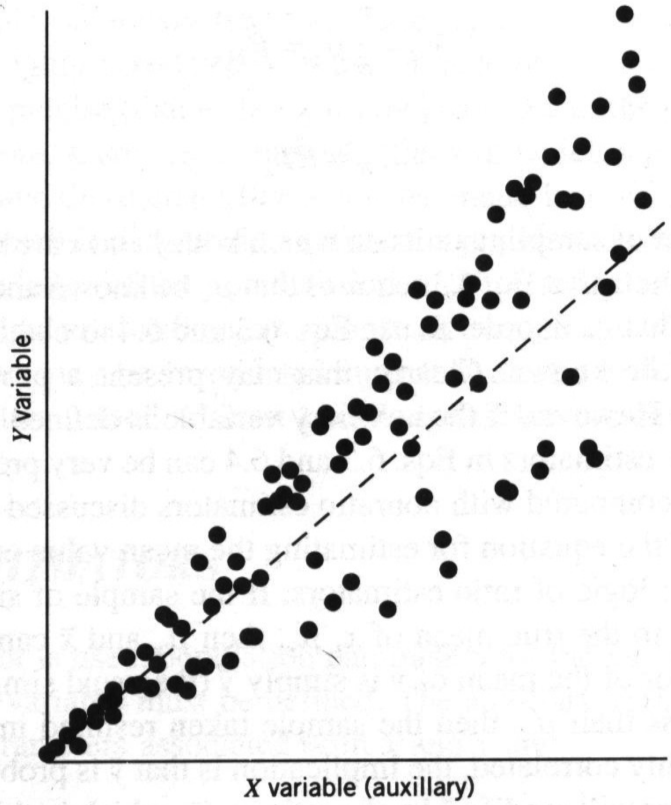
7.2 Stand Inventory using Ratio Estimation

Ratio estimation (and the closely related regression estimation) makes use of an auxiliary variable that is measured on each sampling unit in addition to the variable of interest

Sounds like extra work at first, because more measurements are needed, but if the auxiliary variable is highly correlated with the variable of interest, the extra effort will often result in extra precision on the variable of interest

Consider the case where an estimate of the biomass of all red cedar in a particular forest area is needed, which is very expensive due to destructive sampling involved. Since it's known that basal area is related to biomass, the basal area of all red cedar in that forest might be determined, providing support or additional help in estimating biomass. Biomass is the dependent variable of interest and the basal area is the auxiliary variable.



The population parameters associated with x and y are:

$T_x =$ population total of the auxiliary variable (X in Husch, et al.)

$\mu_x =$ population mean of the auxiliary variable ($\bar{x}$ in Husch, et al.)

$T_y =$ population total of the variable of interest ($\hat{V}$ Husch, et al.)

$\mu_y =$ population mean of the variable of interest ($\hat{V}/N$ in Husch, et al.)

A new population parameter that may be of interest is the ratio defined by

$$R=\frac{\mu_y}{\mu_x}=\frac{N\mu_y}{N\mu_x}=\frac{T_y}{T_x}$$

The meaning of the ratio (its units) obviously depends on how y and x are defined

Some examples

- If Y is height and X is DBH, then R is the Height–Diameter ratio
- If Y is crown length and X is height, then R is Live Crown Ratio
- If Y is Volume and X is basal area, then R is VBAR
- If Y is Volume today and X is volume 5 years ago, then R is proportionate increase (growth) in volume over the five year period

Ratio estimators are used to obtain estimates of $R,\ \mu_y,\ T_y$. The estimators are, respectively,

$$\hat{R}=\frac{\bar{y}}{\bar{x}}=\frac{\frac{1}{n}\sum_{i=1}^{n}y_i}{\frac{1}{n}\sum_{i=1}^{n}x_i}=\frac{\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i} \qquad (\bar{r} \text{ in Husch, et al.})$$

$$\bar{y}_R=\hat{R}\mu_x$$

$$\hat{T}_Y=\hat{R}T_x \qquad \text{(eq. 13-18 in Husch, et al.)}$$

where n is number of sampling units on which both y and x are measured, and all else is defined as before. ($\hat{R}$ is often called the "ratio-of-means" estimate.)

Note that population mean and/or total of X is required – many situations exist that if X is defined wisely, the mean or total of Y will be much more precise compared to non-ratio type estimators

Re-expressing the estimating equation for $\bar{y}_R$ may provide some insight into its logic:

$$\bar{y}_R=\ \hat{R}\mu_x=\ \frac{\bar{y}}{\bar{x}}\mu_x=\ \bar{y}\frac{\mu_x}{\bar{x}}$$

In this form, we see that if $\bar{x}=\mu_x$, they cancel each other and we are left with the usual SRS or SYS estimator of μ_y, which is the sample mean of y

If the sample mean of x is smaller than the population mean of x, then the sample taken gave an underestimate of μ_x. When X and Y are highly correlated, the implication is that $\bar{y}$ is probably underestimating μ_y also. So, $\bar{y}$ is adjusted upward proportionally to reflect the degree of underestimation

If $\bar{x}$ comes out larger than μ_x, then the opposite occurs

Of course, we still need to know how reliable our estimate is, so we need to compute variances and standard errors to construct CI's.

First, calculate:

$$S_u^2 = \frac{\sum_{i=1}^{n} y_i^2 + \hat{R}^2 \sum_{i=1}^{n} x_i^2 - 2\hat{R} \sum_{i=1}^{n} x_i y_i}{n-1}$$

This quantity represents how much variability there is in the LINEAR relationship between Y and X. When X and Y are highly correlated in linear fashion, then this quantity will be small (sum of cross-products will be large!)

Then, variance will be:

[1] For a ratio: $S_{\hat{R}}^2 = \frac{1}{\mu_x^2} \frac{S_u^2}{n} \left(\frac{N-n}{N} \right)$

[2] For a mean: $S_{\bar{y}_R}^2 = \frac{S_u^2}{n} \left(\frac{N-n}{N} \right)$

[3] For a total: $S_{\hat{T}_R}^2 = T_x^2 \frac{1}{\mu_x^2} \frac{S_u^2}{n} \left(\frac{N-n}{N} \right)$ (eq. 13-19 in Husch, et al.)

where, N represents the total number of sampling units in the population.

(Setting $\mu_x^2 S_u^2 = S^2$ in the previous formulas makes them equivalent to those in Husch, et al.)

The relationship between Y and X must be linear for this to work, but does not necessarily have to go through the origin, as long as sample size is large (say, n > 30)

Example. (Estimating a total using Ratio Estimation)

A forester is marking a stand of trees for a variable intensity thinning. The forester wants an estimate of volume for all the marked trees. Every tree that is marked is measured for DBH. On a sample of the marked trees, measurements are taken to enable estimation of total board-foot volume. A total of 500 trees are marked and measured for DBH and a sample of 35 marked trees are measured more intensively for volume.

For this problem, the variable of interest, Y, is board-foot volume and the auxiliary variable, X, is basal area. (Why basal area, not DBH?) Note also that the population of interest is all marked trees, therefore, the true mean of X, basal area per tree is known and was found to be

$$\mu_x = 1.525\ ft^2$$

The following information is also available from the data:

$$\sum_{i=1}^{35} x_i = 57.248 \qquad \sum_{i=1}^{35} y_i = 7063$$

$$\sum_{i=1}^{35} x_i^2 = 107.373 \qquad \sum_{i=1}^{35} y_i^2 = 1{,}781{,}123 \qquad \sum_{i=1}^{35} x_i y_i = 13{,}722.673$$

Thus, $\hat{R} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \dfrac{7063}{57.248} = 123.375$

And, the total is estimated as

$$\hat{T}_Y = \hat{R}T_x = \hat{R}\left(N\mu_x\right) = 123.375\left(500 \bullet 1.525\right) = 94{,}073\ bd.ft$$

For a reliability estimate, calculate

$$S_u^2 = \frac{\sum_{i=1}^{n} y_i^2 + \hat{R}^2 \sum_{i=1}^{n} x_i^2 - 2\hat{R}\sum_{i=1}^{n} x_i y_i}{n-1}$$

$$= \frac{1{,}781{,}123 + \left(123.375\right)^2\left(107.373\right) - 2\left(123.375\right)\left(13{,}722.673\right)}{34}$$

$$= \ 865.289$$

Which leads to variance and standard error of the total,

$$S_{\hat{T}_Y}^2 = T_x^2 \frac{1}{\mu_x^2} \frac{S_u^2}{n} \left(\frac{N-n}{N}\right)$$

$$= \left(500 \bullet 1.525\right)^2 \frac{1}{\left(1.525\right)^2} \frac{865.289}{35} \left(\frac{500-35}{500}\right) \qquad \text{(var. eq'n [3])}$$

$$= 57747991.214$$

$$S_{\hat{T}_Y} = \sqrt{S_{\hat{T}_Y}^2} = 2397.500$$

Let's say we desired a 90% CI which is found by using $t_{0.10(2),34} = 1.691$ so that

$$\hat{T}_Y \pm t_{0.10(2),34} S_{\hat{T}_Y} \quad \Rightarrow \quad 94{,}073 \pm 1.6911\left(2397.5\right)$$

which, when calculated out becomes

$90\%CI\left(90{,}019;\;\; 98{,}127\right)\; bd.ft$ in trees marked for thinning.

Sample Size Formulas in Ratio Estimation

When coupled with SRS or SYS, determining the number of samples required to meet a desired allowable error is straightforward.

Estimating a population Ratio, *R* itself:

$$n = \frac{t^2 S_u^2 N}{E_R^2 \mu_x^2 N + t^2 S_u^2}$$

where, E_R = is the half-width of the desired Confidence Interval on ratio, *R*

If Ratio estimation is being used because population size is unknown, but large enough so that the f.p.c. (finite population correction) can be ignored, then

$$n = \frac{t^2 S_u^2}{E_R^2 \mu_x^2}$$

Regardless of whether N is known or not, S_u^2 and μ_x must be estimated from previous experience or from a pilot survey.

Estimating the population mean, μ_y, use:

For finite population: $$n = \frac{t^2 S_u^2 N}{E_M^2 N + t^2 S_u^2}$$

For infinite population: $$n = \frac{t^2 S_u^2}{E_M^2}$$

where E_M denotes the half-width of the desired CI on the population mean

Estimating a population total, T_y, use:

For finite population: $$n = \frac{t^2 S_u^2 N^2}{E_T^2 + t^2 N S_u^2}$$

For infinite population: $$n = \frac{t^2 S_u^2 N^2}{E_T^2}$$

where W_T denotes the desired Confidence Interval half-width for the total, T_Y

Example.

Let's say we wanted to estimate the total board-foot volume in the variable intensity thinning to within 2,500 bd.ft at 90% confidence. What sample size is needed?

Since this is about an estimate of the total for a finite population, we use the first formula in the last set given above. For the stand we have the following:

$$S_u^2 = 865.289$$

$$N = 500$$

$$E_T = 2500$$

Guessing first that n is really, really big (infinite) we use a z-value of 1.6449

$$n = \frac{1.6449^2 (865.289) 500^2}{(2500)^2 + 1.6449^2 (865.289) 500} = 78.8 \approx 79$$

Rather than interpolate from the table, rounded up a bit and used 80 d.f.; t = 1.6641

$$n = \frac{1.6641^2 (865.289) 500^2}{(2500)^2 + 1.6641^2 (865.289) 500} = 80.4 \approx 81$$

STOP, because, d.f. = 80 corresponds to n = 81.