Theoretical Underpinning of Optimal Allocation under Stratified Sampling

The theoretical underpinning of optimal allocation under stratified sampling hinges on the famous Cauchy-Schwarz Inequality, as defined as, for $a_i > 0, b_i > 0$,

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right) \times \left(\sum_{i=1}^{n} b_i^2\right).$$

Furthermore, the equality holds if and only if, for $i = 1, 2, \dots, n$,

$$\frac{a_i}{b_i} \equiv Constant.$$

Note that the optimal allocation also depends on what population parameter one wishes to estimate (e.g., population mean), as well as the estimate used to estimate the population parameter, and more importantly, the variance of the estimator. Similarly, the optimal allocation also depends on the cost function. In this class, we assume that the cost function is a linear function in sampling unit costs from each of the strata, i.e., the cost function is written as

$$C = c_0 + \sum_{h=1}^{H} c_h n_h.$$

Case 1: Population average, $\bar{Y}$, is of interest and stratified sample mean, $\bar{y}_{st}$, is used to estimate the population mean.

Note the variance of $\bar{y}_{st}$ is equal to

$$\text{var}(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \times \text{var}(\bar{y}_h) = \frac{1}{N^2} \times \left[\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2 \right].$$

Therefore, combing the variance the cost function, essentially, we want the right-hand side of the following inequality as small as possible,

$$\left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h} \right)^2 \leq \left(\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} \right) \times \left(\sum_{h=1}^{H} c_h n_h \right).$$

Recognizing the inequality is precisely the Cauchy-Schwarz Inequality, the right-hand side will equal to the left-hand side if and only if

$$\frac{N_h S_h \Big/ \sqrt{n_h}}{\sqrt{c_h}\sqrt{n_h}} = \frac{N_h S_h}{\sqrt{c_h} n_h} = Constant.$$

This leads to

$$\frac{n_h}{n} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}}, \text{ or}$$

$$n_h = n \times \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}}.$$

Note: Please see the document on stratified sampling optimal allocation formulas for deciding the overall sample size n, whether when the error

bound B is fixed (as to minimizes the cost) or when the cost C is fixed (as to minimizes the error bound).

Case 2: Now suppose that we have a total of two strata and we are actually interested in estimating the difference in the population means (of some variable) between the two strata, say, $\bar{Y}_1 - \bar{Y}_2$. Naturally, the estimator we will use is the difference between the stratum sample samples, i.e., $\bar{y}_1 - \bar{y}_2$.

Note that we have worked on similar problems (also for determining the sample size) under simple random sample. Let's now look at how a different parameter of interest (and consequently different estimator) will change how the optimal allocation will be done in stratified sampling.

Note that

$$\text{var}(\bar{y}_1 - \bar{y}_2) = \left(\frac{N_1 - n_1}{N_1}\right)\frac{S_1^2}{n_1} + \left(\frac{N_{12} - n_2}{N_2}\right)\frac{S_2^2}{n_2} = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) - \left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right).$$

In the same spirit as in Case 1, we have the following set-up of the Cauchy-Schwarz Inequality,

$$\left(S_1\sqrt{c_1} + S_1\sqrt{c_1}\right)^2 \leq \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) \times (c_1 n_1 + c_2 n_2),$$

where the equality holds if and only if

$$\frac{S_1/\sqrt{n_1}}{\sqrt{c_1}\sqrt{n_1}} = \frac{S_1}{n_1\sqrt{c_1}} = \frac{S_2}{n_2\sqrt{c_2}} = Constant.$$

Equivalently,

$$n_1 = n \times \frac{S_1/\sqrt{c_1}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}, n_2 = n \times \frac{S_2/\sqrt{c_2}}{S_1/\sqrt{c_1} + S_2/\sqrt{c_2}}.$$

As for determining overall sample size n is concerned in this case,

(i) Given a fixed error bound B, the following n minimizes the cost

$$n = \frac{\left(\sum_{h=1}^{2} S_h/\sqrt{c_h}\right) \times \left(\sum_{h=1}^{2} S_h\sqrt{c_h}\right)}{\sum_{h=1}^{2} \frac{S_h^2}{N_h} + \frac{B^2}{4}}.$$

(ii) Given a fixed cost C, the following n minimizes the error bound

$$n = \frac{(C - c_0) \times \left(\sum_{h=1}^{2} S_h/\sqrt{c_h}\right)}{\sum_{h=1}^{2} S_h\sqrt{c_h}}.$$