VIVEKANANDA COLLEGE

THAKURPUKUR

KOLKATA-700063

NAAC ACCREDITED ‘A’ GRADE

Topic: Theory of Stratified Random sampling

Course Title: Sample Survey

Paper: CC-8

Unit: 2

Semester: 4

Name of the Teacher: Riddhi Das Majumder

Name of the Department: Statistics

Stratified Random Sampling:

In stratified sampling the popln of N units is first divided into subpopulations of $N_1, N_2, \ldots, N_L$ units, respectively. These sub-popln are non-overlapping, and together they comprise the whole of the popln., so that $N_1+N_2+\cdots+N_L=N$. The subpopulations are called strata. To obtain the full benefit from stratification, the values of the N_h must be known. When the strata have been determined, a sample is drawn from each, the drawing being made independently in different strata. The sample sizes within the strata are denoted by $n_1, \ldots, n_L$, respectively. If a simple random sample is taken in each stratum, the entire procedure is described as stratified random sampling.

Notations: N: Population size L: Number of strata in the population

N_h: Number of units in the stratum h, $h=1,2,\ldots,L$

Y_{hj}: the value of the j^{th} unit in the stratum h. $j=1(1)N_h$; $h=1(1)L$.

n_h: sample size corresponding to the stratum h, $h=1(1)L$.

Y_h: stratum total of the stratum h, $h=1(1)L$.

$\bar{Y}_h$: stratum mean of the stratum h, $h=1(1)L$.

y_{hj}: the value of the j^{th} sampled unit in the stratum h, $j=1(1)n_h$, $h=1(1)L$.

$\bar{y}_h$: stratum sample mean of the stratum h, $h=1(1)L$.

$S_h^2=\frac{1}{N_h-1}\sum_{j=1}^{N_h}[Y_{hj}-\bar{Y}_h]^2$ is true variance of the stratum h and $s_h^2=\frac{1}{n_h-1}\sum_{j=1}^{n_h}[y_{hj}-\bar{y}_h]^2$ is the sample variance of the stratum 'h'.

Further $W_h=\frac{N_h}{N}$ is the stratum weight and $f_h=\frac{n_h}{N_h}$ is the sampling fraction in the stratum 'h'.

(1) Properties of the Estimators:

For the popln. mean per unit, the estimator used in stratified sampling is $\bar{y}_{st}$, where $\bar{y}_{st}=\frac{\sum_{h=1}^{L} N_h\bar{y}_h}{N}=\sum_{h=1}^{L}W_h\bar{y}_h$.

The estimator $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean, $\bar{y}$, can be written as $\bar{y}$ ...

(2) The estimated variance of $\bar{y}_{st}$

Theorem 1: With stratified random sampling, an unbiased estimator of the variance of $\bar{y}_{st}$ is

$$v(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h-n_h)\frac{s_h^2}{n_h}$$

Proof:

If a simple random sample is taken within each stratum, an unbiased estimator of S_h^2 is $s_h^2 = \frac{1}{(n_h-1)}\sum_{j=1}^{n_h}(y_{hj}-\bar{y}_h)^2$.

Note that $E(v(\bar{y}_{st})) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{n_h}E(s_h^2)$

$$= \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{n_h}S_h^2 = V(\bar{y}_{st})$$

Alternative form for computing purposes is

$$v(\bar{y}_{st}) = \sum_{h=1}^{L}\frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{L}\frac{W_h s_h^2}{N}$$

Allocation of sample size to strata:

In stratified sampling the sample sizes n_h in the h^{th} strata are chosen by the sampler. The allocation of the sample sizes for different strata is done in two popular ways.

(i) Proportional allocation (ii) Optimum allocation.

(i) Proportional allocation:

Under proportional allocation the no. of units to be sampled from the stratum h is made proportional to the stratum size

that is, $n_h \propto N_h$, $h = 1, 2, \ldots, L$

$$\Rightarrow \frac{n_h}{N_h} = c, \forall h. \text{ Now } \sum_{h=1}^{L} n_h = n, \Rightarrow c = \frac{n}{\sum_h N_h} = \frac{n}{N}$$

Hence, $n_h = \frac{n}{N}\cdot N_h$, $h = 1(1)L$.

Under proportional allocation, the $V(\bar{y}_{st})$ is given by

$$V_{prop}(\bar{y}_{st}) = \sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{N^2 n_h}S_h^2 = \frac{1}{N}\sum_{h=1}^{L}\left(\frac{N_h}{n_h}-1\right)\cdot\frac{S_h^2}{N}\cdot N_h$$

$$= \frac{1}{N}\left(\frac{N}{n}-1\right)\sum_{h=1}^{L}\frac{N_h}{N}S_h^2 \quad \left[\because \frac{N_h}{n_h} = \frac{N}{n}, \forall h\right]$$

$$= \frac{N-n}{Nn}\sum_{h=1}^{L}\frac{N_h}{N}S_h^2$$

that two theorems apply to stratified sampling in general and not restricted to stratified random sampling.

If in every stratum the [redacted] estimator $\bar{y}_h$ is unbiased, ~~$\bar{y}_{st}$ is an unbiased estimator of the popln mean $\bar{Y}$.~~

1: $E[\bar{y}_{st}] = E[\sum_{h=1}^{L} W_h \bar{y}_h] = \sum_{h=1}^{L} W_h E(\bar{y}_h) = \sum_{h=1}^{L} W_h \bar{Y}_h = \bar{Y}$

since the estimators are unbiased in the individual strata.

and $\bar{Y} = \sum_{h=1}^{L}\sum_{j=1}^{N_h} Y_{hj}/N = \sum_{h=1}^{L} \frac{N_h \bar{Y}_h}{N} = \sum_{h=1}^{L} W_h \bar{Y}_h$.

Th 2. If the samples are independently drawn from the different strata. $V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$ where $V(\bar{y}_h)$ is the variance of an unbiased estimator $\bar{y}_h$ in the stratum h.

Proof: Since samples are drawn independently from different strata, $Cov(\bar{y}_h, \bar{y}_k) = 0$, $h \neq k$.

Therefore $V(\bar{y}_{st}) = V(\sum_{h=1}^{L} W_h \bar{y}_h)$

$$= \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) + 2\sum_{h<k}\sum Cov(\bar{y}_h, \bar{y}_k) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h).$$

Theorem 3: For stratified random sampling, the variance of the estimator $\bar{y}_{st}$ is $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}(1-f_h)$.

Proof: We know that when a simple random sample size n_h is drawn from a popln (strata), $\bar{y}_h$ is an unbiased for $\bar{Y}_h$, and

$V(\bar{y}_h) = \frac{N_h - n_h}{n_h N_h} S_h^2$.

By Theorem (2), $V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{n_h N_h} S_h^2 \left(\text{or } \sum \frac{W_h^2 S_h^2}{n_h}(1-f_h)\right)$

$$= \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{N_h - n_h}{n_h N_h} S_h^2 = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} S_h^2$$

Corollary: (1) If the sampling fractions $f_h = \frac{n_h}{N_h}$ are negligible in all strata. $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}$.

(2) If $\hat{Y}_{st} = N\bar{y}_{st}$ is the estimator of the popln total Y, then $E[\hat{Y}_{st}] = Y$ and $V[\hat{Y}_{st}] = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$

(ii) Optimum Allocation:

The proportional allocations described above do not take into account any factor other than strata sizes. They completely ignore the internal structure of strata like within stratum variability etc., and hence it ~~is desirable to consider an allocation scheme which takes into~~ account these aspects. A guiding principle in the determination of the n_i's is to choose them as to:

(a) Minimize the variance of the estimator for (i) fixed sample size 'n', and (ii) fixed cost.

(b) Minimize total cost for fixed variance.

Since minimum variance or minimum total cost is an optimal property, the allocation of n_h's to the strata in accordance with the above principles is known as optimum allocation.

Theorem: In stratified random sampling with a linear cost function of the form $C = c_0 + \sum_{h=1}^{L} c_h n_h$ where c_0 is the overhead cost and c_h is the cost per unit in h^{th} stratum, the variance of $\bar{y}_{st}$ is minimum for a given cost C and the cost is a minimum for a specified variance $V(\bar{y}_{st})$, when n_h is proportional to $\frac{W_h S_h}{\sqrt{c_h}}$.

Proof: We have $C = c_0 + \sum_{h=1}^{L} c_h n_h$ and $V = V(\bar{y}_{st}) = \sum_h \frac{W_h^2 S_h^2}{n_h} - \sum_h \frac{W_h^2 S_h^2}{N_h}$

Choosing the n_h to minimize V for fixed C or C for fixed V, are both equivalent to minimizing the product

$$V'C' = \left(V + \sum \frac{W_h^2 S_h^2}{N_h}\right)(C - c_0)$$

$$= \left(\sum_h \frac{W_h^2 S_h^2}{n_h}\right)\left(\sum_h c_h n_h\right)$$

Cauchy-Schwarz inequality: $\left(\sum_h a_h^2\right)\left(\sum_h b_h^2\right) \geq \left(\sum a_h b_h\right)^2$.

Choosing $a_h = \frac{W_h S_h}{\sqrt{n_h}}$, $b_h = \sqrt{c_h n_h}$, we get

$$V'C' = \left(\sum_{h=1}^{L} a_h^2\right)\left(\sum_{h=1}^{L} b_h^2\right) \geq \left(\sum_h a_h b_h\right)^2 = \left(\sum_h W_h S_h \sqrt{c_h}\right)^2.$$

Hence, the minimum value of $V'C'$ is $\left(\sum_h W_h S_h \sqrt{c_h}\right)^2$ and the minimum occurs when $a_h / b_h = \frac{W_h S_h}{n_h \sqrt{c_h}} = \lambda$ (constant), $\forall h$.

$$\Rightarrow n_h = \lambda \cdot \frac{W_h S_h}{\sqrt{c_h}}, \ \forall h.$$

Remark: This theorem leads to the following rules of conduct. In a given stratum, take a larger sample if

(1) the stratum is large

(2) the stratum is more variable internally.

(3) sampling is cheaper in the stratum.

If the cost (C) is fixed, substitute the optimum values of n_h in the cost function: $C = C_0 + \sum_{h=1}^{L} C_h n_h$, we get

$$C - C_0 = \lambda \cdot \sum_{h=1}^{L} C_h \frac{W_h S_h}{\sqrt{C_h}} \Leftrightarrow \lambda = \frac{C - C_0}{\sum_h W_h S_h \sqrt{C_h}}.$$

Then the optimum allocation is

$$n_h = \frac{(C - C_0) \cdot \{W_h S_h / \sqrt{C_h}\}}{\sum_h W_h S_h \sqrt{C_h}}$$

(ii) An important special case arises if $C_h = c$, that is, if the cost per unit is the same in all strata. The cost becomes $C = C_0 + c \cdot \sum_{h=1}^{L} n_h \Leftrightarrow \sum_{h=1}^{L} n_h = \frac{C - C_0}{c}$ and optimum allocation for fixed cost (C) reduces to optimum allocation for fixed sample size, as $\sum_{h=1}^{L} n_h = \frac{C - C_0}{c}$ = constant. Then $n_h \propto W_h S_h, \forall h$

$\Leftrightarrow n_h = \lambda' W_h S_h, \forall h$. If $\sum_{h=1}^{L} n_h = n$ (fixed), then

$\lambda' = \frac{n}{\sum_h W_h S_h}$. Hence, in stratified random sampling $V(\bar{y}_{st})$ is minimised for a fixed total size of sample 'n' if

$$n_h = n \frac{W_h S_h}{\sum_h W_h S_h} = n \cdot \frac{N_h S_h}{\sum_h N_h S_h}.$$ and this allocation is known as Neyman allocation. The formula for minimum variance with fixed 'n' is obtained under the Neyman allocation is

$$V_{Ney}(\bar{y}_{st}) = \frac{\left(\sum_h W_h S_h\right)^2}{n} - \frac{\left(\sum_h W_h S_h^2\right)}{N}.$$ Further if $S_h = S, \forall h$, the optimum values of n_h are given by: $n_h = n \cdot \frac{N_h}{N}$, which is the proportional allocation.

(b) If V is fixed, then substitute the optimum allocation of n_h in

$$V = V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{n_h} - \sum \frac{W_h^2 S_h^2}{N_h}$$ and we get

$$V = \sum_h \frac{W_h S_h \sqrt{C_h}}{\lambda} - \sum \frac{W_h^2 S_h^2}{N_h} \Rightarrow \lambda = \frac{\left(\sum_h W_h S_h \sqrt{C_h}\right)}{V + \frac{1}{N} \sum_h W_h S_h^2}.$$

Hence, $$n_h = \left(\frac{\sum_h W_h S_h \sqrt{C_h}}{V + \frac{1}{N} \sum_h W_h S_h^2}\right) \cdot \frac{W_h S_h}{\sqrt{C_h}}$$

Relative Precision of Stratified Random and Simple Random Sampling: In this section a comparison is made between SRS and stratified random sampling with proportional and optimum allocation.

Theorem: If terms in $1/N_h$ are ignored relative to unity,

$$V_{opt} \le V_{prop} \le V_{ran},$$

where V_{ran}, V_{prop} and V_{opt} be the variances of the estimated mean under simple random sampling (SRS), proportional allocation and optimum allocation for a given sample size (n)

Proof: $V_{ran} = (1-f)\frac{S^2}{n}$, $f = \frac{n}{N}$, and $V_{prop} = \frac{(1-f)}{n}\sum_h W_h S_h^2$

Also, $V_{opt} = \frac{(\sum_h W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}$.

Note that $(N-1)S^2 = \sum_{h=1}^{L}\sum_{j=1}^{N_h}(Y_{hj}-\bar{Y})^2 = \sum_h\sum_j (Y_{hj}-\bar{Y}_h)^2 + \sum_h N_h(\bar{Y}_h-\bar{Y})^2$

$= \sum_h (N_h-1)S_h^2 + \sum_h N_h(\bar{Y}_h-\bar{Y})^2 \longrightarrow (*)$

If the terms $\frac{1}{N_h}$ are ignored (negligible) and hence also in $\frac{1}{N}$, then (*) gives

$$S^2 = \sum W_h S_h^2 + \sum W_h(\bar{Y}_h-\bar{Y})^2$$

Hence, $V_{ran} = (1-f)\frac{S^2}{n} = \frac{1-f}{n}\sum W_h S_h^2 + \frac{1-f}{n}\sum W_h(\bar{Y}_h-\bar{Y})^2$

$= V_{prop} + \frac{1-f}{n}\sum W_h(\bar{Y}_h-\bar{Y})^2 \ge V_{prop}$.

Again, $V_{prop} - V_{opt} = \frac{1}{n}\left[\sum_h W_h S_h^2 - (\sum W_h S_h)^2\right] = \frac{1}{n}\sum_h W_h(S_h-\bar{S})^2$

where $\bar{S} = \sum_h W_h S_h$. $\longrightarrow (**)$

Hence, $\boxed{V_{ran} \ge V_{prop} \ge V_{opt}}$

Remark: From (*) and (**), with terms in $\frac{1}{N_h}$ negligible,

$$V_{ran} = V_{opt} + \frac{1}{n}\sum W_h(S_h-\bar{S})^2 + \left(\frac{1-f}{n}\right)\sum_h W_h(\bar{Y}_h-\bar{Y})^2 \longrightarrow (***)$$

Eqn (*) leads to the following important conclusion: "Greater the difference in the stratum means, greater is the gain in precision of stratified sampling with proportional over simple random sampling"

From (**), we conclude that Neyman's allocation gives better estimates than proportional allocation and greater the difference betn. the stratum SD's, more is the gain in precision of Neyman's allocation over proportional allocation.

The expression (***) leads to the conclusion that we change from SRS to optimum allocation with fixed sample size. Considerable ~~[illegible]~~ gained by forming the strata stratified variance between means and variance between SD's are large

...ration ...
...ange from
...e elimi...
...due to
...stratum...

(21)

-cation (A) there are two steps of decrease in the variance as -range from SRS to optimum allocation. The ~~first~~ step (I) is due to the elimination of differences among the stratum means; the step (II) is due to the elimination of the effect of differences among the stratum S.D's.

Stratified Sampling for Proportions:

If we wish to estimate the proportion $P = \frac{A}{N}$ of units in the pop^n that fall into some defined class C. Let $P_h = \frac{A_h}{N_h}$, $h = 1(1)L$, be the proportions of units in C in the h^{th} stratum.

• Define $Y_{hj} = \begin{cases} 1 & \text{if the } j^{th} \text{ unit of the } h^{th} \text{ stratum belongs } C \\ 0, & \text{otherwise.} \end{cases}$, in the population.

Then $A = \sum_{h=1}^{L}\sum_{j=1}^{N_h} Y_{hj}$ and $A_h = \sum_{j=1}^{N_h} Y_{hj}$; and $P_h = \frac{A_h}{N_h} = \bar{Y}_h$,

$\therefore P = \frac{A}{N} = \bar{Y}$.

Define, in a simple random sample of size n_h from the stratum h, $h = 1(1)L$, $y_{hj} = \begin{cases} 1 & \text{if the } j^{th} \text{ unit of the sample from the } h^{th} \text{ stratum belongs } C \\ 0, & \text{otherwise.} \end{cases}$

clearly, $p_h = \frac{a_h}{n_h} = \sum_{j=1}^{n_h} y_{hj}/n_h = \bar{y}_h$, ~~[illegible]~~, is the proportion of units in C in the sample from the h^{th} stratum.

For the proportion in the whole pop^n, the estimator ~~[illegible]~~ appropriate to stratified random sampling is

$$p_{st} = \frac{\sum_{h=1}^{L} N_h p_h}{N} = \frac{\sum N_h \bar{y}_h}{N} = \bar{y}_{st}$$

Theorem: With stratified random sampling, the mean ~~variance~~ of p_{st} is P and the variance of p_{st} is

$$V(p_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2(N_h - n_h)}{N_h - 1}\frac{P_h Q_h}{n_h}.$$

proof:

Note that $E(y_{hj}) = \sum_{j=1}^{N_h} Y_{hj}\cdot\frac{1}{N} = \bar{Y}_h = P_h$

Then $E(p_h) = \frac{1}{n_h}\sum_{j=1}^{n_h} E(y_{hj}) = \frac{1}{n_h}\cdot n_h \cdot P_h = P_h$ and $E(p_{st}) = \frac{\sum_{h=1}^{L} N_h E(p_h)}{N}$

$= \frac{\sum_{h=1}^{L} N_h P_h}{N} = \sum_{h=1}^{L} A_h / N = \frac{A}{N} = P$ [For SRS of size n_h from the h^{th} stratum, $E(\bar{y}_h) = \bar{Y}_h$ $\Rightarrow$ U.B.]

Proof:

Since $\bar{y}_h$ is the sample mean of a simple random sample of size n_h from the h^{th} stratum, hence $E(\bar{y}_h) = \bar{Y}_h$ and $V(\bar{y}_h) = \frac{N_h - n_h}{n_h N_h} S_h^2$

Hence, $E(p_{st}) = E(\bar{y}_{st}) = \frac{\sum_h N_h E(\bar{y}_h)}{N} = \frac{\sum_h N_h \bar{Y}_h}{N} = \bar{Y} = P.$

and $V(p_{st}) = V(\bar{y}_{st}) = \frac{1}{N^2} \sum_h N_h^2 V(\bar{y}_h) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} S_h^2$

Now, $S_h^2 = \frac{1}{N_h - 1} \left\{ \sum_{j=1}^{N_h} Y_{hj}^2 - N_h \bar{Y}_h^2 \right\} = \frac{1}{N_h - 1} \left\{ A_h - N_h P_h^2 \right\}$

$$= \frac{N_h P_h - N_h P_h^2}{N_h - 1} = \frac{N_h}{N_h - 1} \cdot P_h Q_h.$$

Hence $V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \cdot \frac{P_h Q_h}{n_h}.$

Remark: (1) The terms $\frac{1}{N_h}$ are negligible, then $V(p_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} (1 - f_h) P_h Q_h$ can be used. When the fpc can be ignored,

$$V(p_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} P_h Q_h$$

(2) With proportional allocation, $V(p_{st}) = \frac{N-n}{nN} \cdot \frac{1}{N} \sum_h \frac{N_h^2 P_h Q_h}{N_h - 1}$

$$\simeq \frac{1-f}{n} \cdot \sum_{h=1}^{L} W_h P_h Q_h.$$

(3) The unbiased estimator of $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$ is s_h^2

$$= \frac{1}{n_h - 1} \left\{ \sum_{j=1}^{n_h} y_{hj}^2 - n_h \bar{y}_h^2 \right\} = \frac{1}{n_h - 1} (n_h p_h - n_h p_h^2) = \frac{n_h}{n_h - 1} p_h q_h.$$

Hence, an unbiased estimator of $V(p_{st})$ is $\hat{v}(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h (N_h - n_h)}{n_h - 1} p_h q_h$

$$= \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{p_h q_h}{n_h - 1}$$

Again, $\hat{V}_{prop}(p_{st}) = \frac{1-f}{n} \cdot \frac{1}{N} \sum_{h=1}^{L} N_h \cdot \frac{n_h}{n_h - 1} p_h q_h$

$$= \left(\frac{1-f}{n}\right) \sum_{h=1}^{L} W_h \cdot \frac{n_h}{n_h - 1} p_h q_h$$

(4) Minimum Variance for Fixed Cost:

$$\text{Cost} = c_0 + \sum_h c_h n_h.$$

$$n_h = \left(\frac{N_h S_h / \sqrt{c_h}}{\sum N_h S_h / \sqrt{c_h}} \right) n \simeq \left(\frac{N_h \sqrt{P_h Q_h / c_h}}{\sum_h N_h \sqrt{P_h Q_h / c_h}} \right) n.$$

Neyman's optimum allocation: $n_h \simeq \dfrac{n \cdot N_h \sqrt{P_h Q_h}}{\left(\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}\right)}$

Estimation From a Sample of the "Gain due to Stratificatio[n]"

The data available from the sample are the values of N_h, n_h, $\bar{y}_h$ and s_h^2. The unbiased estimator of $V(\bar{y}_{st})$ is $v(\bar{y}_{st})$
$= \sum_{h=1}^{L} W_h^2 s_h^2 / n_h - \sum_{h} W_h s_h^2 / N$.

Th: Given the results of a stratified random sample, an unbias[ed] estimator of V_{ran}, the variance of the mean of a simple random sample from the same popln. is

$$v_{ran} = \frac{N-n}{n(N-1)}\left[\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right]$$

Proof: $V_{ran} = \dfrac{N-n}{nN} S^2 = \dfrac{N-n}{n(N-1)}\left[\dfrac{1}{N}\sum_h\sum_j Y_{hj}^2 - \bar{Y}^2\right]$.

Note that $E[y_{hj}^2] = \sum_{k=1}^{N_h} Y_{hk}^2 \cdot \frac{1}{N_h}$, $\forall j = 1(1)n_h$, $\forall h = 1(1)L$

$= \frac{1}{N_h}\sum_{k=1}^{N_h} Y_{hk}^2$

and $\frac{1}{N} E\left(\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2\right) = \frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\cdot n_h \cdot \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj}^2 = \frac{1}{N}\sum_h\sum_j Y_{hj}^2$

Also, since $v(\bar{y}_{st})$ and $\bar{y}_{st}$ are unbiased estimators of $V(\bar{y}_{st})$ and $\bar{Y}$, respectively, $E[v(\bar{y}_{st})] = V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - \bar{Y}^2$

$\Rightarrow E[\bar{y}_{st}^2 - v(\bar{y}_{st})] = \bar{Y}^2$. $\Rightarrow$ $\bar{y}_{st}^2 - v(\bar{y}_{st})$ is an unbiased estimator of $\bar{Y}^2$.

Hence $V_{ran} = \dfrac{N-n}{n(N-1)}\left\{\dfrac{1}{N}\sum_h\sum_j Y_{hj}^2 - \bar{Y}^2\right\}$

$$= \frac{N-n}{n(N-1)} E\left[\frac{1}{N}\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right]$$

i.e. $v_{ran} = \dfrac{N-n}{n(N-1)}\left\{\dfrac{1}{N}\sum_{h=1}^{L}\dfrac{N_h}{n_h}\sum_{j=1}^{n_h} y_{hj}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st})\right\}$ is an unbiased estimator of V_{ran}.

*

The gain in efficiency due stratification over simple random sampling is $E - 1 = \dfrac{V_{ran}}{V(\bar{y}_{st})} - 1 = \dfrac{V_{ran} - V(\bar{y}_{st})}{V(\bar{y}_{st})}$ and the estimated gain due to stratification is $\hat{E} - 1 = \dfrac{V_{ran} - V(\bar{y}_{st})}{V(\bar{y}_{st})}$