2 Stratified Random Sampling

2.1 Introduction

Description

The population of size N is divided into *mutually exclusive and exhaustive* subpopulations called *strata* of $N_1, N_2, \cdots, N_L$ units where N_l's are *known*.

A *simple random sample* of size n_l is drawn from the l-th stratum, $l = 1, \cdots, L$.

Total sample size : $\displaystyle\sum_{l=1}^{L} n_l = n$.

N_1	N_2	N_3
n_1	n_2	n_3
SRS_1	SRS_2	SRS_3
$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$

Reasons for Stratification:

1. Allow sub-estimates which can then be combined to give an overall estimate, e.g. district and state income levels.

2. Provide administrative convenience as strata may exist in *natural*, e.g. counties and branches.

3. Allow different sampling fractions and methods in different stratum, e.g. small/large business, private/government housing, urban/rural households.

4. Enable efficient estimates if a heterogeneous population is divided into strata that are *internally homogeneous*.
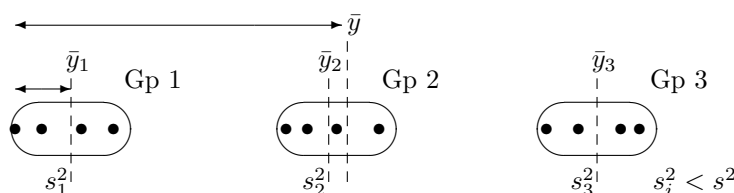
Notation

Population	Sample

Population

N_l

$Y_{li} \quad i = 1, \cdots, N_l$

$$\bar{Y}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} Y_{li}$$

S_l^2

Sample

n_l

$y_{li}, \quad i = 1, \cdots, n_l \quad i\text{-th unit from stratum } h$

$$\bar{y}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} y_{li}$$

s_l^2

Also $W_l = \dfrac{N_l}{N}$ is the stratum weight and $f_l = \dfrac{n_l}{N_l}$ is the sampling fraction for stratum l.

2.2 Estimators of mean and variance

For the population mean $\bar{Y} = \frac{1}{N} \sum_{l=1}^{L} N_l \bar{Y}_l$:

$$\widehat{\bar{Y}} = \bar{y}_{st} = \frac{1}{N} \sum_{l=1}^{L} N_l \bar{y}_l = \sum_{l=1}^{L} W_l \bar{y}_l.$$

For the population total $Y = N\bar{Y}$:

$$\widehat{Y} = N \bar{y}_{st} = \sum_{l=1}^{L} N_l \bar{y}_l = N \sum_{l=1}^{L} W_l \bar{y}_l.$$

Clearly $E(\bar{y}_{st}) = \bar{Y}$ since $E(\bar{y}_l) = \bar{Y}_l$ and the $\bar{y}_l$'s are independent. Hence

$$\text{var}(\widehat{\bar{Y}}) = \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l} \right) \frac{s_l^2}{n_l} = \sum_{l=1}^{L} \frac{W_l^2 \, s_l^2}{n_l} - \frac{1}{N} \sum_{l=1}^{L} W_l s_l^2$$

$$\text{var}(\widehat{Y}) = N^2 \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_l^2}{n_l} = N^2 \sum_{l=1}^{L} \frac{W_l^2 s_l^2}{n_l} - N \sum_{l=1}^{L} W_l s_l^2$$

Estimators and variance estimates for a stratified SRS

Parameter	Estimator	Variance
Ordinary estimator	$s_{yl}^2 = \frac{1}{n_l-1}\big(\sum_{i=1}^{n_l} y_{li}^2 - n_l \bar{y}_l^2 \big), \quad W_l = \frac{N_l}{N}$	
Ratio R	$\widehat{R}_{st} = \frac{1}{\bar{X}} \sum_{l=1}^{L} W_l \bar{y}_l$	$\text{var}(\widehat{R}_{st}) = \frac{1}{\bar{X}^2} \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_{yl}^2}{n_l}$
Mean $\bar{Y}$	$\widehat{\bar{Y}}_{st} = \sum_{l=1}^{L} W_l \bar{y}_l$	$\text{var}(\widehat{\bar{Y}}_{st}) = \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_{yl}^2}{n_l}$
Total Y	$\widehat{Y}_{st} = N \sum_{l=1}^{L} W_l \bar{y}_l$	$\text{var}(\widehat{Y}_{st}) = N^2 \sum_{l=1}^{L} W_l^2 \left(1 - \frac{n_l}{N_l}\right) \frac{s_{yl}^2}{n_l}$

Example (Cochran, 1963) A sample of U.S. colleges was drawn to estimate the enrolments in 1947 from 196 Teachers' Colleges. These colleges were separated into 6 strata. Suppose the following sample information is obtained. Complete the table and estimate total enrollment, its s.e. and coefficient of variation.

l	N_l	n_l	$\bar{y}_l$	$N_l\bar{y}_l$	s_l	$N_l^2(1-\frac{n_l}{N_l})\frac{s_l^2}{n_l}$
1	13	9	523	$13 \times 523 = 6799$	312	$13^2(1-\frac{9}{13})\frac{312^2}{9} = 562432$
2	18	7	324	$18 \times 324 = 5832$	231	$18^2(1-\frac{7}{18})\frac{231^2}{7} = 1509354$
3	26	11	445	$26 \times 445 = 11570$	152	$26^2(1-\frac{11}{26})\frac{152^2}{11} = 819142$
4	42	7	256	$42 \times 256 = 10752$	105	$42^2(1-\frac{7}{42})\frac{105^2}{7} = 2315250$
5	73	14	17	$73 \times 217 = 15841$	92	$73^2(1-\frac{14}{73})\frac{92^2}{14} = 2603889$
6	24	10	135	$24 \times 135 = 3240$	176	$24^2(1-\frac{10}{24})\frac{176^2}{10} = 1040794$
Total	196	58		54034		8850861

Solution: We have $N = 196$ and

$$\widehat{Y}_{st} = \sum_{i=1}^{L} N_l\bar{y}_l = 13 \times 523 + \cdots + 24 \times 135 = 54034$$

$$\text{var}(\widehat{Y}_{st}) = \sum_{i=1}^{L} N_l^2\left(1-\frac{n_l}{N_l}\right)\frac{s_l^2}{n_l}$$

$$= 13^2\left(1-\frac{9}{13}\right)\frac{312^2}{9} + \cdots + 24^2\left(1-\frac{10}{24}\right)\frac{176^2}{10} = 8850861$$

$$\text{sd}(\widehat{Y}_{st}) = \sqrt{8850861} = 2975.03959$$

$$\text{cv}(\widehat{Y}_{st}) = \frac{\text{sd}(\widehat{Y}_{st})}{\widehat{Y}_{st}} = \frac{2975.03959}{54034} = 0.055058659$$

Read Tutorial 10 Q4c and Tutorial 11 Q1a,b.

2.3 Optimal Allocation

We consider

1. Choice of $n_l = nw_l$'s to obtain *maximum precision*, for a *given total sample size* n or, more generally, a given total cost. Here w_l is the desire allocation.

2. For a *specified precision* choice of the minimum n and corresponding n_l's .

1. **Optimization**

   Let K denote a cost function where

   $$C = C_0 + \sum_{l=1}^{L} C_l \, n_l, \qquad C_0 \geq 0, \ C_l > 0, \ l = 1, \cdots, L.$$

   Let

   $$V = \text{Var}(\bar{y}_{st}) = \sum_{l=1}^{L} W_l^2 \frac{S_l^2}{n_l} - \frac{1}{N} \sum_{l=1}^{L} W_l \, S_l^2.$$

   Note that if we take $C_0 = 0, \ C_l = 1, \ l \geq 1$ then C reduces to the total sample size n.

2. **Theorem**

   For a fixed value of C, V is minimized by the choices of n_l where

   $$\boxed{n_l \propto \frac{N_l S_l}{\sqrt{C_l}}, \quad l = 1, \cdots, L.}$$

   The proportionality constant is chosen in accordance with the fixed value of C.

   We should take a larger sample from a statrum if

(1) the stratum size N_l is larger; and/or

(2) the variability in terms of s.d. S_l is greater; and/or

(3) the square root of the cost of sampling a unit $\sqrt{C_l}$ is lower.

3. Best precision for specified total sample size

Optimal allocation:

For fixed total sample size n, precision is greatest if

$$n_l = \frac{N_l S_l / \sqrt{C_l}}{\sum_{l'} N_{l'} S_{l'} / \sqrt{C_{l'}}} \, n = \left(\frac{W_l S_l / \sqrt{C_l}}{\sum_{l'} W_{l'} S_{l'} / \sqrt{C_{l'}}} \right) n, \; l = 1, \cdots, L.$$

Neyman allocation: $C_1 = C_2 = \cdots = C_L$

First determined by Alexander Chuprov (1923), rediscovered by Neyman ($\approx$ 1934).

$$n_l = \frac{N_l S_l}{\sum_{l'} N_{l'} S_{l'}} \, n = \left(\frac{W_l S_l}{\sum_{l'} W_{l'} S_{l'}} \right) n, \qquad l = 1, \cdots, L.$$

Proportional allocation:

If $S_1^2 = S_2^2 = \cdots = S_L^2$, then

$$n_l = \frac{N_l}{\sum_{l'} N_{l'}} \, n = W_l \, n, \qquad l = 1, \cdots, L.$$

4. **Minimal Total Sample Size for Specified Precision**

If we require

$$\Pr\{|\bar{y}_{st} - \bar{Y}| \leq \delta_\mu\} \geq 1 - \alpha,$$

that is

$$\Pr\left\{\frac{|\bar{y}_{st} - \bar{Y}|}{\sqrt{Var(\bar{y}_{st})}} \leq \frac{\delta_\mu}{\sqrt{Var(\bar{y}_{st})}}\right\} \geq 1 - \alpha$$

or

$$\frac{\delta_\mu}{\sqrt{Var(\bar{y}_{st})}} \geq z_{\alpha/2} \quad \text{so} \quad \text{Var}(\bar{y}_{st}) \leq \frac{\delta_\mu^2}{z_{\alpha/2}^2}.$$

For *optimal* allocation, $n_l = n\dfrac{W_l S_l/\sqrt{C_l}}{\sum\limits_{l'=1}^{L} W_{l'}S_{l'}/\sqrt{C_{l'}}}.$

$$
\begin{aligned}
\text{Var}(\bar{y}_{st}) &= \sum_{l=1}^{L} \frac{W_l^2 S_l^2}{n_l} - \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 \qquad \text{(last eqt in P.19)} \\
&= \sum_{l=1}^{L} \frac{(W_l^2 S_l^2)(\sum_{l'} W_{l'}S_{l'}/\sqrt{C_{l'}})}{n W_l S_l/\sqrt{C_l}} - \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 \\
&= \frac{1}{n}\left(\sum_{l=1}^{L} W_l S_l \sqrt{C_l}\right)\left(\sum_{l=1}^{L} \frac{W_l S_l}{\sqrt{C_l}}\right) - \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 \leq \frac{\delta_\mu^2}{z_{\alpha/2}^2}
\end{aligned}
$$

$$
\Rightarrow \frac{1}{n}\left(\sum_{l=1}^{L} W_l S_l \sqrt{C_l}\right)\left(\sum_{l=1}^{L} \frac{W_l S_l}{\sqrt{C_l}}\right) \leq \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 + \frac{\delta_\mu^2}{z_{\alpha/2}^2}
$$

$$
\Rightarrow n/\left(\sum_{l=1}^{L} W_l S_l \sqrt{C_l}\right)\left(\sum_{l=1}^{L} \frac{W_l S_l}{\sqrt{C_l}}\right) \geq \frac{1}{\frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2 + \frac{\delta_\mu^2}{z_{\alpha/2}^2}}
$$

$$
\boxed{n \geq \frac{\left(\sum\limits_{l=1}^{L} \frac{W_l S_l}{\sqrt{C_l}}\right)\left(\sum\limits_{l=1}^{L} W_l S_l \sqrt{C_l}\right)}{V + \frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2} \quad \text{or} \quad n_l = n\frac{\frac{W_l S_l}{\sqrt{C_l}}}{\sum\limits_{l'=1}^{L} \frac{W_{l'}S_{l'}}{\sqrt{C_{l'}}}} \geq \frac{\frac{W_l S_l}{\sqrt{C_l}}\sum\limits_{l'=1}^{L} W_{l'}S_{l'}\sqrt{C_{l'}}}{V + \frac{1}{N}\sum\limits_{l'=1}^{L} W_{l'}S_{l'}^2}.}
$$

where $V = \dfrac{\delta_\mu^2}{z_{\alpha/2}^2}$ is the bound for variance. Ignoring fpc,

$$n \geq \frac{\left(\sum\limits_{l=1}^{L} W_l S_l \sqrt{C_l}\right)\left(\sum\limits_{l=1}^{L} \frac{W_l S_l}{\sqrt{C_l}}\right)}{V} \text{ or } n_l \geq \frac{\frac{W_l S_l}{\sqrt{C_l}} \sum\limits_{l'=1}^{L} W_{l'} S_{l'} \sqrt{C_{l'}}}{V}, \; l = 1, \cdots, L.$$

For *Neyman* allocation, $n_l = n\, W_l S_l / \sum\limits_{l'=1}^{L} W_{l'} S_{l'}$.

$$n \geq \frac{\left(\sum\limits_{l=1}^{L} W_l S_l\right)^2}{\frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2 + V} \text{ or } n_l = n\, \frac{W_l S_l}{\sum\limits_{l'=1}^{L} W_{l'} S_{l'}} \geq \frac{W_l S_l \sum\limits_{l'=1}^{L} W_{l'} S_{l'}}{\frac{1}{N}\sum\limits_{l'=1}^{L} W_{l'} S_{l'}^2 + V}, \; l = 1, \cdots, L.$$

Ignoring fpc, the formula is

$$n \geq \frac{\left(\sum\limits_{l=1}^{L} W_l S_l\right)^2}{V} \text{ or } n_l \geq \frac{W_l S_l \left(\sum\limits_{l'=1}^{L} W_{l'} S_{l'}\right)}{V}, \; l = 1, \cdots, L.$$

For proportional allocation, $n_l = n\, W_l$. Equating S_l fails because S_l in $Var(\bar{y}_{st})$ are not all equal. Instead we should consider

$$\text{Var}(\bar{y}_{st}) \;=\; \sum_{l=1}^{L} \frac{W_l^2 S_l^2}{n_l} - \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 = \sum_{l=1}^{L} \frac{W_l^2 S_l^2}{nW_l} - \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2 \leq V$$

$$n \geq \frac{\sum\limits_{l=1}^{L} W_l S_l^2}{V + \frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2} \text{ or } n_l = n\, W_l \geq \frac{W_l \sum\limits_{l'=1}^{L} W_{l'} S_{l'}^2}{V + \frac{1}{N}\sum\limits_{l'=1}^{L} W_{l'} S_{l'}^2}, \; l = 1, .., L.$$

Sample size determination for a SRS and stratified SRS

SRS for mean $\widehat{\overline{Y}}$	**s.t.** $\Pr(|\bar{y} - \overline{Y}| \leq \delta_\mu) = 0.95$ **and** $S^2 = p(1-p)$ **for prop.** $\widehat{P}$
Sam. n	$n \geq \dfrac{NS^2}{N\delta_\mu^2/z_{\alpha/2}^2 + S^2} \overset{f=0}{\approx} \dfrac{z_{\alpha/2}^2 S^2}{\delta_\mu^2} = \dfrac{S^2}{V}$

Stratified SRS for mean $\widehat{\overline{Y}}$ **s.t.** $\Pr(|\bar{y}_{st} - \overline{Y}| \leq \delta_\mu) = 0.95$ **and** $V = \dfrac{\delta_\mu^2}{z_{\alpha/2}^2}$

Sizes	Optimum	Neyman	Proportional
Strat. n_l	$n_l = n\left(\dfrac{\frac{W_l S_l}{\sqrt{C_l}}}{\sum\limits_{i=1}^{L} \frac{W_i S_i}{\sqrt{C_i}}}\right)$	$n_l = n\left(\dfrac{W_l S_l}{\sum\limits_{i=1}^{L} W_i S_i}\right)$	$n_l = nW_l$
Sam. n	$n \geq \dfrac{\left(\sum\limits_{l=1}^{L} W_l S_l \sqrt{C_l}\right)\left(\sum\limits_{i=1}^{L} \frac{W_i S_i}{\sqrt{C_i}}\right)}{V + \frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2}$	$n \geq \dfrac{\left(\sum\limits_{l=1}^{L} W_l S_l\right)^2}{V + \frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2}$	$n \geq \dfrac{\sum\limits_{l=1}^{L} W_l S_l^2}{V + \frac{1}{N}\sum\limits_{l=1}^{L} W_l S_l^2}$

5. Confidence Intervals for specified total sample size

Provided each n_l is moderately large

$$\frac{\bar{y}_{st} - \mu}{\sqrt{\text{Var}(\bar{y}_{st})}} \sim \mathcal{N}(0,1),$$

at least approximately. Optimal allocation for a given total sample size will give the shortest $(1-\alpha)\,100\%$ CI with length:

$$2\, z_{(1-\frac{\alpha}{2})} \sqrt{\text{Var}\,\bar{y}_{st}} = 2\, z_{(1-\frac{\alpha}{2})} \sqrt{\text{Var}_{opt}\,\bar{y}_{st}}$$

Example (Cochran, 1963) The standard deviation for college enrolments, s, in a given stratum is estimated by the 1943 value.

Stratum	N_l	S_l
1	13	325
2	18	190
3	26	189
4	42	82
5	73	86
6	24	190
	196	

The total enrolment in 1943 was 56,472. For the 1947 study the aim is to obtain a *coefficient of variation* (CV) which is a ratio of standard error to estimate to be approximately 5%. Determine the sample size n required and the Neyman allocation n_l of these values to strata as given previously.

Solution: We have $N = 196$, $\widehat{Y} = 56,472$ and

$$\text{se}(\widehat{Y}) = \text{cv} \times \widehat{Y} = 0.05(56472) = 2824$$
$$\Rightarrow \text{var}(\widehat{Y}) = 2824^2 = 7,974,976 = N^2 V$$

We assume that C_l are all equal. Then

Strat.	N_l	S_l	$N_l S_l$	$N_l S_l^2$	$\omega_l = \frac{N_l S_l}{\sum_{l'} N_{l'} S_{l'}}$	$n_l = \omega_l\, n$	round(n_l)
1	13	325	4225	1373125	0.157	9.13	9
2	18	190	3420	649800	0.127	7.39	7
3	26	189	4914	928746	0.183	10.62	11
4	42	82	3444	282408	0.128	7.44	7
5	73	86	6278	539908	0.234	13.57	14
6	24	190	4560	866400	0.170	9.85	10
Total	196		26841	4640387	1.000	58.00	58

$$n \geq \frac{\left(\sum_{l=1}^{L} W_l S_l\right)^2}{V + \frac{1}{N}\sum_{l=1}^{L} W_l S_l^2} = \frac{\left(\sum_{l=1}^{L} N_l S_l\right)^2}{V N^2 + \sum_{l=1}^{L} N_l S_l^2}$$

$$= \frac{(26841)^2}{7974976 + 4640387} = 57.10809$$

Take $n = 58$.

Read Tutorial 10 Q4a,b and Tutorial 11 Q1e.

2.4 Double sampling for stratification

In stratified SRS, the auxiliary variable X for stratification can be known only after sampling, e.g. income level or education level. In other words, $W_l = \frac{N_l}{N}, l = I, ..., L$ are unknown before sampling.

Double sampling is a *two-phase sampling.* For example, we may call many voters in an opinion poll to identify income level (phase 1 sample), when only a few could be interviewed (phase 2 sample) for purposes of completing a detailed questionnaire.

Phase 1: Take a large sample of size n' for obtaining preliminary information (such as income level of a voter) to identify strata with

$$\widehat{W}_l = \frac{n'_l}{n'}, \quad l = 1, \dots, L$$

denote the proportion of the first sample falling into stratum l. Then $\widehat{W}_l$ is an unbiased estimator of $W_l = \frac{N_l}{N}$.

Phase 2: A much smaller sample of n_l elements are randomly sampled from the n'_l elements identified in stratum l for collecting information on Y. Then the subsample mean and variance, $\bar{y}_l$ and s_l^2, are calculated for each stratum.
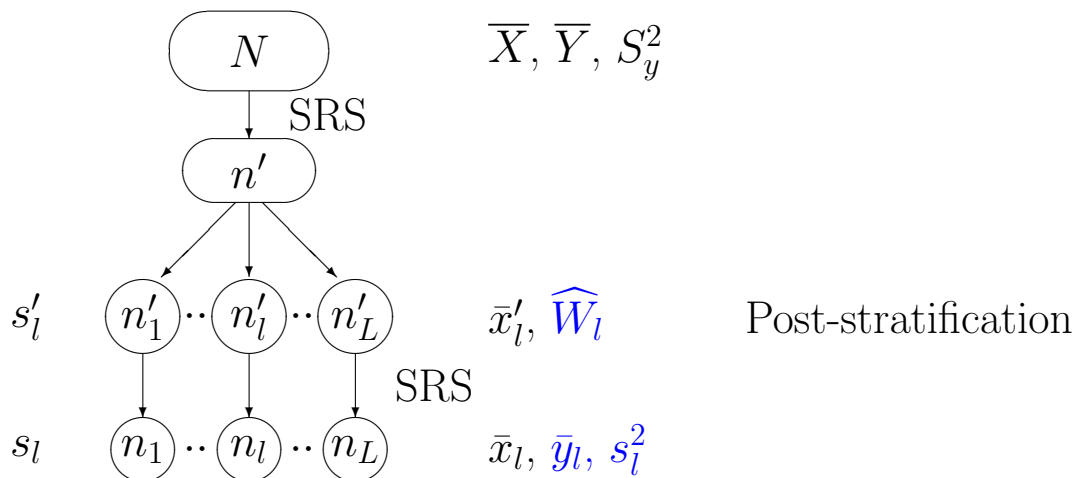
An estimator of the population mean if the phase 2 sampling fractions $\frac{n_l}{N_l}$ are all small and N is large is

$$\widehat{\bar{\bar{Y}}}_{st,ds} = \sum_{l=1}^{L} \widehat{W}_l \bar{y}_l$$

$$\text{var}(\widehat{\bar{\bar{Y}}}_{st,ds}) = \frac{n'}{n'-1} \sum_{l=1}^{L} \left[\left(\widehat{W}_l^2 - \frac{\widehat{W}_l}{n'} \right) \frac{s_l^2}{n_l} + \frac{\widehat{W}_l(\bar{y}_l - \widehat{\bar{\bar{Y}}}_{st,ds})^2}{n'} \right]$$

If n' is so large that $\dfrac{\widehat{W}_l}{n'}$ is negligible, this variance estimate reduces to

$$\text{var}(\widehat{\bar{\bar{Y}}}_{st,ds}) = \sum_{l=1}^{L} \left[\frac{\widehat{W}_l^2 s_l^2}{n_l} + \frac{\widehat{W}_l(\bar{y}_l - \widehat{\bar{\bar{Y}}}_{st,ds})^2}{n'} \right]$$

Ignoring fpc terms, the first part takes the usual form with w'_l replacing $\frac{N_l}{N}$. The second part is the additional variance since we estimate the weights W_l from a sample. It is not necessarily to make the second term small because it means making the $\bar{y}_l$ about equal which is not desirable because stratification is more efficient than SRS when the $\bar{y}_l$ are quite different. Thus, choosing strata that produce different $\bar{y}_l$ still may be better than a SRS, even though double sampling might have to be employed to estimate the W_l.

Example: (College enrollment) From a list of enrollments and faculty sizes for American 4-year colleges and universities, it is desired to estimate the average enrollment (for the 1986-1987 academic year). Private institutions tend to be smaller than public ones, so stratification is in order. While the data indicate the type of college or university, the list is not broken up according to type. A 1-in-10 systematic sample was done to obtain information on type of college and the results are:

Private	Public	Total
$n_1' = 84$	$n_2' = 57$	$n' = 141$

Subsamples of 11 private and 12 public colleges gave the following data on enrollments and faculty size. Estimate the average enrollment for American colleges and universities in 1986-1987.

Private $n_1 = 11$		Public $n_2 = 12$	
Enrollment	Faculty	Enrollment	Faculty
1618	122	7332	452
1140	88	2356	131
1000	65	21879	996
1225	55	935	50
791	79	1293	106
1600	79	5894	326
746	40	8500	506
1701	75	6491	371
701	32	781	108
6918	428	7255	298
1050	110	2136	128
		5380	280
$\bar{y}_1 = 1,680.9$ $s_1 = 1,772.7$		$\bar{y}_2 = 5,852.7$ $s_1 = 5,763.3$	

Solution: From the data given above,

$$\widehat{\overline{Y}}_{st,ds} = \bar{y}_{st,ds} = \sum_{l=1}^{L} \widehat{W}_l \bar{y}_l = \left[\frac{84}{141} \times 1,680.9 + \frac{57}{141} \times 5,852.7 \right]$$

$$= 3,367.4$$

$$\text{var}(\widehat{\overline{Y}}_{st,ds}) = \frac{1}{n_1}(\widehat{W}_1 s_1)^2 + \frac{1}{n_2}(\widehat{W}_2 s_2)^2 +$$

$$\frac{1}{n'}[\widehat{W}_1(\bar{y}_1 - \bar{y}_{st,ds})^2 + \widehat{W}_2(\bar{y}_2 - \bar{y}_{st,ds})^2]$$

$$= \frac{1}{11}\left(\frac{84}{141} \times 1772.7 \right)^2 + \frac{1}{12}\left(\frac{57}{141} \times 5763.3 \right)^2 +$$

$$\frac{1}{141}\left[\frac{84}{141}(1,680.9 - 3,367.4)^2 + \frac{57}{141}(5,852.7 - 3,367.4)^2 \right]$$

$$= 553,748.568 + 29,725.894 = 583,474.463$$

$$\text{se}(\widehat{\overline{Y}}_{st,ds}) = \sqrt{583,474.463} = 763.86$$

The second part of the variance is due to estimating the true stratum weights. The standard deviation of $\widehat{\overline{Y}}_{st,ds}$ is still quite large due to the small sample sizes and large variation among college enrollments, but it is much smaller than the error associated with a single SRS of 23 colleges from the list.

$$\widehat{\overline{Y}}_{srs} = \sum_{l=1}^{L} \frac{n_l}{n} \bar{y}_l = \left[\frac{11}{23} \times 1,680.9 + \frac{12}{23} \times 5,852.7 \right] = 3,857.49$$

$$\text{var}(\widehat{\overline{Y}}_{srs}) = \left(1 - \frac{n}{N} \right)\frac{s^2}{n} = \left(1 - \frac{23}{1410} \right)\frac{22,576,503}{23} = 965,575.4$$

$$\text{se}(\widehat{\overline{Y}}_{srs}) = \sqrt{965,575.4} = 982.637$$

where s is calculated using the combined (1 SRS) sample and $N \simeq 141 \times 10 = 1410$.

Double sampling estimator based on 2-stage sampling

Parameter	Estimator	Variance
Mean $\overline{Y}$	$\widehat{\overline{Y}}_{st,ds} = \displaystyle\sum_{l=1}^{L} w_l \bar{y}_l$ $w_l = \dfrac{n'_l}{n'}$	$\mathrm{var}(\widehat{\overline{Y}}_{st,ds}) = \displaystyle\sum_{l=1}^{L} \left[\dfrac{w_l^2 s_l^2}{n_l} + \dfrac{w_l(\bar{y}_l - \widehat{\overline{Y}}_{st,ds})^2}{n'} \right]$ If n' & N are large and $\dfrac{n_l}{N_l}$ are all small