Stratified Random Samping

Problem 1 ~ A sample of 30 villages is drawn from a total of 300 villages belonging to two districts. The mean and standard deviation of population density of each of the villages are given below:

Districts	No of villages	Mean (μ_j)	Standard deviation (σ_j)
1	200	32	11
2	100	61	42

What are the sample sizes in case of.
i) proportional allocation
and ii) optimum allocation ?
In each case obtain the variance of estimator of the mean population density of all the villages and compare its effeciency with SRSWOR.

Solution ~ We have been provided with the given information:

$k = 2$ (Number of strata).
$n = 30$ (Sample size)
$N_1 = 200$ (Number of villages in district 1)
$N_2 = 100$ (Number of villages in district 2)
$\mu_1 = 32$ (The mean of population density of villages of dis 1)
$\mu_2 = 61$ (The mean of population density of villages of dis 2)
$\sigma_1 = 11$ (standard deviation of pop. density of villag of dist 1)
$\sigma_2 = 42$ (standard deviation of population density of villages of district 2)

i) To find n_1 and n_2 under proportional allocation (sample sizes under proportional allocation)

$$n_i\Big|_{prop} = \frac{N_i}{N} \cdot n.$$

$N = \sum_{i=1}^{2} N_i$, Total number of sampling units in the population.
$n = \sum_{i=1}^{2} n_i$, Total sample size from all the strata.

$$n_1\Big|_{prop} = \frac{200}{300}\times 30 \qquad n_2\Big|_{prop} = \frac{100}{300}\times 30$$

$$\Rightarrow n_1\Big|_{prop} = 20 \qquad n_2\Big|_{prop} = 10.$$

$\therefore$ Sample sizes in cas of proportional allocation is 20 and 10 respectively

ii) Under optimum allocation.

$$n_i\Big|_{opt} = \frac{W_iS_i}{\sum W_iS_i}\cdot n.$$

Here $W_i = \frac{N_i}{N}$ the weight of i th stratum

$$S_i = \sqrt{\frac{N_i}{N_i - 1}}\ \sigma_i.$$

$$W_1 = \frac{200}{300} = \frac{2}{3}, \quad W_2 = \frac{100}{300} = \frac{1}{3}.$$

$$S_1 = \sqrt{\frac{200}{199}}\times 11 = 11.027604,$$

$$S_2 = \sqrt{\frac{100}{99}}\times 42 = 42.211588.$$

Now computing the value of $\sum_{i=1}^{2} W_i S_i = \frac{2}{3}\times 11.027604 + \frac{1}{3}\times 42.211588.$

$$= 21.42226$$

$$n_1\Big|_{opt} = \frac{W_1S_1}{\sum_{i=1}^{2} W_iS_i}\cdot n = \frac{\frac{2}{3}\times 11.027604}{21.42226}\cdot 30 = 10.29546$$

$$\rightarrow n_1\Big|_{opt} = 10.29546 \approx 10$$

$$n_2\Big|_{opt} = \frac{W_2S_2}{\sum_{i=1}^{2} W_iS_i} = \frac{\frac{1}{3}\times 42.211588}{21.42226}\cdot 30 = 19.70454$$

$$\rightarrow n_2\Big|_{opt} = 19.70454 \approx 20$$

$\therefore$ sample sizes in case of optimum allocation is 10 and 20 respectively

We have to obtain the variance of estimator of the mean population density of all villages and compare its efficiency with SRSWOR.

$$\rightarrow \quad Var(\bar{y}_{st})_{prop} = \sum_{i=1}^{2} W_i S_i^2 \left(\frac{1}{n} - \frac{1}{N}\right)$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{2} W_i S_i^2 .$$

$$= \left(\frac{1}{30} - \frac{1}{300}\right) \left[\frac{2}{3} \times (11.027604)^2 + \frac{1}{3}(42.211588)^2 \right]$$

$$= 0.03 \times [675.0114205]$$

$$= 20.25034261.$$

$$\rightarrow \quad Var(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{2} W_i S_i\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{2} W_i S_i^2\right)$$

$$\Rightarrow \quad Var(\bar{y}_{st})_{opt} = \frac{1}{30} \times \left(\frac{2}{3} \times 11.027604 + \frac{1}{3} \times 42.211588\right)^2 - \frac{1}{300}\left(\frac{2}{3} \times (11.027604)^2 + \frac{1}{3} \times (42.211588)^2\right)$$

$$= \frac{1}{30} \times (21.42226)^2 - \frac{1}{300}(675.0114205)$$

$$= 15.29710745 - 2.250038068$$

$$= 13.04707677.$$

So, obtained variance of estimator of the mean population density of all villages using proportional allocation is 20.25034 and using optimum allocation is 13.04708.

<u>Comparing efficiency with respect to SRSWOR</u>

We need to find efficiency E_1 & E_2

when $E_1 = \frac{V_{ran}}{V_{prop}}$ & $E_2 = \frac{V_{ran}}{V_{opt}}$.

$$V_{ran} = \left(1 - \frac{n}{N}\right) \cdot \frac{S^2}{n}$$

We can find S^2 using σ^2

where, $$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2} + \frac{N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}$$

where $d_1 = \mu_1 - \mu$ and $d_2 = \mu_2 - \mu$ $[d_i = \mu_i - \mu,\ i = 1,2]$

and $$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$$

Now, $N_1 = 200$, $N_2 = 100$, $\mu_1 = 32$, $\mu_2 = 61$.

$$\mu = \frac{200 \times 32 + 100 \times 61}{200 + 100}$$

$$\mu = \frac{12500}{300}$$

$$\mu = 41.6667$$

Then $d_1 = \mu_1 - \mu$ $\qquad$ $d_2 = \mu_2 - \mu$

$= 32 - 41.6667$ $\qquad$ $= 61 - 41.6667$

$= -9.6667$ $\qquad$ $= 19.3333$.

Calculating σ^2:

$$\sigma^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2} + \frac{N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}$$

$$\sigma^2 = \frac{200 \times (11)^2 + 100 \times (42)^2}{200 + 100} + \frac{200 \times (-9.6667)^2 + 100(19.3333)^2}{200 + 100}$$

$$\sigma^2 = \frac{200600}{300} + \frac{56066.6667}{300}$$

$$\sigma^2 = 668.6667 + 186.8889$$

$$\sigma^2 = 855.5556$$

Now $S^2 = \frac{N}{N-1}\sigma^2$, on putting values

so $$S^2 = \frac{300}{299} \times 855.5556 = 858.4169454.$$

$$\rightarrow \quad V_{ran} = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = \left(1 - \frac{30}{300}\right)\frac{858.4169454}{30}$$

$$\boxed{V_{ran} = 25.75250836}$$

$$Var(\bar{y}_{st})_{prop} = 20.25034261$$

$$Var(\bar{y}_{st})_{opt} = 13.04707677$$

$$V_{ran} = 25.75250836$$

Now, calculating efficiencies :

$$E_1 = \frac{V_{ran}}{V_{prop}} = \frac{25.75250836}{20.25034216} = 1.27170732$$

$$E_1 = 1.27170732 \approx 1.2717$$

$$E_2 = \frac{V_{ran}}{V_{opt}} = \frac{25.750836}{13.04707677} = 1.973686$$

$$E_2 = 1.973686 \approx 1.9737.$$

Effeciency for proportional allocation with respect to vran SRSWOR is 1.2717.
Efficiency for optimal allocation with respect to vran SRSWOR is 1.9737.
Hence, optimal allocation is more efficient.

Problem 2 ~ In a survey on the area under a crop a total of 186 villages. From each stratum an SRSWOR under proportional allocation was taken and the areas under the ~~area~~ crop in the selected villages were noted. The following are the data obtained from the survey :

Stratum No. (h)	Stratum size (N_h)	Sample size (n_h)	Area under the crop in the sample villages ('00 hectares)
1	72	8	14, 12, 8, 11, 12, 10, 13, 16
2	53	5	27, 20, 21, 22, 30
3	35	4	36, 47, 52, 61
4	26	3	92, 105, 82

Obtain an estimate of the total area under the crop in the district estimate the standard error of the estimator used.

<u>Solution :</u> Let us denote

$\bar{y}_{st} \rightarrow$ Estimate of population mean / mean area under crop.

$N.\bar{y}_{st} \rightarrow$ Estimate of total area under crop.

Firstly, let us calculate the value of $\bar{y}_{st}$ from the provided data.

$$\bar{y}_{st} = \sum_{i=1}^{k} W_i \bar{y}_i$$

Notations:

N : Population size

N_h : Number of units in the stratum h, $h = 1, 2, 3, 4$.

Y_{hj} : The value of the jth unit in stratum h, $j = 1(1)N_h$, $h = 1(1)4$.

n_h : sample size corresponding to the stratum h, $h = 1(1)4$

Y_h : stratum total of the stratum h, $h = 1(1)4$.

$\bar{Y}_h$: stratum mean of the stratum h, $h = 1(1)4$.

y_{hj} : The value of the jth sampled unit in the stratum h.

$\bar{y}_h$: stratum sample mean of the stratum h, $h = 1(1)4$.

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.

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'.

$W_h = \frac{N_h}{N}$ is the stratum weight (hth)

$$\bar{y}_{st} = \sum_{h=1}^{k} W_h \bar{y}_h$$

$$= \frac{72}{186} \times \bar{y}_1 + \frac{53}{186} \times \bar{y}_2 + \frac{35}{186} \times \bar{y}_3 + \frac{26}{186} \bar{y}_4$$

$$\bar{y}_h = \frac{\sum_{h=1}^{n_h} y_h}{n_h}, \quad \bar{y}_1 = \frac{14+12+8+11+12+10+13+16}{8} = \frac{96}{8} = 12$$

$$\bar{y}_2 = \frac{27+20+21+22+30}{5} = \frac{120}{5} = 24$$

$$\bar{y}_3 = \frac{36+47+52+61}{4} = \frac{196}{4} = 49$$

$$\bar{y}_4 = \frac{92+105+82}{3} = \frac{279}{3} = 93$$

$$\bar{y}_{st} = \frac{72}{186} * 12 + \frac{53}{186} \times 24 + \frac{35}{186} \times 49 + \frac{26}{186} \times 93$$

$\bar{y}_{st} = 33.70430108.$, $N\bar{y}_{st} = 186 \times 33.70430108 = 6269.000001$.

★ Obtained estimate of the total area under crop is 6269

After obtaining an estimate of the total area under the crop in the districts, have to estimate of the standard error of the estimator used. Let us make a table depicting values.

Stratum Number (h)	Stratum size (N_h)	sample size (n_h)	$\bar{y}_h$	s_h^2	$W_h^2 s_h^2$	$\frac{W_h^2 s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$
1	72	8	12	6	0.899063475	0.099895941
2	53	5	24	18.5	1.502095618	0.272077696
3	35	4	49	108.6667	3.847748511	0.85200956
4	26	3	93	133	2.598797549	0.766312097

$k = 4$

$$Var(\bar{y}_{st}) = \sum_{h=1}^{k} \frac{W_h^2 S_h^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

an estimate of $var(\bar{y}_{st})$ is $\widehat{var}(\bar{y}_{st}) = \sum_{h=1}^{4} \frac{W_h^2 s_h^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$

$$\widehat{Var}(\bar{y}_{st}) = 0.099895941 + 0.272077696 + 0.85200956 + 0.766312097$$

$$\widehat{Var}(\bar{y}_{st}) = 1.990295294$$

Total estimated area under the crop in district $\hat{Y}_{st} = N\bar{y}_{st}$

$$\widehat{Var}(\hat{Y}_{st}) = N^2 \widehat{Var}(\bar{y}_{st})$$

$$= 186^2 \times 1.990295294 = 68856.25599$$

Estimating the standard error of the estimator used.

$$\widehat{se}(\hat{Y}_{st}) = \sqrt{\widehat{var}(\hat{Y}_{st})}$$

$$= \sqrt{68856.25599} = 262.404756$$

★ so, estimated standard error = 262.404756 of the estimator used.

Problem 3 ~ The following data relate to the number of yearly enrolment in Teachers' Training Colleges stratified into 3 strata. Obtain the gain in precision due to stratification for estimating the average yearly enrolment per college.

Stratum No. (i)	N_i	n_i	$\bar{x}_i$	s_i^2
1	13	9	32.200	2.625
2	18	7	41.638	5.063
3	26	10	19.992	3.549.

Solution ~ The data available from the samples are the values of N_i, n_i, $\bar{y}_i$ (here $\bar{x}_i$) and s_i^2.

· The unbiased estimator of $V(\bar{y}_{st})$ is $v(\bar{x}_{st})$

where $$V(\bar{x}_{st}) = \sum_{i=1}^{K} \frac{W_i^2 S_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

$$v(\bar{x}_{st}) = \widehat{Var}(\bar{x}_{st}) = \sum_{i=1}^{K} \frac{W_i^2 s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

· The unbiased estimator of V_{ran} is v_{ran}.

where, $$V_{ran} = \frac{N-n}{n(N-1)}\left\{\frac{1}{N}\sum_i\sum_j X_{ij}^2 - \bar{X}^2\right\}$$

$$= \frac{N-n}{n(N-1)} E\left[\frac{1}{N}\sum_{i=1}^{L} \frac{N_i}{n_i}\sum_{j=1}^{n_i} x_{ij}^2 - \bar{x}_{st}^2 + v(\bar{x}_{st})\right]$$

i.e. $$v_{ran} = \frac{N-n}{n(N-1)}\left\{\frac{1}{N}\sum_{i=1}^{L}\frac{N_i}{n_i}\sum_{j=1}^{n_i} x_{ij}^2 - \bar{x}_{st}^2 + v(\bar{x}_{st})\right\}$$

is an unbiased estimator of V_{ran}.

Gain in precision due to stratification for estimating the average yearly enrolment per college over simple random sampling.

or The gain in efficiency due to stratification over simple random sampling is $E-1$

$$E - 1 = \frac{v_{ran}}{v(\bar{x}_{st})} - 1 = \frac{v_{ran} - v(\bar{x}_{st})}{v(\bar{x}_{st})}.$$

where,

$$v(\bar{x}_{st}) = \sum_{i=1}^{k} W_i s_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)$$

and.

$$v_{ran} = \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{i=1}^{n_i} x_{ij}^2 - \bar{x}_{st}^2 + v(\bar{x}_{st}) \right]$$

$$s_i^2 = \frac{1}{(n_i-1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$\Rightarrow (n_i - 1) s_i^2 = \sum_{j=1}^{n_i} x_{ij}^2 - n_i \bar{x}_i^2$$

or $$(n_i - 1) s_i^2 + n_i \bar{x}_i^2 = \sum_{j=1}^{n_i} x_{ij}^2$$

$$\therefore \sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}^2 = \sum_{i=1}^{k} \frac{N_i (n_i - 1)}{n_i} s_i^2 + \sum_{i=1}^{k} N_i \bar{x}_i^2 \quad (*)$$

Calculations for gain in precision due to stratification.

<u>Step 1</u>: Finding $\bar{x}_{st} = \sum_{i=1}^{k} W_i \bar{x}_i$

$W_i = \frac{N_i}{N}$, the weight of ith stratum.

$$\bar{x}_{st} = \sum_{i=1}^{3} W_i \bar{x}_i$$

$$= \frac{13}{57} \times 32.200 + \frac{18}{57} \times 41.638 + \frac{26}{57} \times 19.992$$

$$= 29.61185965$$

<u>Step 2</u>: Finding $v(\bar{x}_{st}) = \sum_{i=1}^{k} W_i^2 s_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)$

$$v(\bar{x}_{st}) = \sum_{i=1}^{3} W_i^2 s_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right)$$

$$= \left(\frac{13}{57}\right)^2 \times 2.625 \left(\frac{1}{9} - \frac{1}{13}\right) + \left(\frac{18}{57}\right)^2 \times 5.063 \left(\frac{1}{7} - \frac{1}{18}\right)$$

$$+ \left(\frac{26}{57}\right) \times 3.549 \left(\frac{1}{10} - \frac{1}{18}\right)$$

$$v(\bar{x}_{st}) = 4.668103006 \times 10^{-3} + 0.044078353 + 0.045441181$$
$$= 0.094187637$$

$$v(\bar{x}_{st}) = 0.09419.$$

<u>Step 3</u> : Finding $\sum_{i=1}^{K} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}^2$ from the equation (*)

$$\sum_{i=1}^{3} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{1j}^2 = \sum_{i=1}^{3} \frac{N_i(n_i-1)}{n_i} s_i^2 + \sum_{i=1}^{K=3} N_i \bar{x}_i^2$$

$$\sum_{i=1}^{3} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}^2 = \frac{13(8)}{9} \times 2.625 + 13 \times (32.200)^2$$
$$+ \frac{18(6)}{7} \times 5.063 + 18 \times (41.638)^2$$
$$+ \frac{26(9)}{10} \times 3.549 + 26 \times (19.992)^2.$$

$$\sum_{i=1}^{3} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}^2 = 13509.25333 + 31285.12965 + 10474.72826$$
$$= 55269.11124.$$

<u>Step 4</u> : Finding $v_{ran} = \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum_{i=1}^{K} \frac{N_i}{n_i} \sum_{i=1}^{n_1} x_{ij}^2 - \bar{x}_{st}^2 + v(\bar{x}_{st}) \right]$

$$v_{ran} = \frac{57-26}{26(56)} \left[\frac{1}{57} \times 55269.1124 - (29.61185965)^2 + 0.094187637 \right]$$

$$= 1.977218457. \quad , v_{ran} = 1.977$$

<u>Step 5</u>: Calculating estimated gain in precision : $\frac{v_{ran}}{v(\bar{x}_{st})} - 1$

$$\frac{v_{ran}}{v(\bar{x}_{st})} - 1 = \frac{1.977218457}{0.094187637} - 1$$
$$= 19.9923353.$$

Hence, the gain in precision due to stratification for estimating the average yearly enrolment per college is 19.9923.

Problem 4. Following data show the stratification of all the farms in a country by farm-size as well as the average and standard deviation of acres of corn per farm in each stratum:

Farm size (acres)	Number of farms (N_h)	Average (μ_h)	Standard Deviation (S_h)
- 40	394	5.4	8.3
41 - 80	461	16.3	13.3
81 - 120	390	24.5	15.1
120 - 160	335	34.3	19.8
161 - 200	171	42.1	14.5
201 - 240	115	50.2	5.9
241 -	150	64.0	11.6

For a sample of 100 farms, compute the sample size for each stratum under i) proportional allocation
ii) optimum allocation.

Solution :

i) proportional allocation :

The sample size for each ith stratum under proportional allocation is $n_i\big|_{prop} = W_i \cdot n$

$$= \frac{N_i}{N} \cdot n$$

{W_h is the hth stratum weight} $\left[n_h = \frac{N_h}{N} \cdot n\right]$

$$\cdot\ n_1\big|_{prop} = \frac{394}{2016} \times 100 = 19.54 \approx 20$$

here $n = 100$, $N = 2016$

$n = \sum_{h=1}^{K} n_h$, $N = \sum_{h=1}^{K} N_h$

where N is the total number of sampling units in the population
n is the total sample size from all the strata.

$$\cdot\ n_2\big|_{prop} = \frac{461}{2016} \times 100 = 22.8671 \approx 23$$

- $n_3\big|_{prop} = \frac{390}{2016} \times 100 = 19.3452381 \approx 19$

- $n_4\big|_{prop} = \frac{335}{2016} \times 100 = 16.61706 \approx 17$

- $n_5\big|_{prop} = \frac{171}{2016} \times 100 = 8.482143 \approx 8$

- $n_6\big|_{prop} = \frac{115}{2016} \times 100 = 5.7044 \approx 6$

- $n_7\big|_{prop} = \frac{150}{2016} \times 100 = 7.44047619 \approx 7.$

ii) *Optimum allocation* :

The sample size for each h th starathm under optimum allocation is

$$n_h\big|_{opt} = n \cdot \frac{W_h S_h}{\sum_h W_h S_h} = n \cdot \frac{N_h S_h}{\sum_h N_h S_h}.$$

$$\sum_h N_h S_h = \sum_{h=1}^{7} N_h S_h$$

$$= 394 \times 8.3 + 461 \times 13.3 + 390 \times 15.1 + 335 \times 19.8$$

$$+ 171 \times 14.5 + 115 \times 5.9 + 150 \times 11.6$$

$$= 26821.5$$

- $n_1\big|_{opt} = 100 \times \frac{394 \times 8.3}{26821.5} = 12.19245754 \approx 12$

- $n_2\big|_{opt} = 100 \times \frac{461 \times 13.3}{26821.5} = 22.8596 \approx 23$

- $n_3\big|_{opt} = 100 \times \frac{390 \times 15.1}{26821.5} = 21.95627 \approx 22$

- $n_4\big|_{opt} = \frac{100 \times 335 \times 19.8}{26821.5} = 24.73016 \approx 25$

- $n_5\big|_{opt} = 100 \times \frac{171 \times 14.5}{26821.5} = 9.244449 \approx 9$

- $n_6\big|_{opt} = 100 \times \frac{115 \times 5.9}{26821.5} = 2.529687 \approx 3$

- $n_7\big|_{opt} = 100 \times \frac{150 \times 11.6}{26821.5} = 6.487333 \approx 6$

The computed sample size for each stratum under proportional and optimum allocation can be depicted from the table :

n_i	proportional allocation	optimum allocation
n_1	20	12
n_2	23	23
n_3	19	22
n_4	17	25
n_5	8	9
n_6	6	3
n_7	7	6

Systematic Sampling

Problem 1 ~ Following are the data on number of seedlings in a 80-feet bed :

Bed length in feet							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
26	16	27	37	4	36	20	21
28	9	20	14	5	20	21	26
11	22	25	14	11	43	15	16
16	26	39	24	9	27	14	18
7	17	24	18	25	20	13	11
22	39	25	17	16	21	9	19
44	21	18	14	13	18	25	27
26	14	44	38	22	19	17	29
31	40	55	36	18	24	7	31
26	30	39	29	9	30	30	29

1) Find the variance of the mean of a systematic sample consisting of the seedlings in every 10 feet. Compare this with the variance of sample mean for a SRS of the same size.

Solution : Here we have $k = 10$ and $n = 8$.
Now, each row is a cluster.

Step 1 : Select one of the $k = 10$ rows at random.
Let us select 7th row at random whose observations are 44, 21, 18, 14, 13, 18, 25, 27.

Step 2 : We have to find $\overline{Y}_r$, where r is the random start and $\overline{Y}_r$ is the mean of all the observations in the r'th cluster.

We have selected 7the cluster.

$$\overline{Y}_7 = \frac{44+21+18+14+13+18+25+27}{8}$$

$$= \frac{180}{8}$$

$$= 22.5.$$

Step 3 : We have the selected cluster of size 8. divide this sample of size 8 into $k_1 = 4$ subsambles each of size 2.

Let the four subsamples be:

(44 21) (18 14) (13 18) (25 27)

Step 4 : Let us denote $\overline{Y}_{ir}$ as the mean of the i th subsamble $i = 1(1)k$.

$$\overline{Y}_{1r} = \frac{44+21}{2} = 32.5 \qquad \overline{Y}_{2r} = \frac{18+14}{2} = 16$$

$$\overline{Y}_{3r} = \frac{13+18}{2} = 15.5 \qquad \overline{Y}_{4r} = \frac{25+27}{2} = 26$$

Step 5 : Now, we will be estimating $V(\overline{Y}_{sys})$ by

$$\widehat{V(\overline{Y}_{sys})} \; \frac{1}{k_1(k_1-1)} \sum_{i=1}^{k_1} (\overline{Y}_{ir} - \overline{Y}_r)^2 \qquad (k_1 = 4)$$

$$= \frac{1}{4.3} \left[(32.5-22.5)^2 + (16-22.5)^2 + (15.5-22.5)^2 + (26-22.5)^2 \right]$$

$$= \frac{1}{4.3} [203.5]$$

$$= 16.95833 \quad \text{------*}$$

Step 6 : We have to compare with SRS, For this select 8 observations at random from the population of size 80. Based on the sample and Then find $\overline{y}$ & s_y^2.

26	16	27	37	4	36	20	21
28	9	20	14	5	20	21	26
11	22	25	14	11	43	15	16
16	26	39	24	9	27	14	18
7	17	24	18	25	20	13	11
22	39	25	17	16	21	9	19
44	21	18	14	13	18	25	27
31	40	55	36	18	24	7	31
26	30	39	29	9	3	30	29

We shall Take two - digited numbers from the table of random number. To ensure equal probability for each individual unit we shall take numbers from 01 - 80 (The greatest two digit multiple of 80) and shall ignore the other two digited numbers. We shall divide the number by 80 and take the remainder. The remainder varies from 00 to 79. The remainder 00 will correspond to the 80th observations

Since the sampling is whether SRSWR or SRSWOR, we are not given any information about it so we would consider the sampling without replacement.

Random number (R)	Remainder when divided by 80 R(mod 80)	serial number of the observation selected	corresponding observation
46	46	46	21
52	52	52	14
38	38	38	20
19	19	19	25
84	Rejected		
31	31	31	14
21	21	21	11
50	50	50	39
23	23	23	15

8 selected observations are 21, 14, 20, 25, 14, 11, 39, 15

$$\overline{Y} = \frac{21+14+20+25+14+11+39+15}{8}$$

$$= 19.875$$

$$s_y^2 = \frac{n}{n-1}\sigma_y^2 = \frac{8}{8-1} \times 70.609375$$

$$= 80.69643$$

Step 7 : Find $\widehat{V(\overline{y})} = \frac{s_y^2}{nN}(N-n)$ $\quad (N=80)$

$$= \frac{80.69643}{8 \times 80}(80-8)$$

$$= 9.078348214. \qquad \text{—— **}$$

The variance of the mean of a systematic sample consisting of the seedlings in every 10 feet is 16.95833 which is greater than the variance of sample mean for an SRS of the same size (9.07835)

$$Var(\widehat{\overline{Y}_{sys}}) > Var(\widehat{\overline{y}})_{SRS}.$$

- Precision of systematic sampling method wrt simple random sampling = $\frac{Var(\widehat{\overline{Y}}_{S.R.S})}{Var(\widehat{\overline{Y}})_{sys}} = \frac{9.07835}{16.95833}$

$$= 0.535332583$$

Here, The mean of a simple random sample is more efficient than systematic sampling method as the linear trend is missing.

2) Draw a systematic sample of size $n = 10$ and estimate the population average of the number of seedlings on the basis of the sample drawn.

<u>Solution</u> : Here, $n = 10$, $k = 8$.

Let each column be a cluster i.e 8 clusters.

selecting on cluster at random, let us select 4th cluster with obserbations : 37, 14, 14, 24, 18, 17, 14, 38, 36, 29.

Now, let us find the cluster mean $\overline{Y_r}$

$$= \frac{37+14+14+24+18+17+14+38+36+29}{10}$$

$$= \frac{241}{10}$$

$$= 24.1$$

so, The drawn systematic sample of size = 10 is 37, 14, 14, 24, 18, 17, 14, 38, 36, 29 and the estimate of the population avercarage of the number of seedlings on the basis of the sample drawn is 24.1.

---

3) Draw a circular systematic sample with sampling interval $k = 13$ and sample size $n = 10$ and estimate the population average of the number of seedlings on the basis of sample drawn.

<u>Solution</u> : we have $n = 10$, $k = 13$, Here random start is any number between 1 to 80 then select every 13th observation till a sample of size 10 is formed.

26	16	27	37	4	36	20	21
28	9	20	14	5	20	21	26
11	22	25	14	11	43	15	16
16	26	39	24	9	27	14	18
7	17	24	18	25	20	13	11
22	39	25	17	16	21	9	19
44	21	18	14	13	18	25	27
26	14	44	38	22	19	17	29
31	40	55	36	18	24	7	31
26	30	39	29	9	30	30	29

Let The random start be the 33rd observation
Then 33, 45, 59, 72, 5, 18, 31, 44, 57, 70 are the serial number of the observations selected.
The sample of size 10 is :

7, 21, 44, 31, 4, 22, 14, 17, 26, 24.

$$\overline{Y}_r = \frac{7+21+44+31+4+22+14+17+26+24}{10}$$

where r is the random start.

$$\overline{Y}_r = \frac{210}{10} = 21.$$

so, The drawn circular systematic sample with sampling interval $k=13$ and sample size $n=10$ is 7, 21, 44, 31, 4, 22, 14, 17, 26, 24.

Estimated population average of the number of seedlings on the basis of the sample drawn is 21.

Ratio and Regression Methods

Problem 1 ~ An eye-estimate of the weight of peaches on each tree in an orchard of 200 trees has been done and total weight has been eye-estimated as 12,000 lbs. For a sample of 10 trees the eye-estimated and actual weight of the production of peach has been taken.

S.I No. of trees (sample)	1	2	3	4	5	6	7	8	9	10
Actual weight (lbs)	61	42	50	58	67	45	39	57	71	53
Eye-estimated weight (lbs.)	59	47	52	60	67	48	44	58	76	58

Compute the ratio and regression-estimates of the total actual weight (lbs.) of peaches of all the 200 trees in the orchard and compare the precision of the two estimates.

Solution: Here, let actual weight be Y (Response) and eye-estimated weights be X (auxiliary)

We have been provided with the given information:

$n = 10$

$X = 12{,}000$

$N = 200$

• so, we have the following relation: Ratio estimate

$$\boxed{\hat{Y} = \hat{R} \cdot X} \quad \& \quad \hat{R} = \frac{\bar{y}}{\bar{x}}$$

and we know that the linear regression estimator of the population mean of Y i.e $\bar{Y}$ is given by $\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$

where population regression equation is $Y = \bar{Y} + B(X - \bar{X})$

$B = \rho \frac{S_Y}{S_X}$

The least square estimate of B is $b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

or b can be written as $b = \frac{Cov(y,x)}{Var(x)}$

$$\hat{\bar{Y}}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$$

& $\boxed{\hat{Y}_{lr} = N \cdot \hat{\bar{Y}}_{lr}}$ regression estimate

Now, for computing the values of ratio and regression estimates of the total actual weight (lbs.) of all the 200 trees in the orchard we would first calculate the values of var(x), var(y), Cov(x,y) etc.

For this to calculate, we will make a table containg information

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
59	61	3481	3721	3599
47	42	2209	1764	1974
52	50	2704	2500	2600
60	58	3600	3364	3480
67	67	4489	4489	4489
48	45	2304	2025	2160
44	39	1936	1521	1716
58	57	3364	3249	3306
76	71	5776	5041	5396
58	53	3364	2809	3074
Total $\sum_{i=1}^{10} x_i = 569$	$\sum_{i=1}^{10} y_i = 543$	$\sum_{i=1}^{10} x_i^2 = 3327$	$\sum_{i=1}^{10} y_i^2 = 30483$	$\sum_{i=1}^{10} x_i y_i = 31794$

Now • $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{569}{10} = 56.9$ (From the table)

• $\bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{543}{10} = 54.3$

• $\hat{R} = \frac{\bar{y}}{\bar{x}} = 0.954305799$

$\rightarrow \boxed{\hat{Y} = \hat{R} \cdot X = 0.954305799 \times 12000 = 11451.66959}$

- $Var(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2$

$= \frac{33227}{10} - (56.9)^2$

$= 85.09$

- $Var(y) = \frac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2$

$= \frac{30483}{10} - (54.3)^2$

$= 99.81$

- $Cov(x,y) = E(xy) - E(x)E(y)$

$= \frac{\sum_{i=1}^{10} x_i y_i}{10} - \bar{x}\bar{y}$

$= \frac{31794}{10} - 56.9 \times 54.3$

$= 89.73.$

- $b = \frac{Cov(y,x)}{Var(x)}$

$= \frac{89.73}{85.09}$

$= 1.054530497.$

- $s_x^2 = \frac{n}{n-1} var(x) = \frac{10}{9} \times 85.09 = 94.5444444$

- $s_y^2 = \frac{n}{n-1} var(y) = \frac{10}{9} \times 99.81 = 110.9.$

- $s_{xy} = \frac{n}{n-1} cov(x,y) = \frac{10}{9} \times 89.73 = 99.7.$

so, $\bar{Y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$

$= 54.3 + 1.054530497 \left(\frac{12000}{200} - 56.9\right)$

$= 57.56904454.$

- $\hat{Y}_{lr} = N \cdot \bar{Y}_{lr}$

$= 200 \times 57.56904454$

$= 11513.80891$

$\rightarrow \boxed{\hat{Y}_{lr} = 11513.80891}$

so, The computed ratio estimate of the total actual weight (lbs) of peaches of all the 200 trees in the orchard is 11451.66959.

The computed regression estimate of the total actual weight of peaches of all the 200 trees in the orchard is 11513.80891.

- Now, we have to compute the precisions of the two estimates.

$$MSE(\hat{R}) \simeq \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{X}^2} [S_Y^2 + S_X^2 R^2 - 2R\rho S_X S_Y]$$

$$MSE(\hat{Y}_R) = MSE(X \cdot \hat{R}) \qquad \text{as} \quad \hat{Y} = \hat{R} \cdot X$$

$$= X^2 \, MSE(\hat{R})$$

and $MSE(\hat{R})$ is unbiasedly estimated by $\frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\bar{x}^2}(s_y^2 + s_x^2 \hat{R}^2 - 2\hat{R} s_{xy})$

[estimated by $\left[\left(\frac{1}{n} - \frac{1}{N}\right)(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{xy})\right] / \bar{x}^2$

$$= \frac{\left(\frac{1}{10} - \frac{1}{200}\right)}{(56.9)^2} [110.9 + (0.954305799)^2 \times (94.5444444) - 2 \times (0.954305799) \times 99.7]$$

$$= \frac{0.095}{(56.9)^2} \times [110.9 + 86.1058373 - 190.2885763]$$

$$= \frac{0.095}{(56.9)^2} \times 6.71300743$$

$$= 1.969773091 \times 10^{-4}$$

Now, calculating $\widehat{MSE}(\hat{Y}_R) = X^2 \, MSE(\hat{R})$

$$\widehat{MSE}(\hat{Y}_R) = (12000)^2 \times 1.969773091 \times 10^{-4}$$

$$\boxed{\widehat{MSE}(\hat{Y}_R) = 28364.73252}$$

For calculating $\widehat{MSE}(\hat{Y}_{lr})$, we will have to calculate first $\widehat{MSE}(\bar{y}_{lr})$

$$\widehat{MSE}(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{(n-2)} \sum_{i=1}^{n} [(y_i - \bar{y}) - b(x_i - \bar{x})]^2.$$

Let us make a table containg the information regarding values required for $\widehat{MSE}(\bar{y}_{lr})$

x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$b(x_i-\bar{x})$	$(y_i-\bar{y})-b(x_i-\bar{x})$	$[(y_i-\bar{y})-b(x_i-\bar{x})]^2$
59	61	2.1	6.7	2.214304	4.48569596	20.12140822
47	42	-9.9	-12.3	-10.438862	-1.86113808	3.46384952
52	50	-4.9	-4.3	-5.166709	0.86670944	0.751185245
60	58	3.1	3.7	3.268735	0.43126546	0.1859896
67	67	10.1	12.7	10.649748	2.05025198	4.203533183
48	45	-8.9	-9.3	-9.384431	0.08443142	0.007128665
44	39	-12.9	-15.3	-13.602153	-1.69784659	2.882683039
58	57	1.1	2.7	1.159874	1.54012645	2.371989492
76	71	19.1	16.7	20.139622	-343962249	11.83002892
58	58	1.1	-1.3	1.159874	-2.45987355	6.050977866.

$\Sigma[(y_i-\bar{y})-b(x_i-\bar{x})]^2 = 51.86979$

We got the value $\widehat{MSE}(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{(n-2)} \Sigma[(y_i-\bar{y})-b(x_i-\bar{x})]^2$

$$\widehat{MSE}(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{(n-2)} \times 51.86979$$

$$= \left(\frac{1}{10} - \frac{1}{200}\right)\frac{1}{8} \times 51.86979 = 0.615953756$$

$$\widehat{MSE}(\hat{y}_{lr}) = N^2 \widehat{MSE}(\bar{y}_{lr}) \quad [\text{as } \hat{Y}_{lr} = N \cdot \bar{y}_{lr}]$$

$$= 200^2 \times 0.615953756$$

$$\boxed{\widehat{MSE}(\hat{y}_{lr}) = 24638.15024.}$$

Precision of ratio method with respect to regression estimate

$$= \frac{\widehat{MSE}(\hat{Y}_R)}{\widehat{MSE}(\hat{y}_{lr})} = \frac{28364.73252}{24638.15024} = 1.15125$$

Double Sampling

Problem 1 ~ Following figure relates to a study of a variable Y (in kg.) together with an auxiliary variable x (in ft.):

Population size $(N) = 12908$, $\bar{Y} = 782.5$, $\bar{X} = 88.4$

$S_Y^2 = 45.387$,
$S_x^2 = 39.228$ and
$S_{Yx} = 36.116$.

First - phase sample size $(n') = 1528$ and the sample mean $(\bar{x}') = 85.7$ ft.

Second - phase sample size $(n) = 100$, the sample mean $(\bar{x}) = 86.99$ ft

$\bar{y} = 769.68$ kg and $b = 2.881$.

a) Find an estimate of the population mean of y by ratio method and the variance of the estimator. Also find the relative error of the estimate.

Solution: We have been provided with the information:

$N = 12908$ $\qquad$ $n' = 1528$,

$\bar{Y} = 782.5$ $\qquad$ $\bar{x}' = 85.7$,

$\bar{X} = 88.4$ $\qquad$ $n = 100$,

$S_Y^2 = 45.387$ $\qquad$ $\bar{x} = 86.99$,

$S_x^2 = 39.228$ $\qquad$ $\bar{y} = 769.68$

$S_{xY} = 38.116$ $\qquad$ $b = 2.881$

Firstly we have to find an estimate of the population mean of y by ratio method:

$$\hat{Y}_{Rd} = \frac{\bar{y}}{\bar{x}} \cdot \bar{x}'$$

$$= \frac{769.68}{86.99} \times 85.7$$

$$= 758.2661915$$

an estimate of the population mean of y by ratio method is 758.2661915.

Now calculating the variance of the estimator $\hat{\bar{Y}}_{Rd}$

We have the formula: $MSE - Bias^2 = \hat{Var}$
so, now we need to have the value of $MSE(\hat{\bar{Y}}_{Rd})$ and $Bias(\hat{\bar{Y}}_{Rd})$ in order to get the value of $var(\hat{\bar{Y}}_{Rd})$

- $MSE(\hat{\bar{Y}}_{Rd}) \simeq \left(\frac{1}{n'} - \frac{1}{N}\right)S_Y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(S_Y^2 + R^2S_x^2 - 2R\rho S_x S_y)$

$$MSE(\hat{\bar{Y}}_{Rd}) \simeq \left(\frac{1}{1528} - \frac{1}{12908}\right)45.387 + \left(\frac{1}{100} - \frac{1}{1528}\right) \times \left(45.387 + R^2 \times 39.228 - 2R\rho S_x S_y\right)$$

$$\rho = \frac{S_{xy}}{S_x S_y} \Rightarrow \rho S_x S_y = S_{xy} = 36.116$$

$$R = \frac{\bar{Y}}{\bar{X}} = \frac{782.5}{88.4} = 8.851809955$$

$$MSE(\hat{\bar{Y}}_{Rd}) \simeq \left(\frac{1}{1528} - \frac{1}{12908}\right)45.387 + \left(\frac{1}{100} - \frac{1}{1528}\right) \times (45.387 + (8.851809955)^2 \times 39.228 - 2 \times (8.851809955) \times (36.116))$$

$$= 45.387 \times 5.769789262 \times 10^{-4} + 23.17411238$$
$$= 23.20029972$$

$\therefore MSE(\hat{\bar{Y}}_{Rd}) = 23.20029972$

- $Bias(\hat{\bar{Y}}_{Rd}) \simeq \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)(C_x^2 - \rho C_x C_y)$

For this we have to calculate C_X, C_Y, ρ first

$$C_X = \frac{S_X}{\bar{X}} = \frac{\sqrt{39.228}}{88.4} = 0.070850972$$

$$C_Y = \frac{S_Y}{\bar{Y}} = \frac{\sqrt{45.387}}{782.5} = 8.609568636 \times 10^{-3}$$

$$\rho = \frac{S_{xy}}{S_x S_y} \Rightarrow \rho = \frac{36.116}{\sqrt{39.228}\sqrt{45.387}} \Rightarrow \rho = 0.855925218$$

$$Bias \simeq \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)(C_x^2 - \rho C_x C_Y)$$

$$= 782.5\left(\frac{1}{100}-\frac{1}{1528}\right)(0.070850972^2 - 0.865925218 \times 0.070850972 \times 8.60956863 6\times10^{-3})$$

$$= 0.032891555.$$

$$Var(\hat{\bar{Y}}_{Rd}) = MSE(\hat{\bar{Y}}_{Rd}) - (Bias(\hat{\bar{Y}}_{Rd}))^2$$

$$= 23.20029972 - (0.032891555)^2$$

$$= 23.19921787.$$

The variance of the estimator is 23.19921787. $\approx$ 23.1992

Now to find the relative error of the estimate

We have : Relative error $= \frac{|\hat{\bar{Y}}_{Rd} - \bar{Y}|}{\bar{Y}}$

$$= \frac{|758.2661915 - 782.5|}{782.5}$$

$$= 0.030969723$$

Relative error $= 0.03097$.

so, The relative error of the estimate is 0.03097.

---

b) Find an estimate of the population mean of y by the regression method and the variance of the estimator. Also find the relative error of the estimate.

Solution an estimate of the population mean of y by the regression method is given by

$$\hat{\bar{Y}}_{rd} = \bar{y} - b(\bar{x}_n - \bar{x}_n')$$

$$= 769.68 - 2.881(86.99 - 85.7)$$

$$= 765.96351.$$

so $\hat{\bar{Y}}_{rd} = 765.96351.$

Now, To calculate the value of the variance of the estimator $\hat{\bar{Y}}_{rd}$ is given by the following formula.

$$Var(\hat{\bar{Y}}_{rd}) \simeq \frac{(1-\rho^2)S_Y^2}{n} + \frac{S_Y^2 \rho^2}{n'} - \frac{S_Y^2}{N}.$$

$$= \frac{(1-(0.855925218)^2)S_Y^2}{100} + \frac{45.387 \times (0.855925218)^2}{1528}$$

$$- \frac{45.387}{12908}.$$

$$= 0.139606071$$

$$Var(\hat{\bar{Y}}_{rd}) = 0.139606$$

so the variance of the estimator is 0.139606.

Now, to find the relative error of the estimate.

$$\text{relative error} = \frac{|\hat{\bar{Y}}_{rd} - \bar{Y}|}{\bar{Y}}$$

$$= \frac{|765.97351 - 782.5|}{782.5}$$

$$= 0.021120115.$$

The relative error of the estimate is 0.02112.

c) Hence find a relative measure of precision of one method with respect to the other.

<u>Solution</u> The relative measure of precision of ratio method with respect to the regression method.

$$= \frac{Var(\hat{\bar{Y}}_{rd})}{Var(\hat{\bar{Y}}_{Rd})}$$

$$= \frac{0.139606071}{23.19921787} = 6.017705932 \times 10^{-3}$$

$$= 0.006017706$$

The relative measure of precision of ratio method with respect to the regression model is 0.00602.