CHAPTER 7

Stratified Sampling

7.1 INTRODUCTION

The efficiency of an estimator can be increased by increasing the sample size. It is not wise to select a sample from a large population at a time because it is seldom representative in the sense that the sample units may be distributed unevenly over the population. However, the efficiency can also be increased greatly by dividing the population into homogeneous groups or layers (strata) and then selecting samples from each of the groups separately. Stratified sampling is commonly used in large-scale surveys. In this chapter we will discuss in detail various aspects of stratified sampling.

7.2 DEFINITION OF STRATIFIED SAMPLING

In stratified sampling, the entire population U of N units is divided into a number (K) of mutually exclusive and exhaustive groups, which are called strata. The ith stratum U_i consists of N_i units with $U = \bigcup_{i=1}^{K} U_i$ and $\sum_{i=1}^{K} N_i = N$. From each of the strata, samples of suitable sizes n_i's are selected independently by some suitable sampling design. The sampling procedure for the different strata need not be the same and may depend on the available information about the respective stratum. For example, suppose we want to select a sample of 50,000 households from South Africa to estimate the total number of unemployed persons in South Africa. In this situation, one may divide South Africa into a few mutually exclusive and exhaustive zones such as provinces (strata), and from each of the provinces, samples of households may be selected independently. Suppose further that the number of unemployed persons (x) in the households of some of the provinces is available from the past surveys. Then, using the x values as an auxiliary variable, one can select samples by using the inclusion probability proportional to size (IPPS) sampling procedure for those strata, where x values are available. Thus for stratified sampling, one should carefully consider the problem of forming strata and its number, sampling procedures for different strata, and allocation of sample sizes to the respective stratum.

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00007-8

7.3 ADVANTAGES OF STRATIFIED SAMPLING

Stratified sampling is used in most large-scale surveys because of its various advantages, some of which are described below:

(i) Estimation of subpopulations: In cases where the estimates of the population characteristics are needed not only for the entire population but also for its different subpopulations, one should treat such subpopulations as strata. For example, in a national unemployment survey, the government may be interested in estimating unemployment figures for the entire country as well as at provincial levels. In this case, each province can be taken as a stratum.

(ii) Administrative convenience: The agency conducting the survey may stratify the population such that the survey can be supervised in an efficient manner, e.g., the agency can appoint separate supervisors to conduct survey for each of the strata separately.

(iii) Representativeness of sample: In stratified sampling, formation of strata and allocation of samples to different strata may de done in such a way that the sample can represent the population with respect to the characteristics under study. For instance, if we want to select a sample of students from a school, which represents the different races of South Africa, a simple random sampling without replacement (SRSWOR) sample from the entire school may not be representative. In this situation, a stratified sampling using different racial groups as strata is expected to provide a more representative sample than an SRSWOR sample from the entire school.

(iv) Efficiency: Stratification may increase efficiency of the estimates by forming strata in such a way that each stratum becomes homogeneous with respect to the characteristic under study. Suitable sampling schemes to the respective strata may increase efficiencies of the estimators.

(v) Improved quality of data: Improved quality of data may be obtained by employing different types of investigators to different strata. For example, investigators knowing local languages may be deployed to the rural areas, whereas in urban areas investigators knowing English may be more advantageous.

7.4 ESTIMATION PROCEDURE

Consider a population U consisting of N units, which has been divided into K strata and the ith stratum U_i consists of N_i units and $\sum_{i=1}^{K} N_i = N$. Suppose from the ith stratum a sample s_i of size n_i is selected by some suitable

sampling procedure. Samples from each of the stratum are selected independently. Here we will use the following notations:

y_{ij} = value of the character y under study for the jth unit of the ith stratum, $j = 1,\ldots, N_i$, $i = 1,\ldots, K$; $Y_i = \sum_{j=1}^{N_i} y_{ij} = i$th stratum total; $\overline{Y}_i = Y_i/N_i = i$th stratum mean, $Y = \sum_{i=1}^{K} Y_i$ = population total; $\overline{Y} = \sum_{i=1}^{K}\sum_{j=1}^{N_i} y_{ij}/N = \sum_{i=1}^{K} W_i\overline{Y}_i$ = population mean, $W_i = N_i/N = i$th stratum weight; $S_{yi}^2 = \sum_{j=1}^{N_i}\left(y_{ij} - \overline{Y}_i\right)^2 \Big/ (N_i - 1) = i$th stratum variance, $S_y^2 = \sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(y_{ij} - \overline{Y}\right)^2 \Big/ (N - 1)$ = population variance

	Stratum				
	1	…	i	…	K
	y_{11}	…	y_{i1}	…	y_{K1}
	…	…	…	…	…
	y_{1j}	…	y_{ij}	…	y_{Kj}
	…	…	…	…	…
	y_{1N_1}	…	y_{iN_i}	…	y_{KN_K}
Stratum size	N_1	…	N_i	…	N_K
Stratum weight	W_1	…	W_i	…	W_K
Stratum total	Y_1	…	Y_i	…	Y_K
Stratum mean	$\overline{Y}_1$	…	$\overline{Y}_i$	…	$\overline{Y}_K$
Stratum variance	S_{y1}^2	…	S_{yi}^2	…	S_{yK}^2

7.4.1 Estimation of Population Mean

Let the sample s_i of size n_i be selected from the stratum U_i with probability $p(s_i)$ by using a sampling design $p^{(i)}$. Let $\pi_{j|i}(>0)$ and $\pi_{jk|i}(>0)$ be the inclusion probabilities for the jth unit and jth and kth ($j \neq k$) for the ith stratum, respectively. An unbiased estimator for the ith stratum mean $\overline{Y}_i$ based on the selected sample s_i is given by

$$\widehat{\overline{Y}}_i = \sum_{j \in s_i} b_j(s_i) y_{ij} \tag{7.4.1}$$

where $b_j(s_i)$'s are constants free from y_{ij}'s and satisfy the unbiasedness condition

$$\sum_{s_i \supset j} b_j(s_i) p(s_i) = 1/N_i \text{ for } \forall j \in U_i; \quad i = 1, \ldots, N.$$

Theorem 7.4.1

(i) $\widehat{\overline{Y}}_{st} = \sum_{i=1}^{K} W_i \widehat{\overline{Y}}_i$ is an unbiased estimator for mean $\overline{Y}$

(ii) $V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2 Q_i,$

where

$Q_i = V\left(\widehat{Y}_i\right) = \sum_{j \in U_i} \alpha_{j|i} \gamma_{ij}^2 + \sum_{j \neq} \sum_{k \in U_i} \alpha_{jk|i} \gamma_{ij} \gamma_{ik},$

$\alpha_{j|i} = \sum_{s_i \supset j} b_j^2(s_i) p(s_i) - 1/N_i^2$ and $\alpha_{jk|i} = \sum_{s_i \supset j,k} b_j(s_i) b_k(s_i) p(s_i) - 1/N_i^2$

(iii) An unbiased estimator of $V\left(\widehat{\overline{Y}}\right)$ is

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2 \widehat{Q}_i,$$

where $\widehat{Q}_i = \sum_{j \in s_i} \frac{\alpha_{j|i}}{\pi_{j|i}} \gamma_{ij}^2 + \sum_{j \neq} \sum_{k \in s_i} \frac{\alpha_{jk|i}}{\pi_{jk|i}} \gamma_{ij} \gamma_{ik}$

Proof

(i) $E\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i \, E\left(\widehat{\overline{Y}}_i\right) = \sum_{i=1}^{K} W_i Y_i = \overline{Y}$

(ii) Since samples are selected independently from each stratum, we get

$$V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2 \; V\left(\widehat{\overline{Y}}_i\right),$$

$$\text{Now} \quad V\left(\widehat{\overline{Y}}_i\right) = \sum_{s_i} \left(\sum_{j \in s_i} b_j(s_i) \gamma_{ij} \right)^2 p(s_i) - \overline{Y}_i^2$$

$$= \sum_{j \in U_i} \gamma_{ij}^2 \left(\sum_{s_i \supset j} b_j^2(s_i) p(s_i) \right)$$

$$+ \sum_{j \neq} \sum_{k \in U_i} \gamma_{ij} \gamma_{ik} \left(\sum_{s_i \supset j,k} b_j(s_i) b_k(s_i) p(s_i) \right) - \overline{Y}_i^2 = Q_i$$

$$\text{(iii)} \quad E\left[\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) \right] = \sum_{i=1}^{K} W_i^2 E\left(\widehat{Q}_i\right)$$

$$= \sum_{i=1}^{K} W_i^2 Q_i = V\left(\widehat{\overline{Y}}_{st}\right)$$

7.4.1.1 Arbitrary Fixed Sample Size Design

Let s_i's be selected by some fixed effective size sampling design of size $n_i(FESD(n_i))$ and $b_j(s_i) = \frac{1}{N_i \pi_{j|i}}$. Then, $\widehat{\overline{Y}}_i = \sum_{j \in s_i} \frac{y_{ij}}{N_i \pi_{j|i}}$ and $\overline{Y}_{st} = \sum_{i=1}^{k} W_i \sum_{j \in s_i} \frac{y_{ij}}{N_i \pi_{j|i}}$

$$Q_i = V\left(\widehat{\overline{Y}}_i\right) = \frac{V\left(\widehat{Y}_i\right)}{N_i^2} = \frac{1}{N_i^2}\left[\sum_{j<} \sum_{k \in U_i} \Delta_{jk|i}\left(\frac{y_{ij}}{\pi_{j|i}} - \frac{y_{ik}}{\pi_{k|i}}\right)^2\right]$$

where $\Delta_{jk|i} = \pi_{j|i}\pi_{k|i} - \pi_{jk|i}$.

$$V\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{N^2} \sum_{i=1}^{k} \sum_{j<} \sum_{k \in U_i} \Delta_{jk}\left(\frac{y_{ij}}{\pi_{j|i}} - \frac{y_{ik}}{\pi_{k|i}}\right)^2$$

Since Q_i can be estimated unbiasedly by any of the following

$$\widehat{Q}_i(1) = \frac{1}{N_i^2}\left[\sum_{j \in s_i} \frac{1}{\pi_{j|i}}\left(\frac{1}{\pi_{j|i}} - 1\right) y_{ij}^2 + \sum_{j \neq} \sum_{k \in s_i} \frac{1}{\pi_{jk|i}}\left(\frac{\pi_{jk|i}}{\pi_{j|i}\pi_{k|i}} - 1\right) y_{ij} y_{ik}\right] \text{ and}$$

$$\widehat{Q}_i(2) = \frac{1}{N_i^2} \sum_{j<} \sum_{k \in s_i} \frac{\Delta_{jk|i}}{\pi_{jk|i}}\left(\frac{y_{ij}}{\pi_{j|i}} - \frac{y_{ik}}{\pi_{k|i}}\right)^2,$$

the variance $V\left(\widehat{\overline{Y}}_{st}\right)$ can be estimated unbiasedly by any of the following estimators:

$$\sum_{i=1}^{K} W_i^2 \widehat{Q}_i(1) \text{ and } \sum_{i=1}^{K} W_i^2 \widehat{Q}_i(2).$$

7.4.1.2 Simple Random Sampling Without Replacement

For SRSWOR, $\pi_{j|i} = n_i/N_i$ and $\pi_{jk|i} = n_i(n_i - 1)/\{N_i(N_i - 1)\}$. Substituting the values of $\pi_{j|i}$ and $\pi_{jk|i}$ in Section 7.4.1.1, we get

$$\widehat{\overline{Y}}_i = \overline{y}(s_i) = \sum_{j \in s_i} y_{ij} \Big/ n_i = \text{sample mean for the } i\text{th stratum},$$

$$\widehat{\overline{Y}}_{st} = \sum_{i=1}^{K} W_i\, \overline{y}(s_i),$$

$$Q_i = V(\overline{y}(s_i)) = (1 - f_i) S_{yi}^2 \Big/ n_i,\ f_i = n_i/N_i,\ \widehat{Q}_i(1) = \widehat{Q}_i(2) = \widehat{V}(\overline{y}(s_i))$$

$$= (1 - f_i) s_{yi}^2 \Big/ n_i,\ V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2 (1 - f_i) S_{yi}^2 \Big/ n_i \text{ and}$$

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2 (1 - f_i) s_{yi}^2 \Big/ n_i$$

where
$S_{yi}^2 = \sum_{j\in U_i}\left(y_{ij}-\overline{Y}_i\right)^2\Big/(N_i-1)$ and $s_{yi}^2 = \sum_{j\in s_i}\left(y_{ij}-\overline{y}(s_i)\right)^2\Big/(n_i-1)$.

7.4.1.3 Probability Proportional to Size With Replacement Sampling

In case, s_i's are selected by probability proportional to size with replacement (PPSWR) method of sampling using $p_{j|i}\left(\sum_{j\in U_i} p_{j|i}=1\right)$ as the normed size measure for the jth unit of the ith stratum, we may take $b_j(s_i) = n_j(s_i)/(N_i p_{j|i})$, where $n_j(s_i)$ = number of times jth unit appears in the sample s_i. In this situation we have

$$\widehat{\overline{Y}}_i = \frac{1}{n_i N_i}\sum_j n_j(s_i)\frac{y_{ij}}{p_{j|i}},\ Q_i = V\left(\widehat{\overline{Y}}_i\right) = V_{pps|i}\Big/\left(N_i^2 n_i\right) \text{ and}$$

$$\widehat{Q}_i = \frac{1}{N_i^2\, n_i(n_i-1)}\sum_j n_j(s_i)\left(\frac{y_{ij}}{p_{j|i}} - N_i\widehat{\overline{Y}}_i\right)^2$$

where $V_{pps|i} = \sum_{j\in U_i} p_{j|i}\left(\frac{y_{ij}}{p_{j|i}} - Y_i\right)^2$. The expressions for $\widehat{\overline{Y}}_{st}$, $V\left(\widehat{\overline{Y}}_{st}\right)$, and $\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)$ come out respectively as $\widehat{\overline{Y}}_{st} = \frac{1}{N}\sum_{i=1}^{K}\left(\frac{1}{n_i}\sum_j n_j(s_i)\frac{y_{ij}}{p_{j|i}}\right)$,

$V\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{N^2}\sum_{i=1}^{K}\frac{V_{pps}(i)}{n_i}$, and

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{N^2}\sum_{i=1}^{K}\frac{1}{n_i(n_i-1)}\sum_j n_j(s_i)\left(\frac{y_{ij}}{p_{j|i}} - N_i\widehat{\overline{Y}}_i\right)^2.$$

7.4.1.4 Simple Random Sampling With Replacement

The PPSWR sampling reduces to simple random sampling with replacement (SRSWR) sampling when $p_{j|i} = 1/N_i\ \forall\ j\in U_i$. So, substituting $p_{j|i} = 1/N_i$ in Section 7.4.1.3, we get $\widehat{\overline{Y}}_i = \sum_j n_j(s_i)y_{ij}/n_i = \widetilde{\overline{y}}_i$ = sample mean based on n_i units with repetition, $Q_i = \sigma_{yi}^2/n_i$, $\widehat{Q}_i = \widehat{\sigma}_{yi}^2/n_i$, $\widehat{\overline{Y}}_{st} = \sum_{i=1}^{K} W_i\widetilde{\overline{y}}_i$, $V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2\frac{\sigma_{yi}^2}{n_i}$, and $\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} W_i^2\widehat{\sigma}_{yi}^2/n_i$

where $\sigma_{yi}^2 = \frac{N_i-1}{N_i}S_{yi}^2$ and $\widehat{\sigma}_{yi}^2 = \widetilde{s}_{yi}^2 = \frac{\sum_j n_j(s_i)\left(y_{ij}-\widetilde{\overline{y}}_i\right)^2}{n_i-1}$.

7.4.2 Estimation of Population Proportion

Let π_A denote the proportion of the individuals that possess a certain attribute or group A such as HIV +ve. In this situation we write $y_{ij} = 1$ if jth individual of the ith stratum possesses attribute A and $y_{ij} = 0$ otherwise. Let the total number of individuals belonging to A in the ith stratum and the entire population be denoted, respectively, by $N_{iA} = \sum_{j=1}^{N_i} y_{ij}$ and $N_A = \sum_{i=1}^{K}\sum_{j=1}^{N_i} y_{ij}$. The proportion of individuals that possess the attribute A in the ith stratum and the entire population are respectively denoted by $\pi_{iA} = N_{iA}/N_i$ and $\pi_A = N_A/N = \sum_{i=1}^{K} N_{iA}/N = \sum_{i=1}^{K} W_i\,\pi_{iA}$.

7.4.2.1 Simple Random Sampling Without Replacement

Suppose samples s_i's of sizes n_i's are selected independently from each of the strata by the SRSWOR method and let n_{iA} be the number of individuals belong to A. Then from Section 7.4.1.2 we get the following theorem.

Theorem 7.4.2

(i) $\widehat{\pi}_A = \sum_{i=1}^{K} W_i\,\widehat{\pi}_{iA}$ is an unbiased estimator for mean π_A where. $\widehat{\pi}_{iA} = n_{iA}/n$

(ii) $V(\widehat{\pi}_A) = \sum_{i=1}^{K} W_i^2(1-f_i)\dfrac{N_i}{N_i - 1}\dfrac{\pi_{iA}(1-\pi_{iA})}{n_i}$ with $f_i = n_i/n$

(iii) An unbiased estimator of $V(\widehat{\pi}_A)$ is

$$\widehat{V}(\widehat{\pi}_A) = \sum_{i=1}^{K} W_i^2(1-f_i)\frac{\widehat{\pi}_{iA}(1-\widehat{\pi}_{iA})}{n_i - 1}$$

Proof

The theorem can be proved easily from Section 7.4.1.2 by substituting $y_{ij} = 1$ for $i \in A$ and $y_{ij} = 0$ for $i \notin A$ that is by writing. $\overline{y}(s_i) = \widehat{\pi}_{iA}, S_{yi}^2 = \dfrac{N_i\pi_{iA}(1-\pi_{iA})}{(N_i - 1)}$ and $s_{yi}^2 = \dfrac{\sum_{j\in s_i}\left(y_{ij} - \overline{y}(s_i)\right)^2}{(n_i - 1)} = \dfrac{n_i\,\hat{\pi}_{iA}(1-\hat{\pi}_{iA})}{(n_i - 1)}$.

7.4.2.2 Simple Random Sampling With Replacement

In case samples s_i's are selected by SRSWR method, we get the following theorem from Section 7.4.1.4 by substituting $y_{ij} = 1$ for $i \in A$ and $y_{ij} = 0$ for $i \notin A$.

Theorem 7.4.3

(i) $\widehat{\pi}_A = \sum_{i=1}^{K} W_i \widehat{\pi}_{iA}$ is an unbiased estimator for the mean π_A where $\widehat{\pi}_{iA} = n_{iA}/n$

(ii) $V(\widehat{\pi}_A) = \sum_{i=1}^{K} W_i^2 \frac{\pi_{iA}(1 - \pi_{iA})}{n_i}$

(iii) An unbiased estimator of $V(\widehat{\pi}_A)$ is

$$\widehat{V}(\widehat{\pi}_A) = \sum_{i=1}^{K} W_i^2 \frac{\widehat{\pi}_{iA}(1 - \widehat{\pi}_{iA})}{n_i - 1}$$

7.4.3 Interval Estimation

If the sample size in each stratum is large or the number of strata is large, we can find $(1 - \alpha) \times 100\%$ confidence intervals for the population mean $\overline{Y}$ and proportion π_A by using the central limit theorem as follows:

$$(1 - \alpha) \times 100\% \text{ confidence interval of } \overline{Y}\text{: } \widehat{\overline{Y}}_{st} \pm z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)} \quad (7.4.2)$$

$$(1 - \alpha) \times 100\% \text{ confidence interval of } \pi_A\text{: } \widehat{\pi}_A \pm z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\pi}_A)} \quad (7.4.3)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ % point of a standard normal distribution.

In case the total number of strata and sample sizes from each of the strata is small, we find confidence intervals by replacing $z_{\alpha/2}$ in the formulae above by $t_{\alpha/2,n-K}$, which is the upper $\alpha/2 \times 100\%$ point of t distribution with $n - K$ degrees of freedom.

7.5 ALLOCATION OF SAMPLE SIZE

In this section, we will consider the method of allocation of total sample size n to different strata, i.e., the number of units that should be taken from the different strata. The optimum allocation of sample sizes to different strata are determined either by minimizing the cost of the survey, keeping the efficiency (variance) of the estimator to a certain level or by maximizing the efficiency of the survey (minimizing variance) keeping the cost of the survey to a certain level. Here we will consider the following simple cost function used by Cochran (1977).

$$C = c_0 + \sum_{i=1}^{K} n_i c_i \quad (7.5.1)$$

where C = total cost of the survey, c_0 = overall cost, which is fixed for the survey and does not depend on the sample size, and c_i = cost of surveying each unit of the ith stratum.

Let us assume the situation where the $V\left(\widehat{\overline{Y}}_{st}\right)$ can be expressed in the following form

$$V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} A_i/n_i + B \tag{7.5.2}$$

where A_i and B are independent of the sample sizes n_i

7.5.1 Optimum Allocation for Fixed Cost

Here we minimize the variance $V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K} A_i/n_i + B$ subject to the condition that the total cost of the survey is kept fixed as C_0, i.e.,

$$C_0 = c_0 + \sum_{i=1}^{K} n_i c_i \tag{7.5.3}$$

For this minimization problem, consider

$$\Phi = \sum_{i=1}^{K} A_i/n_i + B - \lambda\left(C_0 - c_0 - \sum_{i=1}^{K} n_i c_i\right) \tag{7.5.4}$$

with λ as a Lagrange multiplier.

Now $\dfrac{\partial \Phi}{\partial n_i} = 0$ implies

$$n_i = \frac{\sqrt{A_i}}{\sqrt{\lambda}\sqrt{c_i}} \tag{7.5.5}$$

On substituting Eq. (7.5.5) in Eq. (7.5.3), we get

$$\frac{1}{\sqrt{\lambda}} = \frac{C_0 - c_0}{\sum_{i=1}^{K} \sqrt{A_i c_i}} \tag{7.5.6}$$

Substituting Eq. (7.5.6) in Eq. (7.5.5), the optimum value of n_i for a given cost $C = C_0$ comes out as

$$opt(n_i|C_0) = n_i(C_0) = \frac{C_0 - c_0}{\left(\sum_{i=1}^{K} \sqrt{A_i c_i}\right)} \sqrt{\frac{A_i}{c_i}} \tag{7.5.7}$$

Finally substituting $n_i = n_i(C_0)$ in Eq. (7.5.2), we get the optimum value of $V\left(\widehat{\overline{Y}}_{st}\right)$ as

$$V_{opt|C_0} = \frac{\left(\sum_{i=1}^{K} \sqrt{A_i c_i}\right)^2}{C_0 - c_0} + B \tag{7.5.8}$$

Hence we get the following theorem:

Theorem 7.5.1

If the total cost of the survey C given in Eq. (7.5.1) is fixed as C_0, then the optimum values of n_i's that minimize the variance $V\left(\widehat{\overline{Y}}_{st}\right)$ given in Eq. (7.5.2) and the minimum value of $V\left(\widehat{\overline{Y}}_{st}\right)$ with the optimum n_i's are given by $n_i(C_0) = \dfrac{C_0 - c_0}{\sum_{i=1}^{K}\sqrt{A_i c_i}}\sqrt{\dfrac{A_i}{c_i}}$ and $V_{opt|C_0} = \dfrac{\left(\sum_{i=1}^{K}\sqrt{A_i c_i}\right)^2}{C_0 - c_0} + B$, respectively.

7.5.2 Optimum Allocation for Fixed Variance

Here the optimum sample size is obtained by minimizing $C = c_0 + \sum_{i=1}^{K} n_i c_i$, the total cost of the survey keeping the variance $V\left(\widehat{\overline{Y}}_{st}\right)$ fixed at V_0. Then, following Section 7.5.1, the optimum value of n_i for a given value of $V = V_0$ is obtained as

$$opt(n_i|V_0) = n_i(V_0) = \frac{\sum_{i=1}^{K}\sqrt{A_i c_i}}{V_0 - B}\sqrt{\frac{A_i}{c_i}} \tag{7.5.9}$$

Finally, putting $n_i = n_i(V_0)$ in Eq. (7.5.2), we get the expression of the optimum cost for the given variance V_0 as

$$C_{opt|V_0} = c_0 + \frac{\left(\sum_{i=1}^{K}\sqrt{c_i A_i}\right)^2}{V_0 - B} \tag{7.5.10}$$

The results above are summarized as follows:

Theorem 7.5.2

If the variance $V\left(\widehat{\overline{Y}}_{st}\right)$ is kept as V_0, then the optimum values of n_i's that minimize the cost C, given in Eq. (7.5.1) and the minimum value of C with the optimum n_i's, are given by $n_i(V_0) = \dfrac{\sum_{i=1}^{K}\sqrt{A_i c_i}}{V_0 - B}\sqrt{\dfrac{A_i}{c_i}}$ and $C_{opt|V_0} = c_0 + \dfrac{\left(\sum_{i=1}^{K}\sqrt{c_i A_i}\right)^2}{V_0 - B}$, respectively.

7.5.3 Simple Random Sampling Without Replacement

For SRSWOR sampling, $A_i = W_i^2 S_{yi}^2$ and $B = -\sum_{i=1}^{K} W_i S_{yi}^2 / N$. Hence using Theorems 7.5.1 and 7.5.2, we get

$$n_i(C_0) = \frac{C_0 - c_0}{\left(\sum_{i=1}^{K} W_i S_{yi}\sqrt{c_i}\right)} \frac{W_i S_{yi}}{\sqrt{c_i}},$$

$$V_{opt|C_0} = \frac{\left(\sum_{i=1}^{K} W_i S_{yi}\sqrt{c_i}\right)^2}{C_0 - c_0} - \frac{\sum_{i=1}^{K} W_i S_{yi}^2}{N}$$

$$n_i(V_0) = \frac{\left(\sum_{i=1}^{K} W_i S_{yi}\sqrt{c_i}\right)}{V_0 + \frac{1}{N}\sum_{i=1}^{K} W_i S_{yi}^2} \cdot \frac{W_i S_{yi}}{\sqrt{c_i}} \text{ and}$$

$$C_{opt|V_0} = c_0 + \frac{\left(\sum_{i=1}^{K} W_i S_{yi}\sqrt{c_i}\right)^2}{V_0 + \sum_{i=1}^{K} W_i S_{yi}^2 / N}$$

7.5.4 Simple Random Sampling With Replacement

For SRSWR sampling $V\left(\widehat{\bar{Y}}_{st}\right) = A_i = W_i^2 \sigma_{yi}^2$ and $B = 0$. Hence Theorems 7.5.1 and 7.5.2 yield

$$n_i(C_0) = \frac{C_0 - c_0}{\left(\sum_{i=1}^{K} W_i \sigma_{yi}\sqrt{c_i}\right)} \frac{W_i \sigma_{yi}}{\sqrt{c_i}}, \quad V_{opt|C_0} = \frac{\left(\sum_{i=1}^{K} W_i \sigma_{yi}\sqrt{c_i}\right)^2}{C_0 - c_0},$$

$$n_i(V_0) = \frac{\left(\sum_{i=1}^{K} W_i \sigma_{yi}\sqrt{c_i}\right)}{V_0} \frac{W_i \sigma_{yi}}{\sqrt{c_i}} \text{ and } C_{opt|V_0} = c_0 + \frac{\left(\sum_{i=1}^{K} W_i \sigma_{yi}\sqrt{c_i}\right)^2}{V_0}.$$

7.5.5 Probability Proportional to Size With Replacement Sampling

For PPSWR sampling $V\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{N^2}\sum_{i=1}^{K}\frac{V_{pps|i}}{n_i}$. In this case $A_i = V_{pps|i}/N^2$ and $B = 0$. Hence,

$$n_i(C_0) = \frac{C_0 - c_0}{\left(\sum_{i=1}^{K}\sqrt{V_{pps|i}\ c_i}\right)}\sqrt{\frac{V_{pps|i}}{c_i}},\quad V_{opt|C_0} = \frac{\left(\sum_{i=1}^{K}\sqrt{V_{pps|i}\ c_i}\right)^2}{N^2(C_0 - c_0)},$$

$$n_i(V_0) = \frac{\sum_{i=1}^{K}\sqrt{V_{pps|i}\ c_i}}{N^2 V_0}\sqrt{\frac{V_{pps|i}}{c_i}} \text{ and } C_{opt|V_0} = c_0 + \frac{\left(\sum_{i=1}^{K}\sqrt{c_i\ V_{pps|i}}\right)^2}{N^2 V_0}.$$

Remark 7.5.1

For both the SRSWOR and SRSWR sampling schemes, the optimum sample sizes are directly proportional to the stratum size and stratum standard deviation but inversely proportional to the square root of the stratum cost per unit of survey. Hence we can propose a rule of thumb for allocation of sample sizes as follows: Take more samples from those strata whose sizes and variances are large and select small sample sizes from the strata where the cost of surveying units is high.

7.5.6 Neyman Optimum Allocation

Consider the situation where the cost of surveying the units of all the strata are the same, i.e., $c_i = c$ for $i = 1,\ldots, K$. In this situation, if the total cost of a survey C is kept fixed as C_0, then the total sample size $\sum_{i=1}^{K} n_i$ becomes fixed and is equal to $\frac{C_0 - c_0}{c} = n$. It can be checked that under SRSWOR and SRSWR sampling, the optimum value of n_i with $c_i = c$ is obtained as

$$n_i(C_0) = n\frac{W_i\ S_{yi}}{\left(\sum_{i=1}^{K} W_i S_{yi}\right)} = n\frac{N_i S_{yi}}{\left(\sum_{i=1}^{K} N_i S_{yi}\right)} \tag{7.5.11}$$

The allocation (7.5.11) is the well known Neyman (1934) allocation, which is usable when both the stratum sizes N_i's and stratum variances S_{yi}^2 are known.

7.5.7 Proportional Allocation

If the cost of surveying units in all the strata and the strata variances are the same, i.e., $c_i = c$ and $S_{yi} = S_y$ for $i = 1,2,\ldots, K$, then Eq. (7.5.11) reduces to

$$n_i(C_0) = nN_i/N = nW_i \tag{7.5.12}$$

The allocation (7.5.12) is known as proportional allocation since the sample sizes in the respective stratum are taken proportional to the stratum sizes when the total sample size is fixed as n. Proportional allocation was introduced by Bowley (1926).

7.6 COMPARISON BETWEEN STRATIFIED AND UNSTRATIFIED SAMPLING

In this section, we will consider the performance of stratified and unstratified sampling under various sampling designs that are commonly used in practice.

7.6.1 Simple Random Sampling Without Replacement

Let an unstratified sample of size n be selected from the entire population U by the SRSWOR method and $\overline{y}_s$ be the sample mean. Then,

$$V(\overline{y}_s) = (1-f)\frac{S_y^2}{n} = V_{ran}(\text{say}) \tag{7.6.1}$$

where

$$f = n/N \text{ and } S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{K}\sum_{j\in U_i}(y_{ij}-\overline{Y})^2 = \frac{1}{N-1}\sum_{i=1}^{K}(N_i-1)S_{yi}^2$$

$$+\frac{1}{N-1}\sum_{i=1}^{K}N_i(\overline{Y}_i-\overline{Y})^2$$

Now assuming N_i's are so large that we can take $1/N_i \cong 0$ for $i = 1,\ldots, K$ and hence also $1/N \cong 0$, we can approximate V_{ran} as

$$V_{ran} = \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2 = \left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{i=1}^{K}W_iS_{yi}^2 + \sum_{i=1}^{K}W_i(\overline{Y}_i-\overline{Y})^2\right] \tag{7.6.2}$$

Let a sample s_i of size n_i be selected by an SRSWOR method from the ith stratum keeping $\sum_{i=1}^{K} n_i = n$ as fixed, then the variance of $\widehat{\overline{Y}}_{st} = \sum_{i=1}^{K} W_i\overline{y}(s_i)$ becomes

$$V\left(\widehat{\overline{Y}}_{st}\right) = \sum_{i=1}^{K}W_i^2(1-f_i)S_{yi}^2/n_i \tag{7.6.3}$$

Under Neyman optimum allocation, $n_i = n_{i0} = n\dfrac{W_i\ S_{yi}}{\left(\sum_{i=1}^{K} W_i S_{yi}\right)}$ and the expression of $V\left(\widehat{\overline{Y}}_{st}\right)$ in Eq. (7.6.3) reduces to

$$V_{opt} = \frac{\left(\sum_{i=1}^{K} W_i\ S_{yi}\right)^2}{n} - \frac{1}{N}\sum_{i=1}^{K} W_i\ S_{yi}^2 \tag{7.6.4}$$

For proportional allocation, $n_i = nN_i/N = nW_i$, the expression (7.6.3) reduces to

$$V_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} W_i\ S_{yi}^2 \tag{7.6.5}$$

Eqs. (7.6.4) and (7.6.5) yield

$$V_{prop} - V_{opt} = \frac{1}{n}\sum_{i=1}^{K} W_i\left[S_{yi} - \left(\sum_{i=1}^{K} W_i S_{yi}\right)\right]^2 \geq 0 \tag{7.6.6}$$

From Eqs. (7.6.2) and (7.6.5), we get

$$V_{ran} - V_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} W_i\left(\overline{Y}_i - \overline{Y}\right)^2 \geq 0 \tag{7.6.7}$$

Finally using Eqs. (7.6.6) and (7.6.7), the following theorem is obtained.

Theorem 7.6.1

If the sample size n is fixed and $1/N_i \cong 0$ for $i = 1,\ldots, K$, then

$$V_{ran} \geq V_{prop} \geq V_{opt}$$

7.6.2 Probability Proportional to Size With Replacement Sampling

Let $x_{ij}(>0)$ be the measure of size for the jth unit of the ith stratum and $p_{ij} = x_{ij}/X$ be the corresponding normed size measure with $X = \sum_{i=1}^{K} X_i$ and $X_i = \sum_{j=1}^{N_i} x_{ij}$. Suppose an unstratified sample of size n is selected from the

population by PPSWR sampling scheme using p_{ij} as normed size measure for the jth unit of the ith stratum. Then the Hansen–Hurwitz (HH) estimator for population mean $\overline{Y}$ is given by

$$\widehat{\overline{Y}}_{hh} = \frac{1}{nN}\sum_{r=1}^{n}\frac{y(r)}{p(r)}$$

where $y(r) = y_{ij}$ and $p(r) = p_{ij}$ if the rth draw produces the jth unit of the ith stratum.

The variance of the HH estimator $\widehat{\overline{Y}}_{hh}$ is given by

$$V\left(\widehat{Y}_{hh}\right) = \frac{1}{N^2 n}\left(\sum_{i=1}^{K}\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}} - N^2\overline{Y}^2\right) \tag{7.6.8}$$

Consider a stratified sampling where sample s_i of size n_i is selected from the ith stratum for $i = 1,\ldots, K$ by PPSWR method using normed size measure $p_{ij}^* = p_{ij}/P_i$ for the jth unit of the ith stratum, where $P_i = X_i/X$; $i = 1,2,\ldots, K, j = 1,\ldots, N_i$. Then the HH estimator for the mean $\overline{Y}$ based on the stratified sample is given by

$$\widehat{\overline{Y}}_{hh}^{st} = \frac{1}{N}\sum_{i=1}^{K}\widehat{Y}_i$$

where $\widehat{Y}_i = \frac{1}{n_i}\sum_{r=1}^{n_i}\frac{y(r,i)}{p(r,i)}$ is an unbiased estimator for the total Y_i, $y(r,i) = y_{ij}$ and $p(r,i) = p_{ij}^*$ if the rth draw produces the jth unit of the ith stratum with probability p_{ij}^*.

The variance of $\widehat{\overline{Y}}_{hh}^{st}$ is given by

$$V\left(\widehat{\overline{Y}}_{hh}^{st}\right) = \frac{1}{N^2}\sum_{i=1}^{K}\frac{1}{n_i}\left(\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}^*} - Y_i^2\right) \tag{7.6.9}$$

Consider Arnab allocation (1991), where n_i's are allocated with proportional X_i's, the ith stratum total of the auxiliary variable, i.e., $n_i = nX_i/X = nP_i$. In this case, Eq. (7.6.9) reduces to

$$\begin{aligned} V\left(\widehat{\overline{Y}}_{hh}^{st}\right) &= \frac{1}{N^2}\left[\sum_{i=1}^{K}\frac{1}{nP_i}\left(\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}^*} - Y_i^2\right)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^{K}\frac{1}{n}\left(\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}} - \frac{Y_i^2}{P_i}\right)\right] \end{aligned} \tag{7.6.10}$$

Eqs. (7.6.8) and (7.6.10) yield

$$V\left(\widehat{Y}_{hh}\right) - V\left(\widehat{\overline{Y}}^{st}_{hh}\right) = \frac{1}{N^2 n}\left(\sum_{i=1}^{K}\frac{Y_i^2}{P_i} - Y^2\right)$$
$$= \frac{1}{N^2 n}\sum_{i=1}^{K} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 \qquad (7.6.11)$$
$$\geq 0$$

Eq. (7.6.11) derived by Arnab (1991) leads to the following theorem:

Theorem 7.6.2

For a given sample size n, the stratified PPSWR sampling is more efficient than unstratified PPSWR sampling if sample sizes are allocated to each of the strata proportional to the aggregate of the measure of size and in this case

$$V\left(\widehat{Y}_{hh}\right) - V\left(\widehat{\overline{Y}}^{st}_{hh}\right) = \frac{1}{N^2 n}\sum_{i=1}^{K} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 \geq 0$$

7.6.3 Inclusion Probability Proportional to Size Sampling Scheme

Consider an unstratified sample s of size n selected from the entire population U using a suitable IPPS sampling design with inclusion probability $\pi^u_{j|i} = n\, x_{ij}/X$ attached to the jth unit of the ith stratum. In this case the Horvitz–Thompson estimator (HTE) for the population total Y is given by

$$\widehat{Y}^u_{ht} = \sum_i \sum_{j\in s} \frac{y_{ij}}{\pi^u_{j|i}}$$
$$= \frac{X}{n}\sum_i \sum_{j\in s} \frac{y_{ij}}{x_{ij}}$$

Suppose from the ith $(i = 1,\ldots, K)$ stratum a sample s_i of size $n_i\left(\sum_{i=1}^{K} n_i = n\right)$ is selected with inclusion probability $\pi_{j|i} = n_i x_{ij}/X_i$ for the jth unit of the ith stratum. Then the HTE for Y based on the stratified sample is given by

$$\widehat{Y}^{st}_{ht} = \sum_i \sum_{j\in s_i} \frac{y_{ij}}{\pi_{j|i}}$$
$$= \sum_i X_i \sum_{j\in s_i} \frac{y_{ij}}{n_i x_{ij}}$$

Let us suppose that the study variable y be related to the auxiliary variable x through the following superpopulation model ξ

$$E_\xi(y_{ij}) = \beta x_{ij},\ V_\xi(y_{ij}) = \sigma^2 x_{ij}^2 \text{ and } C_\xi(y_{ij}, y_{kl}) = 0 \ \text{ for} (i,j) \neq (k,l) \quad (7.6.12)$$

where E_ξ, V_ξ, and C_ξ denote the operator for expectation, variance, and covariance with respect to the model ξ, respectively.

The expected variances of $\widehat{Y}_{ht}^{u}$ and $\widehat{Y}_{ht}^{st}$ under the superpopulation model ξ are given, respectively, by

$$\begin{aligned} E_\xi V\left(\widehat{Y}_{ht}^{u}\right) &= E_\xi E\left(\widehat{Y}_{ht}^{u} - Y\right)^2 \\ &= \sigma^2 \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij}^2 \left(\frac{1}{\pi_{j|i}^{u}} - 1\right) \\ &= \sigma^2 \left(\frac{X^2}{n} - \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij}^2\right) \end{aligned}$$

and

$$E_\xi V\left(\widehat{Y}_{ht}^{st}\right) = E_\xi E\left(\widehat{Y}_{ht}^{st} - Y\right)^2 = \sigma^2 \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij}^2 \left(\frac{1}{\pi_{j|i}} - 1\right)$$

$$= \sigma^2 \left(\sum_{i=1}^{K} \frac{X_i^2}{n_i} - \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij}^2\right)$$

Now noting,

$$\text{(i)} \sum_{i=1}^{K} \frac{X_i^2}{n_i} = \sum_{i=1}^{K} n_i \left(\frac{X_i}{n_i}\right)^2 \geq \frac{\left(\sum_{i=1}^{K} X_i\right)^2}{\sum_{i=1}^{K} n_i} = \frac{X^2}{n}$$

and

$$\text{(ii)} \sum_{i=1}^{K} \frac{X_i^2}{n_i} = \frac{X^2}{n} \text{ if } n_i = nX_i/X,$$

we find $E_\xi V\left(\widehat{Y}_{ht}^{st}\right) \geq E_\xi V\left(\widehat{Y}_{ht}^{u}\right)$.

Furthermore, $E_\xi V\left(\widehat{Y}_{ht}^{st}\right) = E_\xi V\left(\widehat{Y}_{ht}^{u}\right)$ if $n_i = nX_i/X$. Hence we have the following theorem.

Theorem 7.6.3

Under the superpopulation model (7.6.12), the unstratified IPPS sampling design is more efficient than the stratified IPPS sampling design based on

the same sample size. Both of them are equally efficient if $n_i = nX_i/X$ for $i = 1,\ldots, K$.

Comparisons between the stratified and unstratified sampling were studied in details by Hanurav (1965), Ramachandran and Rao, T.J. (1974), Cassel et al. (1977), and Rao, T.J. (1968, 1977a,b, 1983), among others.

Example 7.6.1

4000 households of a city were stratified into three strata: high-, middle-, and low-income groups. A stratified sample of 400 households was selected from the entire population by using the SRSWOR method. The following table gives the sample sizes, sample mean, and sample variances of the household income and the number of earning members of the sampled households.

(i) Estimate the mean household income of the city and 95% confidence interval of the mean income.

(ii) Estimate the proportion of earning members of the city and 95% confidence interval of this proportion.

Income Group (i)	Number of households N_i	Sampled households n_i	Number of earning members in the sample z_i	Sample mean income (\$) $\bar{y}(s_i)$	Sample standard deviation of income (\$) s_{yi}
High	500	50	10	25,000	300
Middle	1500	150	20	15,000	200
Low	2000	200	20	2500	400
Total	4000	400	50	—	

We first calculate the following:

Strata	W_i	$W_i\bar{y}(s_i)$	$W_i^2(1-f_i)s_{yi}^2/n_i$	$\hat{\pi}_{iA}$	$W_i\hat{\pi}_{iA}$	$\frac{W_i^2(1-f_i)\hat{\pi}_{iA}(1-\hat{\pi}_{iA})}{n_i-1}$
High	0.125	3125	25.3125	0.2000	0.025	0.00004
Middle	0.375	5625	33.7500	0.1333	0.050	0.0001
Low	0.500	1250	180.0000	0.1000	0.050	0.0001
Total	1	10,000	239.0625	—	0.125	0.00024

(i) Here the estimated mean household income. $\widehat{\bar{Y}}_{st} = \sum_{i=1}^{4} W_i\bar{y}(s_i)$ $= \$10,000$.

Estimated standard error of $\widehat{\overline{Y}}_{st}$ is $\sqrt{\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)} = \sqrt{\sum W_i^2(1-f_i)s_{yi}^2/n_i}$ = \$15.461.

95% confidence interval of the mean income is $10{,}000 \pm 1.96 \times 15.461 =$ (\$ 9969.695, \$ 10,030.30)

(ii) Estimated population proportion of earning member is $\widehat{\pi}_A = \sum_{i=1}^{4} W_i \widehat{\pi}_{iA} = 0.125$.

Estimated standard error of $\widehat{\pi}_A$ is $\sqrt{\sum W_i^2(1-f_i)\widehat{\pi}_{iA}(1-\widehat{\pi}_{iiA})/(n_i-1)}$ = 0.0156

95% confidence interval of π_A is $0.125 \pm 1.96 \times 0.0156 = (0.094, 0.156)$.

Example 7.6.2

A survey was conducted to determine the average farm size cultivating maize in the KwaZulu- Natal (KZN) province of South Africa. The farmers are classified into five strata according to their farm sizes. The total cost of the survey was limited to \$ 50, 000. The fixed cost of the survey was budgeted as \$ 10,000 to cover expenses for administration and report writing. The variable costs for collection of data per farm, including transport and filling schedule, are given in the following table along with other information such as the number of farms, average farm sizes, and SD of the farm sizes for respective stratum.

(a) Determine the sample sizes that should be selected from each stratum under (i) proportional, (ii) Neyman allocation, and (iii) optimum allocations.

(b) Compute the relative efficiencies of the Neyman and optimum allocation with respect to the proportional allocation for estimating the population mean when samples from each stratum are selected by SRSWOR method.

Stratum (i)	Number of farms (N_i)	Mean area under maize (acre) ($\overline{Y}_i$)	SD of area under maize (S_i)	Cost of collection of data per farm (c_i)
1	500	80	5.25	100
2	400	100	10.75	100
3	300	120	15.50	150
4	200	150	20.50	175
5	100	200	25.00	200

Here the cost function is

$$C = C_o + \sum_{i=1}^{5} c_i n_i \tag{7.6.13}$$

where $C = 50{,}000$ and $C_o = 10{,}000$.

(a) (i) Proportional allocation:

Under proportional allocation $n_i = nN_i/N$, where $N = \sum_{i=1}^{5} N_i$ and n should be obtained from the equation $C - C_o = n\sum_{i=1}^{5} c_i N_i/N$, i.e., $n = \dfrac{(C - C_o)}{\sum_{i=1}^{5} c_i N_i/N}$ and hence n_i under proportional allocation is

$$n_{i0}(p) = \frac{(C - C_o)}{\sum_{i=1}^{5} c_i N_i} N_i$$

(ii) Neyman allocation: Under this allocation $n_i = nN_iS_i \Big/ \sum_{i=1}^{5} N_iS_i$ and Eq. (7.6.13) yields $C - C_o = n\sum_{i=1}^{5} c_i N_i S_i \Big/ \sum_{i=1}^{5} N_i S_i$, i.e., $n = \dfrac{(C - C_o)\left(\sum_{i=1}^{5} N_i S_i\right)}{\sum_{i=1}^{5} c_i N_i S_i}$ and hence n_i under Neyman allocation is

$$n_{i0}(ney) = \frac{(C - C_o) N_i S_i}{\sum_{i=1}^{5} c_i N_i S_i}$$

(iii) Under optimum allocation

$$n_{i0}(opt) = \frac{(C - C_o) N_i S_i \Big/ \sqrt{c_i}}{\sum_{i=1}^{5} \sqrt{c_i} N_i S_i}$$

The values of n_i's under proportional, Neyman and optimum allocation are obtained as follows:

i	N_i	S_i	c_i	N_ic_i	$n_{i0}(p) = \dfrac{(C-C_o)N_i}{\sum c_iN_i}$	$N_iS_ic_i$	$n_{i0}(ney) = \dfrac{(C-C_o)N_iS_i}{\sum c_iN_iS_i}$	$N_iS_i\sqrt{c_i}$	$n_{i0}(opt) = \dfrac{(C-C_o)N_iS_i/\sqrt{c_i}}{\sum \sqrt{c_i}N_iS_i}$
1	500	5.25	100	50,000	105	262,500	40	26,250	49
2	400	10.8	100	40,000	84	430,000	66	43,000	80
3	300	15.5	150	45,000	63	697,500	71	56,950.64	70
4	200	20.5	175	35,000	42	717,500	63	54,237.9	57
5	100	25	200	20,000	21	500,000	38	35,355.34	33
	1500			190,000	315	2,607,500	278	215,793.9	289

(b) The unbiased estimator for the population mean $\overline{Y}$ is $\widehat{\overline{Y}}_{st} = \sum_{i=1}^{5} N_i \overline{y}(s_i)/N$. The variance of $\widehat{\overline{Y}}_{st}$ under proportional, Neyman, and optimum allocations are respectively given by

$$V_{prop} = \sum_{i=1}^{5} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{n_{i0}(p)} - \frac{1}{N_i}\right) S_i^2 = 0.4659,$$

$$V_{ney} = \sum_{i=1}^{5} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{n_{i0}(ney)} - \frac{1}{N_i}\right) S_i^2 = 0.4043 \text{ and}$$

$$V_{opt} = \sum_{i=1}^{5} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{n_{i0}(opt)} - \frac{1}{N_i}\right) S_i^2 = 0.394.$$

The relative efficiency of Neyman and optimum allocations are $\frac{V_{prop}}{V_{ney}} \times 100 = 115.23\%$ and $\frac{V_{prop}}{V_{opt}} \times 100 = 118.26\%$, respectively.

7.7 CONSTRUCTION OF STRATA

It is obvious that the strata should be constructed in such a way that each stratum becomes homogeneous with respect to the character under study so that strata means differ significantly from each other, but at the same time variation within the strata becomes as small as possible. To make the strata homogeneous, one needs to have some prior knowledge of the population such as past survey data or auxiliary information related to the study variable. As far as the number of strata is concerned, more is better because the efficiency of stratified sampling increases as the number of strata increases. According to Dalenius (1953) conjecture $V_{k+1}\left(\widehat{\overline{Y}}_{st}\right)$, the variance of the estimator of the population mean based on $k+1$ strata is approximately equal to $\left(\frac{k}{k+1}\right)^2 V_k\left(\widehat{\overline{Y}}_{st}\right)$. Dalenius (1953) conjecture was supported by Cochran (1961) through empirical studies. Hence the gain in efficiency of increasing the number of strata becomes marginal after a certain stage. Various researchers have paid attention to construction of strata and details have been given by Sukhatme et al. (1984), Murthy (1977), and Cochran (1977), among others. Here we will highlight the following methods of optimum points of stratification when there is plan to divide a population into a predetermined number of strata K.

7.7.1 Optimum Points of Stratification

Suppose that we want to stratify a population into K strata, where K is prespecified. Let y_0 and y_K be the maximum and minimum value

of y, respectively. Let the interval (y_0, y_K) be divided at the points $y_1, \ldots, y_{K-1}$ $(y_1 < \cdots < y_{K-1})$ to form K strata. The problem is to determine the points of division so that the variance of the estimator of the population mean attains a minimum. Clearly the optimum points of division depend on the sampling design and the estimators used. For simplicity, let us assume (i) each stratum size is large and (ii) the study variable has a continuous probability density function $f(y)$ so that we can approximate

$$W_i = \int_{y_{i-1}}^{y_i} f(y)dy, \overline{Y}_i = \mu_i = \frac{1}{W_i} \int_{y_{i-1}}^{y_i} yf(y)dy,$$

$$S_{yi}^2 = \sigma_i^2 = \frac{1}{W_i} \int_{y_{i-1}}^{y_i} (y - \mu_i)^2 f(y)dy, i = 1, \ldots, k.$$

The expressions of $V\left(\widehat{\overline{Y}}_{st}\right)$ under proportional and optimum allocation for SRSWR sampling are given respectively as follows:

$$V_{prop} = \frac{\sum_{i=1}^{K} W_i S_{yi}^2}{n} = \frac{\sum_{i=1}^{K} W_i \sigma_i^2}{n} \text{ and } V_{opt} = \frac{\left(\sum_{i=1}^{K} W_i S_{yi}\right)^2}{n}$$
$$= \frac{\left(\sum_{i=1}^{K} W_i \sigma_i\right)^2}{n}$$

7.7.1.1 Proportional Allocation

Under proportional allocation

$$V_{prop} = \frac{\sum_{i=1}^{K} W_i \sigma_i^2}{n}$$

$$= \frac{\sum_{i=1}^{K} \int_{y_{i-1}}^{y_i} (y - \mu_i)^2 f(y)dy}{n}$$

$$= \frac{\sum_{i=1}^{K} \int_{y_{i-1}}^{y_i} y^2 f(y)dy - \sum_{i=1}^{k} W_i \mu_i^2}{n}$$

$$= \frac{1}{n} \int_{y_0}^{y_K} y^2 f(y)dy - \frac{1}{n} \sum_{i=1}^{k} \frac{1}{W_i} \left(\int_{y_{i-1}}^{y_i} yf(y)dy \right)^2$$

Now

$$\frac{\partial V_{prop}}{\partial \gamma_i} = -\frac{1}{n}\frac{\partial}{\partial \gamma_i}\sum_{i=1}^{k}\frac{1}{W_i}\left(\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)^2$$

$$= -\frac{1}{n}\left[\frac{\partial}{\partial \gamma_i}\frac{1}{W_i}\left(\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)^2 + \frac{\partial}{\partial \gamma_i}\frac{1}{W_{i+1}}\left(\int_{\gamma_i}^{\gamma_{i+1}} \gamma f(\gamma)d\gamma\right)^2\right]$$

(since the other terms do not involve γ_i)

$$= -\frac{1}{n}\left[-\left(\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)^2\left((W_i)^{-2}\frac{\partial W_i}{\partial \gamma_i}\right) + (W_i)^{-1}\right.$$

$$\left.\times\left(2\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)\left(\frac{\partial}{\partial \gamma_i}\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)\right]$$

$$-\frac{1}{n}\left[\left(\int_{\gamma_i}^{\gamma_{i+1}} \gamma f(\gamma)d\gamma\right)^2\left((W_{i+1})^{-2}\frac{\partial W_{i+1}}{\partial \gamma_i}\right) + (W_{i+1})^{-1}\right.$$

$$\left.\times\left(2\int_{\gamma_i}^{\gamma_{i+1}} \gamma f(\gamma)d\gamma\right)\left(\frac{\partial}{\partial \gamma_i}\int_{\gamma_i}^{\gamma_{i+1}} \gamma f(\gamma)d\gamma\right)\right]$$

$$= -\frac{1}{n}\left[-\left(\mu_i^2\frac{\partial W_i}{\partial \gamma_i}\right) + 2\mu_i\frac{\partial}{\partial \gamma_i}\left(\int_{\gamma_{i-1}}^{\gamma_i} \gamma f(\gamma)d\gamma\right)\right.$$

$$\left.-\left(\mu_{i+1}^2\frac{\partial W_{i+1}}{\partial \gamma_i}\right) + 2\mu_{i+1}\frac{\partial}{\partial \gamma_i}\left(\int_{\gamma_i}^{\gamma_{i+1}} \gamma f(\gamma)d\gamma\right)\right]$$

Furthermore, noting

$$\frac{\partial W_i}{\partial y_i} = f(y_i), \frac{\partial W_{i+1}}{\partial y_i} = -f(y_i), \frac{\partial}{\partial y_i}\left\{\int_{y_{i-1}}^{y_i} y f(y)dy\right\}$$

$$= y_i f(y_i), \frac{\partial}{\partial y_i}\left\{\int_{y_i}^{y_{i+1}} y f(y)dy\right\} = -y_i f(y_i)$$

and equating $\frac{\partial V_{prop}}{\partial y_i} = 0$, the optimum points of stratification are obtained as

$$y_i = \frac{\mu_i + \mu_{i+1}}{2} \text{ for } i = 1, \ldots, K-1 \tag{7.7.1}$$

Since μ_i's are unknown, y_i's can be determined by using an iterative procedure when $f(y)$ is known.

7.7.1.2 Optimum Allocation

The optimum points of stratification y_i's are obtained by solving the equation

$$\frac{\partial V_{opt}}{\partial y_i} = 2\frac{\left(\sum_{i=1}^{K} W_i\sigma_i\right)}{n}\left\{\frac{\partial(W_i\sigma_i)}{\partial y_i} + \frac{\partial(W_{i+1}\sigma_{i+1})}{\partial y_i}\right\} = 0 \tag{7.7.2}$$

Now

$$\frac{\partial(W_i\sigma_i)}{\partial y_i} = \sigma_i\frac{\partial W_i}{\partial y_i} + W_i\frac{\partial\sigma_i}{\partial y_i} \tag{7.7.3}$$

Further differentiating both sides of $W_i\sigma_i^2 = \int_{y_{i-1}}^{y_i}(y-\mu_i)^2 f(y)dy$ with respect to y_i, we get

$$f(y_i)\sigma_i^2 + 2W_i\sigma_i\frac{\partial\sigma_i}{\partial y_i} = (y_i-\mu_i)^2 f(y_i) \tag{7.7.4}$$

Eqs. (7.7.3) and (7.7.4) yield

$$\frac{\partial(W_i\sigma_i)}{\partial y_i} = \frac{f(y_i)}{2}\left[\sigma_i + \frac{(y_i-\mu_i)^2}{\sigma_i}\right] \tag{7.7.5}$$

Similarly we get

$$\frac{\partial(W_{i+1}\sigma_{i+1})}{\partial y_i} = -\frac{f(y_i)}{2}\left[\sigma_{i+1} + \frac{(y_i - \mu_{i+1})^2}{\sigma_{i+1}}\right] \tag{7.7.6}$$

Finally putting Eqs. (7.7.5) and (7.7.6) in Eq. (7.7.2) and setting $\frac{\partial V_{opt}}{\partial y_i} = 0$, the optimum points of stratifications y_i's are obtained from the following equation

$$\sigma_i + \frac{(y_i - \mu_i)^2}{\sigma_i} = \sigma_{i+1} + \frac{(y_i - \mu_{i+1})^2}{\sigma_{i+1}} \text{ for } i = 1, \ldots, K-1 \tag{7.7.7}$$

here also y_i's cannot be solved easily from the equation above because μ_i's and σ_i's are unknown. However, if $f(y)$ is known, y_i's may be obtained by using an iterative method when the number of strata is small.

7.7.2 Dalenius and Hodges's Approximation

Dalenius and Hodges (1959) have given a simple formula to determine the optimum points of stratification under the following assumptions:

(i) The stratum weight W_i is approximated as

$$W_i = \int_{y_{i-1}}^{y_i} f(y)dy \cong f(y_i)\int_{y_{i-1}}^{y_i} dy = f(y_i)(y_i - y_{i-1}) \tag{7.7.8}$$

(ii) The distribution of y is uniform for all the strata so that

$$\sigma_i^2 = \frac{1}{W_i}\int_{y_{i-1}}^{y_{i1}} (y - \mu_i)^2 f(y)dy = \frac{(y_i - y_{i-1})^2}{12}$$

So, under the approximations (i) and (ii) of uniform distribution over the strata, the expression of $V_{opt} = \frac{\left(\sum_{i=1}^{K} W_i\sigma_i\right)^2}{n}$ reduces to $\frac{\left(\sum_{i=1}^{K} F_i^2\right)^2}{12n}$, where $F_i = (y_i - y_{i-1})\sqrt{f(y_i)}$

Now

$$V_{opt} = \frac{\left(\sum_{i=1}^{K} F_i^2\right)^2}{12n} \geq \frac{\left(\sum_{i=1}^{K} F_i\right)^4}{12nK^2} \tag{7.7.9}$$

Thus V_{opt} attains a minimum when

$$F_i = \frac{1}{K}\sum_{i=1}^{K}(y_i - y_{i-1})\sqrt{f(y_i)} \cong \frac{1}{K}\sum_{i=1}^{K}\int_{y_{i-1}}^{y_i}\sqrt{f(y)}dy = \frac{1}{K}\int_{y_0}^{y_K}\sqrt{f(y)}dy$$

i.e., when

$$F_i = \int_{y_{i-1}}^{y_i}\sqrt{f(y)}dy = \frac{1}{K}\int_{y_0}^{y_K}\sqrt{f(y)}dy = \overline{F}(\text{say}) \tag{7.7.10}$$

Thus the Dalenius and Hodges (1959) rule is to find the points of stratification, which yield equal cumulative values of the square root of the frequency function $f(y)$.

Example 7.7.1

The following table gives the frequency distribution of daily wages of 820 workers in a certain factory.

Wages ($)	Frequency (f)
Less than 50	80
51−100	120
101−150	250
151−200	120
201−300	100
301−450	70
451−600	50
Above 601	30

It is decided to divide the workers in five strata. Use the Dalenius and Hodges rule to find the optimal strata.

Here we first prepare the cumulative frequency table of $\sqrt{f}$ as follows:

Wages ($)	Frequency (f)	$\sqrt{f}$	Cumulative $\sqrt{f}$
Less than 50	80	8.94	8.94
51−100	120	10.95	19.89
101−150	250	15.81	35.7
151−200	120	10.95	46.65
201−300	100	10.00	56.65
301−450	70	8.37	65.02
451−600	50	7.07	72.09
Above 600	30	5.48	77.57

The computed value of $\overline{F} = 77.57/5 = 15.51$, $2\overline{F} = 31.03$, $3\overline{F} = 46.54$, $4\overline{F} = 62.06$ and $5\overline{F} = 77.57$. Thus the five strata may be formed as follows:

Strata	Wages ($)	Size
I	0−100	200
II	101−150	250
III	151−200	120
IV	201−450	170
V	Above 450	80

7.7.3 Other Methods

Dalenius and Gurney (1951) suggested that strata should be formed so that $W_i\,\sigma_i$ should be constant, whereas Mahalanobis (1952) recommended that $W_i\,\mu_i$ should be constant. Cochran (1961) suggested that the coefficient of variations for all the strata should be the same i.e.,

$$\frac{\sigma_1}{\mu_1} = \cdots = \frac{\sigma_i}{\mu_i} = \cdots = \frac{\sigma_K}{\mu_K} \tag{7.7.11}$$

Furthermore, if the distribution of y within each stratum is uniform so that

$$\mu_i = \frac{y_{i-1} + y_i}{2} \text{ and } \sigma_i = \frac{y_i - y_{i-1}}{\sqrt{12}}$$

Eq. (7.7.11) leads to

$$\frac{y_i - y_{i-1}}{y_i + y_{i-1}} = \frac{y_{i+1} - y_i}{y_{i+1} + y_i} \tag{7.7.12}$$

Gunning and Horgan (2004) derived the following recurrence relation using Eq. (7.7.12)

$$\begin{gathered} y_i^2 = y_{i+1}\ \ y_{i-1} \\ \text{i.e., } y_i = ar^i \text{ for } i = 0, \ldots, K \end{gathered} \tag{7.7.13}$$

Now putting $i = 0$ and $i = K$, we get $a = y_0$ and $r = (y_K/y_0)^{\frac{1}{K}}$. Substituting $a = y_0$ and $r = (y_K/y_0)^{\frac{1}{K}}$ in Eq. (7.7.13), we get

$$y_i = y_0 \left(\frac{y_K}{y_0}\right)^{\frac{i}{K}} \tag{7.7.14}$$

Thus for $K = 4$, $\gamma_0 = 5$, and $\gamma_4 = 50{,}000$, Gunning and Horgan (2004) obtained the optimum points of stratification as $\gamma_1 = 50$, $\gamma_2 = 500$, and $\gamma_3 = 5000$. They showed that the proposed method performs reasonably well for positively skewed populations.

7.8 ESTIMATION OF GAIN DUE TO STRATIFICATION

Suppose from the ith stratum a sample s_i of size n_i is selected by some suitable sampling procedure with $\pi_{l|i}$ and $\pi_{lq|i}$ as the inclusion probabilities for the lth unit and lth and qth unit of the ith stratum l, $q = 1,\ldots, N_i$, $i = 1,\ldots, K$, $l \neq q$. The HTE of $\overline{Y}$ based on the stratified sample is given by

$$\widehat{\overline{Y}}_{ht}^{st} = \frac{1}{N}\sum_{i=1}^{K}\left(\sum_{l\in s_i}\frac{\gamma_{il}}{\pi_{l|i}}\right) \tag{7.8.1}$$

The variance of $\widehat{\overline{Y}}_{ht}^{st}$ and its unbiased estimates are respectively given by

$$V\left(\widehat{\overline{Y}}_{ht}^{st}\right) = \frac{1}{2N^2}\sum_{i=1}^{K}\sum_{l\neq}^{N_i}\sum_{q=1}^{N_i}\left(\pi_{l|i}\pi_{q|i} - \pi_{lq|i}\right)\left(\frac{\gamma_{il}}{\pi_{l|i}} - \frac{\gamma_{iq}}{\pi_{q|i}}\right)^2 \tag{7.8.2}$$

and

$$\widehat{V}\left(\widehat{\overline{Y}}_{ht}^{st}\right) = \frac{1}{2N^2}\sum_{i=1}^{K}\sum_{l\neq}^{N_i}\sum_{q\in s_i}^{N_i}\left(\frac{\pi_{l|i}\pi_{q|i} - \pi_{lq|i}}{\pi_{lq|i}}\right)\left(\frac{\gamma_{il}}{\pi_{l|i}} - \frac{\gamma_{iq}}{\pi_{q|i}}\right)^2 \tag{7.8.3}$$

Suppose an unstratified sample s of size n is selected from the entire population using some sampling design with $\widetilde{\pi}_{l|i}$ and $\widetilde{\pi}_{lq|i,j}$ (>0) as the inclusion probabilities for the lth unit of the ith stratum, and lth unit of the ith stratum and qth unit of the jth stratum, respectively; $l = 1,\ldots, N_i$; $q = 1,\ldots, N_j$; $i, j = 1,\ldots, K$. In this case the expressions for the HTE for $\overline{Y}$ and its variance are respectively given as follows:

$$\widehat{\overline{Y}}_{ht}^{u} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\frac{\gamma_{ij}}{\widetilde{\pi}_{j|i}}I_{sj|i} \tag{7.8.4}$$

where $I_{sj|i} = 1$, if jth unit of the ith stratum is selected in the sample s and $I_{sj|i} = 0$ otherwise.

$$V\left(\widehat{\overline{Y}}_{ht}^{u}\right) = \frac{1}{N^2}\left[\sum_{i=1}^{K}\left\{\sum_{l=1}^{N_i}\frac{1}{\widetilde{\pi}_{l|i}}\gamma_{il}^2 + \sum_{l'=1}^{N_i}\sum_{l=1}^{N_i}\left(\frac{\widetilde{\pi}_{ll'|ii}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{l'|i}}\right)\gamma_{il}\gamma_{il'}\right\} + \sum_{i\neq}^{K}\sum_{j=1}^{K}\left\{\sum_{l=1}^{N_i}\sum_{q=1}^{N_j}\left(\frac{\widetilde{\pi}_{lq|ij}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|j}}\right)\gamma_{il}\gamma_{jq}\right\} - Y^2\right] \tag{7.8.5}$$

The gain in efficiency of using stratified sampling over unstratified sampling is given by

$$G = V\left(\widehat{\overline{Y}}_{ht}^{u}\right) - V\left(\widehat{\overline{Y}}_{ht}^{st}\right) \tag{7.8.6}$$

here we will find the formula for estimating G through a sample $s^*(=s_1 \cup s_2 \cup \cdots \cup s_K)$ selected by the stratified sampling procedure.

The quantities $\sum_{i=1}^{K}\sum_{l=1}^{N_i}\frac{1}{\widetilde{\pi}_{l|i}}y_{il}^2$ and Y^2 can by estimated unbiasedly by

$$\sum_{i=1}^{K}\sum_{l\in s_i}\frac{1}{\pi_{l|i}\widetilde{\pi}_{l|i}}y_{il}^2 \text{ and } N^2\left[\left(\widehat{\overline{Y}}_{ht}^{st}\right)^2 - \widehat{V}\left(\widehat{\overline{Y}}_{ht}^{st}\right)\right] \tag{7.8.7}$$

respectively, where $\widehat{V}\left(\widehat{\overline{Y}}_{ht}^{st}\right)$ is an unbiased estimator of $V\left(\widehat{\overline{Y}}_{ht}^{st}\right)$.

Similarly, $\sum_{k=1}^{K}\sum_{l'=1}^{N_i}\sum_{l=1}^{N_i}\frac{\widetilde{\pi}_{ll'|ii}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{l'|i}}y_{il}y_{il'}$ and $\sum_{i\neq}^{K}\sum_{j=1}^{N}\left\{\sum_{l=1}^{N_i}\sum_{q=1}^{N_j}\frac{\widetilde{\pi}_{lq|ij}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|j}}y_{il}y_{jq}\right\}$ can be estimated unbiasedly by

$$\sum_{k=1}^{K}\sum_{l'\neq}\sum_{l\in s_i}\frac{\widetilde{\pi}_{ll'|ii}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{l'|i}\pi_{ll'|i}}y_{il}y_{il'} \text{ and } \sum_{i\neq}^{K}\sum_{j=1}^{K}\left\{\sum_{l\in s_i}\sum_{q\in s_j}\frac{\widetilde{\pi}_{lq|ij}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|j}\pi_{l|i}\pi_{q|j}}y_{il}y_{jq}\right\} \tag{7.8.8}$$

The expressions Eqs. (7.8.7) and (7.8.8) yield an unbiased estimator for $V\left(\widehat{\overline{Y}}_{ht}^{u}\right)$ as

$$\widehat{V}\left(\widehat{\overline{Y}}_{ht}^{u}\right) = \frac{1}{N^2}\left[\begin{array}{l}\sum_{i=1}^{K}\left(\sum_{l\in s_i}\frac{1}{\widetilde{\pi}_{l|i}\pi_{l|i}}y_{il}^2 + \sum_{l\neq}\sum_{q\in s_i}\frac{\widetilde{\pi}_{lq|ii}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|i}\pi_{lq|i}}y_{il}y_{iq}\right) \\ +\sum_{i\neq}^{K}\sum_{j=1}^{K}\left(\sum_{l\in s_i}\sum_{q\in s_j}\frac{\widetilde{\pi}_{lq|ij}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|j}\pi_{l|i}\pi_{q|j}}y_{il}y_{jq}\right)\end{array}\right] - \left[\left(\widehat{\overline{Y}}_{ht}^{st}\right)^2 - \widehat{V}\left(\widehat{\overline{Y}}_{ht}^{st}\right)\right] \tag{7.8.9}$$

Finally using Eqs. (7.8.6) and (7.8.9) an unbiased estimator for the gain in efficiency of stratified sampling over the unstratified is obtained as

$$\widehat{G} = \widehat{V}\left(\widehat{\overline{Y}}^{u}_{ht}\right) - \widehat{V}\left(\widehat{\overline{Y}}^{st}_{ht}\right)$$

$$= \frac{1}{N^2}\left[\begin{array}{l}\sum_{i=1}^{K}\left(\sum_{l\in s_i}\frac{1}{\widetilde{\pi}_{l|i}\pi_{l|i}}y_{il}^2 + \sum_{l\neq}\sum_{q\in s_i}\frac{\widetilde{\pi}_{lq|ii}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|i}\pi_{lq|i}}y_{il}y_{iq}\right)\\ +\sum_{i\neq}^{K}\sum_{j=1}^{K}\left(\sum_{l\in s_i}\sum_{q\in s_j}\frac{\widetilde{\pi}_{lq|i,j}}{\widetilde{\pi}_{l|i}\widetilde{\pi}_{q|j}\pi_{l|i}\pi_{q|j}}y_{il}y_{jq}\right)\end{array}\right] - \left(\widehat{\overline{Y}}^{st}_{ht}\right)^2 \tag{7.8.10}$$

7.8.1 Simple Random Sampling Without Replacement

In case both the stratified and unstratified samples are selected by SRSWOR method, we get

$$\pi_{l|i} = n_i/N_i, \pi_{lq|i} = n_i(n_i-1)/N_i(N_i-1); \widetilde{\pi}_{l|i} = n/N,$$

$$\widetilde{\pi}_{lq|i,j} = n(n-1)/N(N-1); \widehat{\overline{Y}}^{st}_{ht} = \sum_{i=1}^{K} W_i\overline{y}(s_i),$$

$$\overline{y}(s_i) = \frac{1}{n_i}\sum_{j\in s_i} y_{ij}, \widehat{\overline{Y}}^{u}_{ht} = \frac{1}{n}\sum_{i}\sum_{j\in s} y_{ij} = \overline{y}(s),$$

$$V\left(\widehat{\overline{Y}}^{st}_{ht}\right) = \sum_{i=1}^{K} W_i^2(1-f_i)S_{yi}^2\Big/n_i,$$

$$\widehat{V}\left(\widehat{\overline{Y}}^{st}_{ht}\right) = \sum_{i=1}^{K} W_i^2(1-f_i)s_{yi}^2\Big/n_i, V\left(\widehat{\overline{Y}}^{u}_{ht}\right) = \frac{(1-f)}{n}S_y^2, f_i = n_i/N_i,$$

$$f = n/N \text{ and } (N-1)S_y^2 = \sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(y_{ij}-\overline{Y}\right)^2.$$

Now $V\left(\widehat{\overline{Y}}^{u}_{ht}\right) = \frac{(1-f)}{n}\frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(y_{ij}-\overline{Y}\right)^2$ can be unbiasedly estimated through stratified sample by

$$\widehat{V}\left(\widehat{\overline{Y}}^{u}_{ht}\right) = \frac{(1-f)}{n}\frac{1}{N-1}\left[\sum_{i=1}^{K}\frac{N_i}{n_i}\sum_{j\in s_i} y_{ij}^2 - N\left\{\left(\widehat{\overline{Y}}^{st}_{ht}\right)^2 - \widehat{V}\left(\widehat{\overline{Y}}^{st}_{ht}\right)\right\}\right].$$

After simplification we get

$$\widehat{V}\left(\widehat{\overline{Y}}_{lt}^{u}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} W_i s_{yi}^2 + \frac{N-n}{n(N-1)} \sum_{i=1}^{K}\left[W_i\left(\overline{y}(s_i) - \widehat{\overline{Y}}_{lt}^{st}\right)^2 - W_i(1-W_i)s_{yi}^2 \Big/ n_i\right] \tag{7.8.11}$$

So, the estimated gain in efficiency is

$$\widehat{G} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} W_i s_{yi}^2 + \frac{N-n}{n(N-1)}\sum_{i=1}^{K}\left[W_i\left(\overline{y}(s_i) - \widehat{\overline{Y}}_{lt}^{st}\right)^2 - W_i(1-W_i)s_{yi}^2 \Big/ n_i\right] - \sum_{i=1}^{K} W_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_{yi}^2 \tag{7.8.12}$$

7.8.2 Probability Proportional to Size With Replacement Sampling

Suppose an unstratified sample of size n is selected by PPSWR method using normed size measure p_{ij} for the jth unit of the ith stratum. Then the HH estimator of the population mean $\overline{Y}$ is given by

$$\widehat{\overline{Y}}_{hh}^{u} = \frac{1}{nN}\sum_{r=1}^{n}\frac{y(r)}{p(r)} \tag{7.8.13}$$

where $y(r) = y_{ij}$ and $p(r) = p_{ij}$ if the rth draw selects the jth unit of the ith stratum.

The variance of $\widehat{\overline{Y}}_{hh}^{u}$ is given by

$$V\left(\widehat{\overline{Y}}_{hh}^{u}\right) = \frac{1}{N^2 n}\left(\sum_{i=1}^{K}\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}} - N^2\overline{Y}^2\right) \tag{7.8.14}$$

Let the HH estimator based on a stratified sample of size n be

$$\widehat{\overline{Y}}_{hh}^{st} = \frac{1}{N}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{r=1}^{n_i}\frac{y(r,i)}{p(r,i)}$$

where $y(r,i) = y_{ij}$, $p(r,i) = p_{ij}/P_i = p_{ij}^*$ if the rth draw selects the jth unit of the ith stratum and $P_i = \sum_{j=1}^{N_i} p_{ij}$.

Unbiased estimators of $\frac{1}{N^2 n}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}} = \frac{1}{N^2 n}\sum_{i=1}^{K}\frac{1}{P_i}\sum_{j=1}^{N_i}\frac{y_{ij}^2}{p_{ij}^*}$ and $\frac{\overline{Y}^2}{n}$ based on the stratified sample are given by $\frac{1}{N^2 n}\sum_{i=1}^{K}\frac{1}{n_i P_i}\sum_{r=1}^{n_i}\left(\frac{y(r,i)}{p(r,i)}\right)^2$ and

$\frac{\left(\hat{\bar{Y}}_{hh}^{st}\right)^2 - \hat{V}\left(\hat{\bar{Y}}_{hh}^{st}\right)}{n}$, respectively, where $\widehat{V}\left(\widehat{\bar{Y}}_{hh}^{st}\right) = \frac{1}{N^2}\sum_{i=1}^{K}\frac{1}{n_i(n_i-1)}$ $\sum_{r=1}^{n_i}\left(\frac{y(r,i)}{p(r,i)} - \widehat{Y}_i\right)^2$ is an unbiased estimator of $V\left(\widehat{\bar{Y}}_{hh}^{st}\right)$ and $\widehat{Y}_i = \frac{1}{n_i}$ $\sum_{r=1}^{n_i}\frac{y(r,i)}{p(r,i)}$. Hence an unbiased estimator for $V\left(\widehat{\bar{Y}}_{hh}^{u}\right)$ based on the stratified sample is given by

$$\widehat{V}\left(\widehat{\bar{Y}}_{hh}^{u}\right) = \frac{1}{N^2 n}\sum_{i=1}^{K}\frac{1}{n_i P_i}\sum_{r=1}^{n_i}\left(\frac{y(r,i)}{p(r,i)}\right)^2 - \frac{1}{n}\left[\left(\widehat{\bar{Y}}_{hh}^{st}\right)^2 - \widehat{V}\left(\widehat{\bar{Y}}_{hh}^{st}\right)\right]$$

The expression of the estimated gain in efficiency of stratified sampling over unstratified sampling under PPSWR sampling design is

$$\begin{aligned}\widehat{G} &= \widehat{V}\left(\widehat{\bar{Y}}_{hh}^{u}\right) - \widehat{V}\left(\widehat{\bar{Y}}_{hh}^{st}\right) \\ &= \frac{1}{N^2 n}\left[\sum_{i=1}^{K}\frac{1}{n_i P_i}\sum_{r=1}^{n_i}\left(\frac{y(r,i)}{p(r,i)}\right)^2 - N^2\left(\widehat{\bar{Y}}_{hh}^{st}\right)^2\right] - \left(1-\frac{1}{n}\right)\widehat{V}\left(\widehat{\bar{Y}}_{hh}^{st}\right)\end{aligned} \tag{7.8.15}$$

Example 7.8.1

The garment factories of the KZN province of South Africa are classified into three strata. To estimate the average earnings of the garments, samples of factories are selected by SRSWOR method from each of the strata. Estimate the average earnings of garments per factory and compute its standard error. Estimate the gain in efficiency of stratified sampling over simple random sampling of the same sample size.

Factories	Number of factories (N_i)	Sample size (n_i)	Earnings (000 $) ($y_{ij}$)
Small	50	5	150, 100, 75, 125,175
Medium	100	10	750, 800, 950, 1000, 750, 750, 1200, 1500, 1200, 1500
Large	75	8	3000, 4000, 5000, 4500

To estimate the gain in efficiency, we prepare the following table.

Factories	N_i	n_i	Sample mean (000 $) $\bar{y}_i$	Sample SD s_{yi} (000 $)	$W_i s_{yi}^2$	$W_i(\bar{y}_i - \bar{y}_{st})^2$	$W_i(1 - W_i)\frac{s_{yi}^2}{n_i}$
Small	50	5	125	39.52	347.22	672,800	54.01
Medium	100	10	1040	297.02	39,209.88	302,500	2178.32
Large	75	8	4125	583.91	243,055.55	1,702,533.333	20,254.63
Total	225	23	–	–	282,612.65	2,677,833.333	22,486.97

Estimated mean income per factory $= \bar{y}_{st} = \sum_{i=1}^{3} N_i \bar{y}_i / N = \$\,1,865,000$

Estimated variance of stratified sampling $= \widehat{V}(\bar{y}_{st}) = \sum_{i=1}^{3} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_{yi}^2 = 10,629.35$

Estimated variance of the sample mean based on the stratified sampling is

$$\widehat{V}(\bar{y}_{wor}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_i W_i s_{yi}^2 + \frac{N-n}{(N-1)n} \left\{ \sum_i W_i (\bar{y}_i - \bar{y}_{st})^2 - \sum_i W_i (1 - W_i) s_{yi}^2 \Big/ n_i \right\}$$

$$= (1/23 - 1/225) \times 282,612.65 + \frac{225 - 23}{(224)23} \times (2,677,833.33 - 22,486.97)$$

$$= 115,142.5$$

So, the percentage gain in efficiency of stratified sampling $= \left\{ \widehat{V}(\bar{y}_{wor}) \big/ \widehat{V}(\bar{y}_{st}) - 1 \right\} \times 100 = 983.25\%$

Example 7.8.2

To estimate the total production of apples (in 00 kg) in a region, the farms were classified into four strata. From each of the strata, samples of farms

were selected by PPSWR method using farm size (in acres) as measure of size variable. The details are given as follows:

Types of farm	Number of farms (N_i)	Sample size	Average farm size (in acre) $\overline{X}_i$	Production (y) and farm size (x)				
Small	30	4	3	x	2	4	2	4
				y	25	30	20	40
Medium	50	8	4	x	2	5	5	4
				y	30	40	60	30
				x	4	4	5	4
				y	50	50	60	40
Large	20	4	5	x	5	4	5	4
				y	60	80	60	50

(i) Estimate the total production of apples and estimate the standard error of the estimator used.

(ii) Estimate the gain in efficiency of stratified sampling over unstratified sampling that might be selected by the PPSWR method of the same sample size of 16.

The HH estimate for the total production of apple Y based on the stratified sample is given by

$$\widehat{Y}_{hh}^{st} = \sum_{i=1}^{3} \widehat{Y}_i = 900 + 2237.5 + 1412.5 = 4550$$

where $\widehat{Y}_i = \dfrac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}/p_{ij}^*$, $p_{ij}^* = x_{ij}/X_i$ and $X_i = N_i \overline{X}_i$

Estimated standard error of $\widehat{Y}_{hh}^{st} = \sqrt{\sum_{i=1}^{3} \widehat{V}\left(\widehat{Y}_i\right)}$ with

$$\widehat{V}\left(\widehat{Y}_i\right) = \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{n_i} \left(\frac{y_{ij}}{p_{ij}^*} - \widehat{Y}_i \right)^2$$

We find $\widehat{V}\left(\widehat{Y}_1\right) = 8437.50$, $\widehat{V}\left(\widehat{Y}_2\right) = 31,763.39$ and $\widehat{V}\left(\widehat{Y}_3\right) = 38,489.58$

Hence standard error of $\widehat{Y}_{hh}^{st}$ is $\sqrt{\sum_{i=1}^{3} \widehat{V}\left(\widehat{Y}_i\right)} = \sqrt{78,690.47} = 280.52$

The HH estimator based on an unstratified PPSWR sample of size n is given by

$$\widehat{Y}_{hh}^{u} = \frac{1}{n} \sum_{i=1}^{3} \sum_{j=1}^{N_i} y_{ij}/p_{ij} \text{ with } p_{ij} = x_{ij}/X \text{ and } X = \sum_{i=1}^{3} X_i = 390.$$

The variance of $\widehat{Y}^{u}_{hh}$ is $V\left(\widehat{Y}^{u}_{hh}\right) = \frac{1}{n}\left(\sum_{i=1}^{3}\sum_{j=1}^{N_i} y_{ij}^2 \Big/ p_{ij} - Y^2\right) = \frac{1}{n}\left(\sum_{i=1}^{3}\frac{X}{X_i}\sum_{j=1}^{N_i} y_{ij}^2 \Big/ p_{ij}^{*} - Y\right)^2$

An unbiased estimate of $\sum_{i=1}^{3}\frac{X}{X_i}\sum_{j=1}^{N_i} y_{ij}^2 \Big/ p_{ij}^{*}$ is $\sum_{i=1}^{3}\frac{X}{X_i}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2 \Big/ (p_{ij}^{*})\right) =$ 3,619,687.5 +10,196,062.5 + 8,231,438 = 22,047,188 and an unbiased estimate of Y^2 is $\left(\widehat{Y}^{st}_{hh}\right)^2 - \widehat{V}\left(\widehat{Y}^{st}_{hh}\right) = (4550)^2 - 78,690.47 = 20,623,809.53$. Hence an unbiased estimate of $V\left(\widehat{Y}^{u}_{hh}\right)$ based on the stratified sample is $\widehat{V}\left(\widehat{Y}^{u}_{hh}\right)$ (22,047,188−20,623,809.53)/ 16 = 88,961.15. An estimate of gain in efficiency is $\left(\frac{\hat{V}\left(\hat{Y}^{u}_{hh}\right)}{\hat{V}\left(\hat{Y}^{st}_{hh}\right)} - 1\right) \times 100\% = 13.05\%$.

7.9 POSTSTRATIFICATION

The principle of stratification is to divide the entire population into a number of homogenous strata and then select samples from the each of the strata by some suitable sampling procedure. Sometimes, stratification is not possible before selection of sample for various reasons, e.g., (i) variability of the character under study is not known, (ii) in a multicharacter survey, information about more than one character is studied at a time. In this case, stratification with respect to one character may make the strata homogeneous with respect to that character only, but it makes them more heterogeneous with respect to the other character, and (iii) stratum sizes and sampling frame for each stratum may not be available. For example, in an HIV prevalence study, reliable data on the number of people in different age groups may be available from the past census or surveys, but it is difficult to get a list of people belonging to different age groups. In this situation, one can select a sample from the entire population first and then stratify the sample, looking at the values of the study variable under consideration. The technique of stratifying the population after the selection of the sample is known as poststratification.

Let a sample s of size n be selected from a finite population of size N by SRSWOR method, and let the sample be poststratified into K strata. Let n_i be the number of units falling into the ith stratum and $\overline{y}_i$ be the sample mean, $i = 1,\ldots, K$. Here n_i's are random variables subject to $\sum_{i=1}^{K} n_i = n$. We assume that n is so large that $P(n_i = 0) \cong 0$ for every i.

Theorem 7.9.1

(i) $\overline{y}_{pst} = \sum_{i=1}^{K} W_i \overline{y}_i$ is unbiased for the population mean $\overline{Y}$ when stratum weights W_i's are known.

(ii) $$V\left(\overline{y}_{pst}\right) = \sum_{i=1}^{K} W_i^2 \left[E\left(\frac{1}{n_i}\right) - 1\right] S_{yi}^2$$

$$\cong \frac{N-n}{Nn} \sum_{i=1}^{K} W_i S_{yi}^2 + \frac{1}{n^2} \sum_{i=1}^{K} (1 - W_i) S_{yi}^2$$

(iii) An unbiased estimator of $V\left(\overline{y}_{pst}\right)$ is

$$\widehat{V}\left(\overline{y}_{pst}\right) = \sum_{i=1}^{K} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_{yi}^2$$

where s_{yi}^2 is the sample variance for the ith stratum.

Proof

(i) $$E\left(\overline{y}_{pst}\right) = E\left\{ \sum_{i=1}^{K} W_i E\left(\overline{y}_i \middle| n_i\right) \right\} = E\left(\sum_{i=1}^{K} W_i \overline{Y}_i \right) = \overline{Y}$$

(ii) $$V\left(\overline{y}_{pst}\right) = E\left\{ V\left(\overline{y}_{pst} \middle| n_i\right) \right\} + V\left\{ E\left(\overline{y}_{pst} \middle| n_i\right) \right\}$$

$$= E\left\{ \sum_{i=1}^{K} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_{yi}^2 \right\} + V\left(\sum_{i=1}^{K} W_i \overline{Y}_i \right)$$

$$= \sum_{i=1}^{K} W_i^2 \left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i}\right] S_{yi}^2$$

Now writing $\in_i = \dfrac{n_i - \gamma_i}{\gamma_i}$ with $\gamma_i = E(n_i)$, we approximate

$$E\left(\frac{1}{n_i}\right) = \frac{1}{\gamma_i} E(1 + \in_i)^{-1} \cong \frac{1}{\gamma_i} E\left(1 - \in_i + \in_i^2\right) = \frac{1}{\gamma_i}\left(1 + \frac{V(n_i)}{\gamma_i^2}\right) \tag{7.9.1}$$

Furthermore, noting that n_i follows a hypergeometric distribution with $\gamma_i = E(n_i) = nW_i$ and $V(n_i) = nW_i(1 - W_i)\dfrac{N - n}{N - 1} \cong nW_i(1 - W_i)$ (ignoring f.p.c), we get

$$E\left(\frac{1}{n_i}\right) \cong \frac{1}{nW_i} + \frac{1 - W_i}{(nW_i)^2} \tag{7.9.2}$$

The approximation (7.9.2) was derived by Stephan (1945). Substituting Eq. (7.9.2) in Theorem 7.9.1, we find an approximate expression of $V\left(\overline{y}_{pst}\right)$ as follows:

$$V\left(\overline{y}_{pst}\right) = \frac{N - n}{Nn}\sum_{i=1}^{K} W_i S_{yi}^2 + \frac{1}{n^2}\sum_{i=1}^{K}(1 - W_i)S_{yi}^2 \tag{7.9.3}$$

(iii) $$E\left[\widehat{V}\left(\overline{y}_{pst}\right)\right] = E\left[\sum_{i=1}^{K} W_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)E\left(s_{yi}^2 \middle| n_i\right)\right]$$

$$= \sum_{i=1}^{K} W_i^2\left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i}\right]S_{yi}^2$$

The first term in Eq. (7.9.3) is the variance of the stratified sample mean $\overline{y}_{st}$ under proportional allocation. The second term is a positive quantity, which may be defined as the cost of poststratification. Thus post-stratification is always less efficient than stratified sampling with proportional allocation. However, for large n, the second term is expected to be much smaller compared to the first one. Hence, poststratified sampling is almost as efficient as stratified sampling if the sample size is very large.

7.10 EXERCISES

7.10.1 Consider a population of $N = nk$ units where the variable under study y possesses a linear trend $y_j = \alpha + \beta j$, $i = 1,\ldots, N$. The population is stratified into n strata of k units each. The ith $(=1,\ldots, k)$ stratum consists of units with labels $(i - 1)k + 1,\ldots, ik$. Let $\overline{y}_{st}$, $\overline{y}_{wor}$, and $\overline{y}_{sys}$ be the sample means each of size n, selected by stratified simple random sampling with one unit per stratum, SRSWOR sampling, and systematic sampling from the entire population, respectively. Prove that $V\left(\overline{y}_{st}\right) \le V\left(\overline{y}_{sys}\right) \le V\left(\overline{y}_{wor}\right)$.

7.10.2 Suppose an infinite population $f(y) = \theta e^{-y\theta}$; $y > 0$ (θ is a known positive constant) is divided into two strata by taking a point y_0 in the range of the study variable y. Samples of sizes n_1 and $n_2(=n - n_1)$ are selected from stratum one and stratum two, respectively, by SRSWR method. Find the optimum point of stratification, which minimizes the variance of the estimator of the population mean $\overline{Y}$ under (i) proportional and (ii) optimum allocation, when the sample size n is fixed. Compute the relative efficiency of the optimal allocation over the proportional allocation.

7.10.3 Consider a population consisting of two strata with sizes N_1 and N_2. Let V_{prop} and V_{opt} denote the variances of the estimator of the population mean for the proportional and optimum allocation under the cost function $C = c_0 + c_1 n_1 + c_2 n_2$, respectively, based on SRSWR samples from each of the stratum. Show that when the total cost of the survey C is fixed, strata variances are equal

$$\frac{V_{prop}}{V_{opt}} = \frac{N(N_1 c_1 + N_2 c_2)}{\left(N_1\sqrt{c_1} + N_2\sqrt{c_2}\right)^2}$$

7.10.4 A population is stratified into K strata. From the $i(=1,\ldots, K)$th stratum a sample of n_i units is selected from the N_i units by SRSWOR method and let $\overline{y}_i$ be the sample mean. Show that the optimum n_i that minimizes the variance of $T = \sum_{i=1}^{K} l_i \overline{y}_i$ for a given cost C_0 under the cost function $C = c_0 + \sum_{i=1}^{K} c_i n_i$ is $n_i = \dfrac{(C_0 - c_0)}{\left(\sum_{i=1}^{k} |l_i| S_i \sqrt{c_i}\right)} \dfrac{|l_i| S_i}{\sqrt{c_i}}$. Hence, show that the optimum values of n_i and n_j that minimizes the variance of $\overline{y}_i - \overline{y}_j$ are given by $n_i = \dfrac{(C_0 - c_0)}{\left(\sum_{i=1}^{k} S_i \sqrt{c_i}\right)} \dfrac{S_i}{\sqrt{c_i}}$ and $n_j = \dfrac{(C_0 - c_0)}{\left(\sum_{i=1}^{k} S_j \sqrt{c_j}\right)} \dfrac{S_j}{\sqrt{c_j}}$, respectively.

7.10.5 The following data relate to the household size, monthly income, and ownership of the house of 50 households in South Africa.

Serial number of HH	HH size	HH income ($)	Ownership of house	Serial number of HH	HH size	HH income ($)	Ownership of house
1	1	18,562	0	26	5	11,526	1
2	1	10,864	1	27	5	10,778	0
3	4	8456	1	28	5	17,843	1
4	5	11,990	0	29	1	17,183	0
5	4	16,449	1	30	5	6543	0
6	4	13,563	1	31	5	10,009	0
7	3	2151	0	32	1	19,592	0
8	1	20,786	0	33	4	20,473	1
9	4	16,738	0	34	1	6662	1
10	4	18,003	0	35	1	7052	0
11	5	22,073	0	36	1	16,071	0
12	1	8104	0	37	4	14,222	0
13	3	1565	0	38	2	24,092	0
14	4	20,049	1	39	5	12,970	1
15	5	10,770	0	40	1	23,993	1
16	1	14,678	1	41	1	21,026	0
17	1	5674	1	42	1	17,053	0
18	1	19,169	0	43	4	11,736	1
19	1	24,088	0	44	3	5816	0
20	2	8938	0	45	2	2167	0
21	3	19,090	1	46	3	23,211	1
22	4	13,141	1	47	2	22,293	1
23	3	23,849	1	48	5	9189	0
24	1	9728	1	49	1	5443	1
25	1	12,605	1	50	4	23,163	1

Stratify the 50 households with their income (a) less than $10000 (stratum 1), (b) $10001 to $15000 (stratum 2), and (c) above $15000 (stratum 3).

(i) From each of the stratum select a sample of five households by SRSWOR method. Estimate the average income of the 50 households. Find the standard error of your estimator. Compare the relative efficiency of your estimator with an unstratified sample mean based on an SRSWOR of the same sample size of 15.

(ii) Select samples of sizes 5 from stratum 1, 6 from stratum 2, and 5 from stratum 3 by using SRSWR, SRSWOR, and PPSWR methods, respectively, taking household size as a measure of size variable. From the selected samples, find an unbiased estimate of the proportion of persons owning a house. Estimate the standard error of the estimate and obtain a 90% confidence interval for the proportion of households owning a house.

(iii) Give the allocation of a total of 16 households using proportional and Neyman allocation.

7.10.6 The following sample of 23 factories selected by SRSWOR method from 50 small, 75 medium, and 25 large factories in a

certain region. Information regarding number of workers, output, and membership of a medical aid is also listed.

Small factories				Medium factories			
Serial number of factories	Number of workers	Daily output	Medical aid facility	Serial number of factories	Number of workers	Daily output	Medical aid facility
1	21	885	1	1	29	1021	1
2	20	981	0	2	48	1184	0
3	18	762	1	3	40	1198	0
4	22	539	1	4	50	1112	1
5	14	777	0	5	47	1404	0
6	13	648	0	6	33	1175	0
7	17	972	1	7	50	1286	0
8	16	919	0	8	48	1319	1
				9	45	1358	1
				10	50	18,668	1

Large factories			
Serial number of factories	Number of workers	Daily output	Medical Aid facility
1	70	1791	1
2	58	2455	1
3	89	2659	1
4	79	2342	0
5	74	1743	1

(i) Estimate the average daily output of the 150 factories of the region and find a 90% confidence interval of the average output.

(ii) Estimate the proportion of factories having a medical aid facility. Find a 95% confidence interval of this proportion.

(iii) Estimate the average number of workers per factory and obtain its standard error. Estimate the gain in efficiency of stratified sampling over unstratified sampling for estimating the average number of workers.

(iv) Estimate the average daily output of the factories having a medical aid facility when it is known that 40% of the factories provide medical facilities. Estimate the standard error of your estimator.

7.10.7 To estimate the total production of maize in a certain region, the farms were classified in to four strata. It is decided that a total of 50 farms will be selected using (i) proportional, (ii) Neyman, (iii) allocation proportional to the average size of farm, and (iv) optimal allocation. Compare the relative efficiencies of the allocations for estimating the total output if SRSWOR method is used from each of the strata using the following past survey data:

Strata	Number of farms	Average size of the farm (in acre)	Standard deviation of maize production (in 000 kg)	Cost of surveying a factory (in $)
I	50	10	10	50
II	75	15	20	60
III	100	20	20	75
IV	25	25	25	100

7.10.8 The following table gives a summary result of an agricultural survey for estimating the production of milk using SRSWR method of sampling from the each type of farms.

Types of farm	Number of farms	Sampled farms	Production of milk per sampled farm (00 in liter)	Sample standard deviation of production
Small	50	10	125	30
Medium	100	25	200	30
Large	175	30	200	50

(i) Estimate total production of milk and its standard error.

(ii) Estimate the gain in efficiency of the stratified sampling with respect to unstratified sampling based on SRSWR method of the same sample size.

(iii) Compare relative efficiencies of estimating the total production of milk of the present allocation with (i) proportion and (ii) Neyman's allocation keeping the total sample size fixed.

7.10.9 To estimate the average marks of a STA112 course, which comprises of four groups of students in the University of Botswana, samples from each of the groups were selected by PPSWR method using the CA marks as a measure of size variable. Details are given in the following table:

(i) Estimate the average examination marks and find a 90% confidence interval of the average marks.

Groups	Number of students	Sampled students	Average CA		Marks on CA (x) and Exam (y) obtained from the students selected in the sample									
A	75	10	60	x	50	60	60	80	80	25	30	80	70	40
				y	55	50	70	90	95	30	50	70	75	20
B	50	5	75	x	30	70	60	45	30					
				y	40	75	55	60	50					
C	75	8	60	x	90	25	40	80	35	65	60	20		
				y	98	30	50	70	50	50	80	30		
D	75	5	65	x	50	20	10	60	70					
				y	60	25	40	70	90					

(ii) Estimate the gain in efficiency of the stratified sampling over unstratified PPSWR sampling from the entire population with the same sample size.

(iii) Based on the selected sample, compare efficiencies of the present allocation with (a) proportional allocation and (b) allocation proportional to the average CA marks for estimating the average examination marks.

7.10.10 A sample of 200 students was selected from a class of 800 students by SRSWOR method, and their marks in the STAT 121 course were recorded. The students were then classified according to their grades. The following table gives the sample means and standard deviations along with the grades of 800 students obtained from the university record.

		Sampled students		
Grade	Number of students	Number	Average marks	Standard deviation
A	80	20	82	10
B	120	35	72	15
C	350	100	65	12
D	150	30	52	10
E	100	15	30	20

(i) Estimate the average marks of the STAT 121 course.

(ii) Give an appropriate estimate of the standard error of your estimator used in "(i)."

7.10.11 A simple random sample of 500 households was selected from an urban community of 5000 households. The following table shows a summary of the survey along with the distribution of the age of the community from the past census report.

		Sampled people			
Age group	Number of people (obtained from census)	Number	Average income ($)	Standard deviation ($)	HIV prevalence rate
Below 18	1000	80	500	100	0.10
19–25	1500	150	2500	50	0.25
26–45	1500	170	8000	40	0.40
46–60	700	80	12,000	60	0.30
Above 60	300	20	1500	80	0.10

(i) Estimate the mean income of the community and also obtain the 90% confidence interval of the mean income.

(ii) Estimate the HIV prevalence rate for the entire community and its standard error.

7.10.12 The following table gives the income distribution of 1500 households of a certain district. Make five strata using (i) Dalenius and Hodges and (ii) Gunning and Horgan methods.

Household income (US$)	Below 1000	1001 to 2000	2001 to 3000	3001 to 4000	4001 to 5000	5001 to 7000	Above 7000
Frequency	200	350	400	250	150	100	50