Chapter 3

Review of Literature

Summary

In this chapter, we review the literature on the problem of allocation of sample size to strata in stratified sampling specially under superpopulation models by earlier authors. First the pioneering works of Hanurav and Rao are reviewed and are followed by contributions of other authors in the field. Power allocations and Lui's (1992-93) [31] unconditional expectation approach are also briefly mentioned at the end.

3.1 TN Optimum Allocation

Allocation of sample size to strata consists of distributing the total sample size among different strata and is one of the major problems in stratified sampling. Bowley (1926) [2] suggested allocation proportional to stratum sizes but this does not take into consideration the within stratum variability.
Tschuprow (1923) [80] was the first to derive the optimum allocation for allocation of the total sample size to different strata which minimises the sampling variance of the usual Unbiased Expansion Estimator (UEE) of the population total. But unfortunately his proof remained unknown until Neyman (1934) [40] derived the same minimum variance optimum allocation independently in his celebrated paper. That is the reason why the optimum allocation is popularly known as Neyman’s Optimum Allocation. We would however prefer to call the allocation by Tschuprow-Neyman (TN) Optimum Allocation (TNOA) in this thesis.

TNOA was derived for simple random sampling with or without replacement (SRSWR or SRSWOR) under the following simple cost function with equal cost per unit in all the strata.

\[ C_t = C_{t0} + c_t n \]  

(3.1.1)

where \( C_t \) is the total cost of the survey, \( C_{t0} \) the overhead cost and \( c_t \) is the cost per unit of collecting information on the sampled units. For SRSWR under the cost function (3.1.1) the TNOA is given by

\[ n_{i}^{TN} = n \cdot \frac{N_i \sigma_i}{\sum N_i \sigma_i} \]  

(3.1.2)

Later Mahalanobis (1944) [34] considered a more general cost function with unequal cost per unit in different strata, which is given by
\[ C_t = C_{t0} + \sum_i c_{ti} n_i \]  \hspace{1cm} (3.1.3)

where \( c_{ti} \) is the cost per unit of survey in the \( i^{th} \) stratum.

Mahalanobis (1944) [34] optimum allocation for SRSWR within each stratum under the cost function (3.1.3) is given by

\[ n_i^M = \frac{n_i \sigma_i / \sqrt{c_{ti}}}{\sum_i N_i \sigma_i / \sqrt{c_{ti}}} \]  \hspace{1cm} (3.1.4)

The TNOA follow as a particular case of Mahalanobis optimum allocation when the cost per unit is the same in all strata. Stuart (1954) [70] noted that the optimum allocation can easily be derived by using Cauchy-Schwarz inequality.

The TNOA is usually derived using calculus method of Lagrange's multiplier. It is not however intuitively easy to interpret the allocation as to what the product of the stratum size \( N_i \) and the stratum standard deviation \( \sigma_i \) means. Alternative derivations of the TNOA which does not require the calculus method of Lagrange's multiplier were given by Rao (1984, 94) [51], [55] and Gupt and Rao (1994) [15].

42
3.2 Model Based Approach to Allocation Problem

Hanurav (1965) [23] was the first to use the superpopulation model approach for studying the problem of allocation of sample size to strata followed by Rao (1968) [49]. The TNOA involves unknown parameters viz., within stratum standard deviations (WSSDs) of the study variable \( \mathcal{Y} \). In order to compute the TNOA one needs at least the proportionate values of the unknown WSSDs of the study variate \( \mathcal{Y} \).

In practice, the known WSSDs of an available auxiliary variable are usually substituted in place of the unknown WSSDs. The justification for the assumption that the unknown WSSDs of the study variate are not far from the known WSSDs of the auxiliary variable was examined first by Hanurav (1965) [23] in the light of the prior distributions under the following superpopulation model \( \Delta_2 \), the class of prior distributions \( \delta_2 \) satisfying

\[
\begin{align*}
(i) & \quad \mathcal{E}_{\delta_2}(Y_{ij}|X_{ij}) = \beta X_{ij} \\
(ii) & \quad \mathcal{V}_{\delta_2}(Y_{ij}|X_{ij}) = \sigma^2 X_{ij}^2 \\
(iii) & \quad \mathcal{C}_{\delta_2}(Y_{ij}|Y_{ij'},X_{ij},X_{ij'}) = 0.
\end{align*}
\]

He found that the substitution of \( \sigma_i(x) \) in place of \( \sigma_i \) of TNOA is justified only if the coefficients of variation of the auxiliary variable \([CV(x)]\) are more or less equal in all strata. He further opined that if this condition is far from
being fulfilled then there is not much point in using the relative values of $\sigma_i^2(x)$'s in place of $\sigma_i^2$'s in TNOA. The result due to Hanurav (1965) [23] is stated in the following theorem:

**Theorem 3.2.1.** [Hanurav (1965)] [23] The Tschuprow-Neyman optimum allocation for stratified SRSWR design assuming large stratum sizes reduces under the model $\Delta_2$ specified by (3.2) to the following allocation.

$$n_i(2) \propto N_i \sigma_i(x)$$  \hspace{1cm} (3.2.1)

where $\sigma_i(x)$ is the WSSD in the $i^{th}$ stratum of the auxiliary variable $X$ provided its coefficient of variation

$$C_i(x) = \sigma_i(x)/\bar{X}_i$$  \hspace{1cm} (3.2.2)

are equal in all strata.

**Definition 3.2.2.** The allocation (3.2.1) is called Hanurav's Auxiliary Variable Optimum Allocation (HAVOA).

Rao (1968) [49] considered the problem under a general superpopulation model. Let $\Delta_\gamma$ be the class of all prior distributions $\delta_\gamma$ for which
(i) \[ \mathcal{E}_\delta(Y_{ij}|X_{ij}) = \beta X_{ij} \]

(ii) \[ \mathcal{V}_\delta(Y_{ij}|X_{ij}) = \sigma^2 X_{ij}^2 \]

(iii) \[ C_\delta(Y_{ij}, Y_{ij'}|X_{ij}, X_{ij'}) = 0. \]

The results are stated in the following two theorems due to Rao (1968) [49].

**Theorem 3.2.3.** [Rao(1968)] [49] Under the superpopulation model \( \Delta_\delta \), Tschuprow-Neyman optimum allocation reduces to allocation proportional to \( \sqrt{T_i^2 - \gamma_i(g)} \), where

\[ T_i = \sum_j X_{ij} \]

is the total of the auxiliary variable for the \( i^{th} \) stratum and

\[ \gamma_i(g) = N_i \left( \sum_j X_{ij}^2 - \sum_j X_{ij}^2 \right) \]

provided \( \gamma_i(g) > 0 \), when the corrected coefficient of variation of the \( X \)-character are equal in all strata.

**Theorem 3.2.4.** [Rao(1968)] [49] Tschuprow-Neyman optimum allocation reduces to allocation proportional to the stratum totals of the auxiliary vari-
able \( X \) under \( \Delta_2 \), when the coefficients of variation of \( X \)-character are equal in all strata.

Further Rao (1968) [49] considered the problem of allocation of sample size to strata when sample is selected with inclusion probability proportional to size (\( \pi PS \)) design within each stratum using Horvitz-Thompson (HT) Estimator (HTE) for estimating the population total under the superpopulation model \( \Delta_g \). Suppose a \( \pi PS \) (\( \pi \lambda \), the probability of inclusion of \( \lambda \)th unit, being proportional to size \( X \)) sample of size \( n \) is taken from the \( i \)th stratum, such that

\[
\sum_{i=1}^{k} n_i = n.
\]

Let \( \pi_{ij} \) denote the probability of inclusion of \( j \)th unit of the \( i \)th stratum in the sample. The Horvitz-Thompson estimator of the population total is given by

\[
\hat{Y}_{st,HT} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{y_{ij}}{\pi_{ij}}.
\]

We state below the three theorems due to Rao (1968)[49]:

**Theorem 3.2.5.** [Rao(1968)][49] Under the superpopulation model \( \Delta_g \), the allocation of the sample size to the strata which minimizes the expected variance of \( \hat{Y}_{HT} \) is given by
\[ n_i = n \sqrt{\frac{X_i \sum_{j=1}^{N_i} X_{ij}^{g-1}}{\sum_{i=1}^{k} X_i \sum_{j=1}^{N_i} X_{ij}^{g-1}}} \]  

(3.2.4)

where \( X_i \) is the total of \( X \)-values of the \( i \)th stratum.

**Definition 3.2.6.** The allocation of sample size to strata given in equation (3.2.4) is called the \( \Delta_g \)-model optimum allocation (\( \Delta_g \)-MOA) for the strategy (St.\( \pi \) PS: \( \hat{Y}_{HT} \)).

**Theorem 3.2.7.** [Rao (1968)] [49] Under the superpopulation model \( \Delta_2 \), allocation of the sample size to strata proportional to the stratum totals of the \( X \)-variate minimizes the expected variance of \( \hat{Y}_{HT} \).

Finally Rao (1968) [49] has also compared unstratified \( \pi PS \) sampling with stratified \( \pi PS \) sampling under \( \Delta_g \)-optimum allocation. The result is given in the following theorem:

**Theorem 3.2.8.** [Rao (1968)] In the sense of expected variance, under \( \Delta_g \), unstratified \( \pi PS \) sampling is inferior to \( \pi PS \) stratified sampling with \( \Delta_g \)-optimum allocation and for \( g = 2 \), both the schemes are equivalent.

Rao (1977) [50] further considered a more general superpopulation model and derived the allocation of sample size which minimizes the expected variance of the strategy (St.\( \pi \) PS : \( \hat{Y}_{St.HT} \)) under the class of prior
distributions $\Delta_{(v)} = \{\delta_{(v)}\}$ for which we have,

\[
\begin{align*}
\mathcal{E}_{\delta_{(v)}}(Y_{ij} | X_{ij}) &= \beta X_{ij} \\
\mathcal{V}_{\delta_{(v)}}(Y_{ij} | X_{ij}) &= v(X_{ij}) \\
C_{\delta_{(v)}}(Y_{ij}, Y_{i'j'} | X_{ij}X_{i'j'}) &= 0.
\end{align*}
\]

The TNOA under the model $\Delta_{(v)}$ is given by

\[
n_i = n \frac{\sqrt{X_i \left( \sum_{j=1}^{N_i} \frac{v(X_{ij})}{X_{ij}} \right)}}{\sum_{i=1}^{k} \sqrt{X_i \left( \sum_{j=1}^{N_i} \frac{v(X_{ij})}{X_{ij}} \right)}}.
\]

**Definition 3.2.9.** The allocation given by (3.2.6) is called Rao’s $\Delta_{(v)}$ Model Optimum Allocation ($\Delta_{(v)}$-RMOA) for the strategy (St.$\pi$PS : $\hat{Y}_{STHT}$).

Rao (1977) [50] also generalized the Rao’s (1968) [49] theorem 3.2.7 for the allocation (3.2.6) as follows:

**Theorem 3.2.10.** [Rao(1977)] [50] In the sense of expected variance under the above model, unstratified $\pi$PS sampling strategy (with Horvitz-Thompson(HT) estimator) is inferior to stratified $\pi$PS sampling strategy (with the corresponding HT estimator) with $\delta_{(v)}$-optimum allocation.
Vijayan (1971) [cf. Rao (1977)] [50] also studied this problem for special cases of \( v(X_{ij}) \). It is, however, not known from the above result that under what condition unstratified \( \pi PS \) sampling is still inferior to stratified \( \pi PS \) sampling when one deviates from the above \((\Delta_{(v)}-RMOA)\) allocation. With this objective, Ramachandran and Rao (1974a) [44] investigated whether stratified \( \pi PS \) sampling with various non-optimal allocations is likely to be worthwhile. They considered deviations from the above allocation i.e., \( \delta_{(v)}\)-optimum allocation with \( v(X_{ij}) \propto X_{ij}^q \) which is realistic and of practical interest.

Ramachandran and Rao (1974b) [45] studied the problem of allocation when probability proportional to size with replacement \((PPSWR)\) is adopted within each stratum. The \( j^{th} \) unit of the \( i^{th} \) stratum is selected with initial probability

\[
P_{ij} = \frac{X_{ij}}{X_i}, \quad \text{where} \quad X_i = \sum_{j=1}^{N_i} X_{ij}. \tag{3.2.7}
\]

The Hansen-Hurwitz \((HH)\) Estimator \((HHE)\) of the population total based on the stratified PPSWR sampling design is given by

\[
\hat{Y}_{st.HH} = \sum_{i=1}^{k} \frac{X_i}{n_i} \sum_{j=1}^{n_i} \frac{y_{ij}}{x_{ij}}. \tag{3.2.8}
\]

Their result is given in the following theorem:

**Theorem 3.2.11.** [Ramachandran and Rao (1974b)] [45] Under the super-
population model $\Delta_g$, the allocation of sample size to strata that minimizes the expected variance of the strategy (SS. PPSWR: $\hat{Y}_{St.HH}$) is given by

$$n_i = n \frac{\sqrt{\sum_j X_{ij}^g (X_i X_i^{-1} - 1)}}{\sum_{i=1}^k \sqrt{\sum_j X_{ij}^g (X_i X_i^{-1} - 1)}}.$$ 

Their result on comparison between unstratified PPSWR and stratified PPSWR sampling designs is given in the following theorem:

**Theorem 3.2.12.** [Ramachandran and Rao (1974b)] [45] In the sense of expected strategy variance, under $\Delta_g$ unstratified PPSWR sampling strategy is worse than stratified PPSWR sampling strategy with $\Delta_g$-optimum allocation for estimating $Y$.

Earlier Raj (1963) [42] compared unstratified PPS sampling strategy with stratified PPSWR sampling strategy when AV-proportional allocation is used in terms of exact variance. Ramachandran and Rao (1974b) [45] generalized Desraj’s results by comparing unstratified PPSWR sampling strategy with stratified PPSWR sampling strategy when various non-optimal allocations are used in the sense of expected variance under the model $\Delta_g$.

We study the problem of allocation of sample size to strata for stratified PPSWR sampling strategy with actual variance under the model $\Delta_g$ in Chapter 5.
3.3 Power Allocations

Tschuprow-Neyman optimum allocation minimizes the sampling variance or equivalently the coefficient of variation of the overall population estimator. But it may result in estimates for some strata with higher variance than that for others. This may happen because the strata may vary considerably in population size or importance. This situation may not be desirable as the stratum estimates may also be needed with reasonably equal precision besides the overall population estimate with high precision. On the other hand an alternative allocation that achieves equal SCV for stratum estimates may produce overall population estimate with much higher CV than that under TNOA. In such a situation, Bankier (1988) [1] proposed a compromise allocation between these two types of allocations. He considered to minimize the loss function

\[ F = \sum_{i=1}^{k} \left( X_i^q C_i(\hat{Y}_i) \right)^2 \]

subject to the constraint \( \sum n_i = n \)

where \( X_i \) is some measure of size or importance of the \( i^{th} \) stratum and \( q \) is a constant in the range \( 0 \leq q \leq 1 \). This gives the Bankier allocation

\[ n_i \propto C_i(\hat{Y}_i) X_i^q. \]

The choice of \( q \) results in different allocations. For example when \( q = 1 \) and
$X_i = Y_i$ this reduces to the TNOA and when $q = 0$, this results in an allocation where the SCV's for the different strata are almost equal if the $C_i(y)$ do not vary significantly from stratum to stratum and the finite population correction (fpc) factors can be ignored. Thus allocation with $q$ between 0 and 1 can be viewed as a compromise between TNOA and the almost equal SCV's allocation. This allocation is also known as Power Allocation and $q$ is called the power of this allocation.

Rao (2001) [56] has extended Bankier allocation under the general cost function (3.1.3) and obtained what he called Mahalanobis-Bankier allocation:

$$n_i^{MB} \propto X_i^q \frac{C_i(y)}{\sqrt{c_{ti}}}.$$  

He further considered these allocations under the superpopulation model. These allocations are expected under the model $\Delta_2$ as follows:

$$\mathcal{E}_2(n_i^B) \propto C_i(x) X_i^q \propto X_i^q$$

if stratum SCV's of $\mathcal{X}$ are equal in all strata.

$$\mathcal{E}_2(n_i^{MB}) \propto C_i(x) \frac{X_i^q}{\sqrt{c_{ti}}} \propto \frac{X_i^q}{\sqrt{c_{ti}}}$$

if stratum SCV's of $\mathcal{X}$ are equal in all strata.
3.4 Unconditional Expectation Approach

Lui (1992-93) [31] considered the problem of estimation of sample size for ratio estimator and allocation of sample sizes to strata for separate ratio estimator under the superpopulation model by a slightly different approach. He assumes that the bivariate finite population on two characters \( X, Y \) \( \{(X_i, Y_i), \ i = 1, 2, \ldots, N\} \) is realization of a pair of random variables that are independently and identically distributed from a superpopulation model \( \Delta_t \).

For fixed \( X_i \)

\[
E_{\delta_1}(Y_i|X_i) = RX_i
\]

\[
V_{\delta_1}(Y_i|X_i) = \sigma^2 X_i
\]

\[
C_{\delta_1}(Y_i, Y_j|X_i, X_j) = 0.
\]

It is also assumed that the mean and the variance of the distribution of \( X_i \) either from a pilot study or previous survey i.e.,

\[
E_{\delta_1}(X_i) = \mu_x \text{ known}
\]

\[
V_{\delta_1}(X_i) = \sigma^2_x \text{ known for } \forall i = 1, 2, \ldots, N.
\]

It is well known that the ratio estimator
\[
\hat{Y}_R = \frac{\bar{y}}{\bar{x}} \cdot X
\]

based on a simple random sample of size \(n\) \(\{(x_i, y_i), \ i = 1, 2, \ldots, n\}\) is the best linear unbiased estimator (BLUE) of the population total \(Y\) under the above superpopulation model (Royall (1970)) [60].

Unconditional expectation of the sampling variance of \(\hat{Y}_R\), \(V(\hat{Y}_R)\) given by equation (2.5.6) as derived under the above model was demonstrated that

\[
E_{\delta_1} \left\{ V(\hat{Y}_R) \right\} = \left[ E_{\delta_1} E_{\delta_1} \left\{ V(\hat{Y}_R|X_i) \right\} \right]
\]

(3.4.1)

\[
= \frac{N^2(N-n)\sigma^2}{n(N-1)} \approx \frac{N^2\sigma^2\mu}{n}\n\]

if \(N\) were so large that \(\frac{N-n}{N-1} \approx 1\).

Using Chetyshev's inequality and on the basis of the expression (3.4.1), Lui estimated sample size for a given precision.

Further the problem of allocation of sample size to strata for separate ratio estimator was considered under the following superpopulation model: For the \(i^{th}\) stratum,
\[
\mathcal{E}_{\delta_i}(Y_{ij}|X_{ij}) = R_i X_{ij}
\]

\[
\mathcal{V}_{\delta_i}(Y_{ij}|X_{ij}) = \sigma_i^2 X_{ij}
\]

(3.4.2)

\[
\mathcal{C}_{\delta_i}(Y_{ij}, Y_{ij'}|X_{ij}, X_{ij'}) = 0
\]

with

\[
\mathcal{E}_{\delta_i}(X_{ij}) = \mu_i(x)
\]

and

\[
\mathcal{V}_{\delta_i}(X_{ij}) = \sigma_i^2(x)
\]

When SRSWOR is used within each stratum, the variance of the separate ratio estimator

\[
\hat{Y}_{SR} = \sum_i \frac{\bar{y}_i}{x_i} X_i
\]

is given by (2.5.12).
The unconditional expectation of the $V(Y_{SR})$ under the above model from the eq.(3.4.1) was demonstrated that

$$E_{\delta_1}\{V(\hat{Y}_{SR})\} = \sum_i \frac{N_i^2(N_i - n_i) \sigma_i^2 \mu_i(x)}{n_i(N_i - 1)}. \quad (3.4.3)$$

If the strata are so large that

$$\frac{N_i - n_i}{N_i - 1} \approx 1 \forall i,$$

then the variance (3.4.3) can be approximated by

$$E_{\delta_1}\{V(\hat{Y}_{SR})\} \approx \sum_i \frac{N_i^2}{n_i} \sigma_i^2 \mu_i(x). \quad (3.4.4)$$

The TNOA minimizing the $E_{\delta_1}\{V(\hat{Y}_{SR})\}$ for a given total sample size $n = \sum_i n_i$ under the above model was shown to be

$$n_i = n \frac{N_i \sqrt{\frac{N_i}{N_i - 1} \sigma_i^2 \mu_i(x)}}{\sum N_i \sqrt{\frac{N_i}{N_i - 1} \sigma_i^2 \mu_i(x)}}.$$

However, if the strata are so large that $\frac{N_i}{N_i - 1} \approx 1$ then the TNOA can be approximated by

$$n_i = n \frac{N_i \sigma_i \sqrt{\mu_i(x)}}{\sum N_i \sigma_i \sqrt{\mu_i(x)}}.$$
Note that when \( \sigma_i \)'s are all equal then the TNOA reduces to allocation proportional to \( N_i \sqrt{\mu_i(x)} \). This is one of the rules recommended by Cochran (1977) [8]. Lui pointed out that Cochran however did not explicitly pointed out that \( n_i \propto N_i \sqrt{\mu_i(x)} \) would be optimal allocation only when \( \sigma_i \)'s are constant for all strata.

The main drawback of Lui's approach is that it assumes the knowledge of superpopulation parameters \( \sigma_i^2 \)'s which may not be available at the planning stage of the survey. The distribution of each \( X_{ij} \)'s is assumed known particularly each \( X_{ij} \) has mean \( \mu_i(x) \) and variance \( \sigma_i^2(x) \) which are unknown. Therefore conditional approach of Hanurav (1965) [23] and Rao (1968,77) [49], [50] is followed to study the problem of allocation of sample size under superpopulation model approach for different sampling strategies.