Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 5, pp. 879–892
DOI : 10.47974/JSMS-1390

Allocations in stratified modified PPS sampling technique

Anupama Goyal [†]
Anju Goyal *
Sangeeta Arora [§]
Department of Statistics
Panjab University
Chandigarh 160014
India

Abstract

Stratified sampling enhances the efficiency of estimators by dividing the heterogenous population into the homogeneous strata. In the literature, Stratified Modified Probability Proportional to Size (Stratified MPPS) sampling is a method that selects units within each stratum using MPPS sampling. MPPS sampling is used when the data follows Zipf's law. In this paper, the allocation methods of sample size among the strata are applied to Stratified MPPS sampling to minimize the variance. Proportional allocation assigns sample sizes to strata in proportion to their sizes and Neyman allocation considers both stratum size and variability among strata to allocate the sample size among various strata. An Empirical study based on real-world datasets, compares the proposed allocation methods to estimate the population total in statistical practice.

Subject Classification: *62D05, 62F40, 62G09, 94A20.*

Keywords: *With and without replacement sampling, Probability proportional to size (PPS), Stratified sampling, Proportional and neyman allocation, Modified PPS (MPPS), Zipf's law.*

1. Introduction

Stratified sampling reduces sampling variability and enhances precision by ensuring adequate representation of each stratum in the sample. Goodman and Kish [5] proposed the technique of controlled selection to increase the probability of selection of preferred samples. The allocation of sample units across strata is a critical aspect of Stratified sampling design. Various allocation

[†] *E-mail:* **anupamagoyal6198@gmail.com**
* *E-mail:* **anju.statistics@gmail.com (Corresponding Author)**
[§] *E-mail:* **sarora131@gmail.com**

strategies have been proposed, including proportional allocation, Neyman allocation etc. The selection of an allocation strategy is influenced by various factors, such as objectives of research, population characteristics, and availability of resources. Proportional allocation is often preferred when strata vary in their sizes [2], [16], as it ensures adequate representation of smaller strata such that

$$n_i = np_i,$$

where n_i is the sample size drawn from i^{th} stratum of size N_i, n is the total sample size drawn from total population N and p_i is the proportion of strata size $\frac{N_i}{N}$, $i = 1, 2, \dots, k$.

Neyman allocation is theoretically optimal but may be impractical as it needs accurate auxiliary information [12] and is given by

$$n_i = \frac{np_i\sigma_{iz}}{\sum_{i=1}^{k} p_i\sigma_{iz}},$$

where σ_{iz} is the standard deviation between $z's$ in the sample for the i^{th} stratum. Sukhatme [14] proposed a method when variances for i^{th} stratum is not known then estimated variances are used based on equal allocations and it is called modified Neyman allocation. Evans [3] observed when σ_{iz}^2 are significantly equal then proportional allocations are more efficient then modified Neyman Allocation. The preliminary test of significance had been carried out by Sukhatme and Tang [15] and if the variances are significantly equal then the proportional allocation is recommended to be used otherwise modified Neyman allocation gives more efficient results.

Probability Proportional to Size (PPS) sampling has been suggested by Hansen and Hurvitz [9] for finite populations. PPS sampling is used when the unit size is proportional to the study variable, ensuring that each unit's selection probability is directly proportional to its size. In literature, to draw a sample with PPS without replacement Rao Hartley Cochran (RHC) design has been introduced by Rao, Hartley, and Cochran, [13]. In this design, the population is divided into number of groups equal to sample size using SRSWOR and from each group a unit from each group is selected using PPSWR sampling to draw a sample of PPSWOR sample of certain size.

Avadhani and Sukhatme [1] extended the approach to Controlled sampling in case of varying probability of selection of units within each stratum. Gupt and Rao [8] examined optimal allocation for distributing the sample size among strata in Stratified Probability Proportional to Size with Replacement (Stratified PPSWR) sampling within each stratum.

Zipf introduced a distribution where the most frequent unit appears approximately twice as often as the second most frequent, three times as often as the third, and so on [17]. This phenomenon, known as Zipf's law, states that a unit's rank is inversely related to its frequency. It was initially identified through the analysis of word repetition patterns in the English language. Further, Zipf's law is applied on cities and their population ranked in decreasing order of the size of cities [4].

In 2018, Joshi and Rajarshi suggested a new sampling technique named Modified Probability Proportional to Size (MPPS) sampling to draw a sample when the data follow Zipf's law [11]. They also suggested Modified Rao Hartley Cochran (MRHC) sampling in case of without replacement selection of a sample. In this approach, the population is divided into two groups after arranging the sampling units in descending order based on their frequency, ensuring no loss of generality. The first group, containing the highest frequencies, is entirely included in the sample, while the remaining units are selected from the second group using PPSWR for the MPPSWR design and RHC sampling for the MRHC design. The MPPSWR sampling technique is also developed in case to find the proportion and an application to find the proportion of PA (Persian-Arabic) loanwords by Joshi and Rajarshi [10]. Further, this sampling technique is applied to estimate the proportion of Indeclinable words in Hitopadesha [7].

In 2024, Goyal, Arora and Goyal introduced Stratified Modified Rao Hartley Cochran (Stratified MRHC) sampling for without-replacement case and Stratified Modified Probability Proportional to Size (Stratified MPPSWR) sampling for with-replacement scenarios [6]. In this approach, a heterogeneous population is split into homogeneous strata, with samples from each stratum selected using MPPSWR for with-replacement sampling and MRHC for without-replacement sampling, as outlined by Joshi and Rajarshi [11].

In this paper, we consider the problem of proportional and Neyman allocation for Stratified MPPS sampling and compare the allocation of sample sizes among k strata for the sampling technique. The efficiency of these allocations is also studied along with a real-life illustration.

The structure of this paper is as follows; in section 2, allocation of sample size among k strata using proportional and Neyman allocation for Stratified MPPSWR sampling is done. The comparison of these allocations has been given under certain conditions. In section 3, proportional and Neyman allocations are used for Stratified MRHC design followed by the comparison of the variances under these allocations. In section 4, these allocations are applied on real life situations and the results comes out to be gain due to Neyman allocation over proportional allocation. In section 5, the concluding remarks of comparison of Neyman and proportional allocation are given.

2. Allocations in Stratified Modified PPS with Replacement Design

Suppose a finite heterogenous population of total size M, consisting of N distinct sampling units. These units are divided into k homogeneous strata, each containing N_i units, with the total size for i^{th} stratum denoted as M_i, where $i = 1, 2, \ldots, k$ such that $M = M_1 + M_2 + \cdots + M_k$. A sample of size n_i is selected from the i^{th} stratum, which has a population size of N_i, resulting in a total sample of size n. Let $y_{i1}, y_{i2}, \ldots, y_{iN_i}$ represent the characteristic values of N_i units in the i^{th} stratum and let M_{ij} denote the size of j^{th} unit in the i^{th} stratum, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, N_i$. Then, the known probability of selection of j^{th} unit from i^{th} stratum is $f_{ij} = \frac{M_{ij}}{M_i}$ according to PPSWR sampling such that $\sum_{j=1}^{N_i} f_{ij} = 1 \; \forall i$.

We wish to draw a sample using Stratified MPPSWR sampling where from each stratum sample is drawn using MPPSWR sampling as the data within each stratum is assumed to follow Zipf's law [6].

In the i^{th} stratum, a sample of size n_i is split into two groups after N_i units are arranged in decreasing order of their size such that $M_{i1} > M_{i2} > \cdots > M_{iN_i}$, without loss of generality. Then, the population within each stratum be split into two groups, T_{i1} and T_{i2} with sizes N_{i1} and N_{i2} respectively.

(i) The first group, T_{i1} which consists of higher frequencies and has a size of $n_{i1} = N_{i1}(< N)$, is fully included in the sample with probability 1. Here, n_{i1} represents the sample size chosen from the first part of i^{th} stratum.

(ii) The second part of the sample, n_{i2} is selected from the remaining population T_{i2}, which consists of $N_{i2} = N_i - N_{i1}$ units, using PPSWR sampling. Consequently, the overall sample size for the i^{th} stratum is given by $n_i = n_{i1} + n_{i2}$. This procedure ultimately results in a Stratified MPPSWR sample of size n.

Suppose that,

$$q_{i1} = \sum_{j=1}^{N_{i1}} f_{ij}\, ; \;\; q_{i2} = \sum_{j=N_{i1}+1}^{N_i} f_{ij}\, ; \;\; Y_{i1} = \sum_{j=1}^{N_{i1}} y_{ij}\, ; \;\; Y_{i2} = \sum_{j=N_{i1}+1}^{N_i} y_{ij} \quad (1)$$

then the selection probability of n_{i2} units from the second group T_{i2}, which includes the j^{th} unit (where $j = N_{i1+1} + N_{i1+2} + \cdots + N_i$) in the i^{th} stratum, is given by $\frac{f_{ij}}{q_{i2}}$ using PPSWR sampling. Let,

$$z_{ij} = \frac{y_{ij}}{f_{ij}/q_{i2}}.$$

For the population total Y, the unbiased estimator under Stratified MPPSWR sampling is expressed as

$$\hat{y}_{smppf} = \sum_{i=1}^{k} \frac{N_i}{N}\left(\sum_{j=1}^{n_{i1}} y_{ij} + \frac{q_{i2}}{n_{i2}} \sum_{j=1, j\epsilon T_{i2}}^{n_{i2}} \frac{y_{ij}}{f_{ij}}\right), \quad (2)$$

where $\hat{y}_{ismppf2} = \frac{q_{i2}}{n_{i2}} \sum_{j=1, j\epsilon T_{i2}}^{n_{i2}} \frac{y_{ij}}{f_{ij}}$ such that $E(\hat{y}_{ismppf2}) = Y_{i2}$.

The first group from each stratum is selected with probability 1, resulting in no variation, while the variance arises solely from the second group of the sample in each stratum and is given by

$$V(\hat{y}_{smppf}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2}\left[\frac{1}{n_{i2}} \sum_{j=N_{i1}+1}^{N_i} \frac{f_{ij}}{q_{i2}} \left(\frac{y_{ij}}{f_{ij}/q_{i2}} - Y_{i2}\right)^2\right]. \quad (3)$$

The unbiased estimator of the variance of Stratified MPPSWR sampling is given by

$$\hat{v}(\hat{y}_{smppf}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2}\left[\frac{1}{n_{i2}(n_{i2}-1)} \sum_{j=1, j\epsilon T_{i2}}^{n_{i2}} \left(\frac{y_{ij}}{f_{ij}/q_{i2}} - \hat{y}_{ismppf2}\right)^2\right]. \quad (4)$$

Comparison of Stratified MPPSWR sampling and Stratified PPSWR sampling establishes the condition for determining the optimum value of n_{i1} [6]. The value is obtained under the condition

$$V(\hat{y}_{sppf}) - V(\hat{y}_{smppf}) \geq 0, if \left(\frac{1}{n_i} - \frac{q_{i2}}{n_{i2}}\right) \geq 0. \tag{5}$$

2.1 *Allocation of Sample Size among Various Strata*

The variance of Stratified MPPSWR sampling depends on the sample size drawn from each stratum, specifically sample size distribution among different strata. In this regard, we will explore proportional and Neyman allocation in the context of Stratified MPPSWR sampling. In Stratified MPPSWR sampling, variance can be written as

$$V(\hat{y}_{smppf}) = \sum_{i=1}^{k} p_i^2 \left[\frac{\sigma_{iz}^2}{n_{i2}}\right], \tag{6}$$

where $p_i = \frac{N_i}{N}$ and $\sigma_{iz}^2 = \sum_{j=N_{i1}+1}^{N_i} \frac{f_{ij}}{q_{i2}} \left(\frac{y_{ij}}{f_{ij}/q_{i2}} - Y_{i2}\right)^2$.

The gain in efficiency due to Stratified MPPSWR sampling is minimum when the factor is equals to zero as given in equation (5) i.e.,

$$\left(\frac{1}{n_i} - \frac{q_{i2}}{n_{i2}}\right) = 0 \quad \Rightarrow n_{i2} = n_i q_{i2} \tag{7}$$

(i) Proportional Allocation

In proportional allocation, n_i is directly proportional to N_i which implies $n_i = KN_i$, where K is the proportionality constant.

$$\Rightarrow n_i = KN_i$$

$$\Rightarrow \sum_{i=1}^{k} n_i = K \sum_{i=1}^{k} N_i$$

$$\Rightarrow n_i = \frac{nN_i}{N} = np_i$$

Putting the value of n_i in equation (7) we get $n_{i2} = np_i q_{i2}$.

Then, the variance under Stratified MPPSWR sampling using proportional allocation by putting the value of n_{i2} in equation (6) is given by

$$V(\hat{y}_{smppf})_P = \sum_{i=1}^{k} p_i^2 \left[\frac{1}{q_{i2} n p_i} \sigma_{iz}^2\right]$$

$$= \frac{1}{n} \sum_{i=1}^{k} \frac{p_i \sigma_{iz}^2}{q_{i2}}. \tag{8}$$

(ii) Neyman Allocation

In Neyman allocation, n_i is directly proportional to $p_i \sigma_{iz}$ which implies $n_i = Kp_i \sigma_{iz}$, where K is the proportionality constant.

$$\Rightarrow n_i = Kp_i \sigma_{iz}$$

$$\Rightarrow \sum_{i=1}^{k} n_i = K \sum_{i=1}^{k} p_i \sigma_{iz}$$

$$\Rightarrow n_i = \frac{np_i \sigma_{iz}}{\sum_{i=1}^{k} p_i \sigma_{iz}}$$

Putting the value of n_i in equation (7) we get $n_{i2} = \frac{np_i\sigma_{iz}q_{i2}}{\sum_{i=1}^{k} p_i\sigma_{iz}}$.

Then, the variance under Stratified MPPSWR sampling using Neyman allocation is given by

$$V(\hat{y}_{smppf})_N = \sum_{i=1}^{k} p_i^2 \left[\frac{\sum_{i=1}^{k} p_i\sigma_{iz}}{q_{i2}np_i\sigma_{iz}}\sigma_{iz}^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}\left[\sum_{i=1}^{k} p_i\sigma_{iz}\right]$$
$$= \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}\sigma_w, \tag{9}$$

where $\sigma_w = \sum_{i=1}^{k} p_i\sigma_{iz}$.

2.2 *Comparison of Stratified MPPSWR sampling under Proportional and Neyman Allocation*

For fixed n, the variance under proportional allocation can be written using equation (8) and (9) as

$$V(\hat{y}_{smppf})_P = \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}\sigma_w + \left[\frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}^2}{q_{i2}} - \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}\sigma_w\right]$$
$$= V(\hat{y}_{smppf})_N + \left[\frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}(\sigma_{iz} - \sigma_w)\right], \tag{10}$$

where $\sigma_w = \sum_{i=1}^{k} p_i\sigma_{iz}$.

For Neyman allocation to be better than proportional allocation the factor $(\sigma_{iz} - \sigma_w)$ must be greater than or equal to 0. But we observe in practical situations if for one stratum it is positive then for the other it will be negative. But overall, the factor $\left[\sum_{i=1}^{k}\frac{p_i\sigma_{iz}}{q_{i2}}(\sigma_{iz} - \sigma_w)\right]$ comes out to be positive as it is the weighted sum of the factor $(\sigma_{iz} - \sigma_w)$.

It can be visualised that if the standard deviation is same for all strata, then the gain due to Neyman allocation is reduced in comparison to the gain due to proportional allocation under the condition due to modification given in equation (5) when the gain in variance is least. If all σ_{iz} are equal i.e., $\sigma_{iz} = \sigma$, then two systems of allocation become equally efficient.

$$V(\hat{y}_{smppf})_P - V(\hat{y}_{smppf})_N = \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma}{q_{i2}}\left(\sigma - \sum_{i=1}^{k} p_i\sigma\right)$$
$$= \frac{1}{n}\sum_{i=1}^{k}\frac{p_i\sigma}{q_{i2}}\left(\sigma - \sigma\sum_{i=1}^{k} p_i\right) = 0$$

as we know $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k}\frac{N_i}{N} = 1$.

$$V(\hat{y}_{smppf})_P = V(\hat{y}_{smppf})_N$$

Hence, the proportional allocation comes out to be equal to Neyman allocation when the strata variances are equal.

3. Allocations in Stratified MPPS without Replacement Sampling Design

Goyal et al. introduced Stratified MRHC sampling for heterogeneous data which follow Zipf's law within each stratum [6]. The population is split into k homogeneous strata. Consequently, MRHC sampling is used to draw samples from each stratum in the case of without replacement sampling. Using MRHC design, the population total Y is estimated which is highly correlated to the size of the population units using a sample of size n_i selected from the population of size N_i from the i^{th} stratum.

In this sampling technique, i^{th} stratum is separated into two groups after sorting the units in descending order based on their size. Each stratum is split into two groups, T_{i1} and T_{i2}. $N_{i1} = n_{i1}$ units with higher frequency are fully included in the sample with probability 1 from the first group i.e., T_{i1}. From the population size $N_{i2} = N - N_{i1}$ i.e., from the second group T_{i2}, remaining sample of size $n_{i2} = n - n_{i1}$ is selected using RHC design of PPSWOR sampling. The population N_{i2} in i^{th} stratum is split into n_{i2} SRSWOR random groups and from each random group a unit is selected with RHC design. The j^{th} unit in the i^{th} stratum from the second group has the selection probability $\frac{f_{ij}}{\tau_{is,j}}$, where $\tau_{is,j}$ represents the total of f_{ij}'s i.e., the initial probabilities of the s^{th} random group in the i^{th} stratum $(i = 1, 2, \dots, k, j = 1, 2, \dots, N_i)$. Then, under this scheme, the unbiased estimator of population total is

$$\hat{Y}_{sMRHC} = \sum_{i=1}^{k} \frac{N_i}{N} \left[\sum_{j=1}^{n_{i1}} y_{ij} + \sum_{j=1, j\epsilon T_{i2}}^{n_{i2}} \frac{y_{ij}}{f_{ij}/\tau_{is,j}} \right] \tag{11}$$

The variance of the estimator when the n_{i2} groups formed of equal size such that $N_{i2} = n_{i2} Z_{i2}$, where Z_{i2} is a positive integer is given by

$$V(\hat{Y}_{sMRHC}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2} \left(1 - \frac{n_{i2}-1}{N_{i2}-1}\right) \frac{1}{n_{i2}} \sum_{j=N_{i1}+1}^{N_i} \frac{f_{ij}}{q_{i2}} \left(\frac{y_{ij}}{f_{ij}/q_{i2}} - Y_{i2} \right)^2 \tag{12}$$

The comparison of Stratified MRHC sampling with Stratified RHC sampling gives the condition under which the optimum value of n_{i1} is determined [6]. The value is obtained under the condition $V(\hat{Y}_{sMRHC}) - V(\hat{Y}_{sRHC}) \geq 0$, if

$$\left[\left(\frac{N_i - n_i}{N_i - 1} \right) \frac{1}{n_i} - \left(1 - \frac{n_{i2}-1}{N_{i2}-1}\right) \frac{q_{i2}}{n_{i2}} \right] \geq 0 \tag{13}$$

3.1 *Allocation of Sample Size among Various Strata in case of Stratified MRHC sampling*

The variance of the Stratified MRHC sampling is influenced by the sample size allocated to each stratum, determining how the total sample is distributed among the various strata. We will discuss proportional allocation and Neyman allocation in the case of Stratified MRHC sampling. The variance of Stratified MRHC sampling when $N_{i2} = n_{i2} Z_{i2}$, where Z_{i2} is a positive integer given in equation (12) can be written as

$$V(\hat{Y}_{sMRHC}) = \sum_{i=1}^{k} p_i^2 \left(1 - \frac{n_{i2}-1}{N_{i2}-1}\right) \frac{1}{n_{i2}} \sigma_{iz}^2$$

$$= \sum_{i=1}^{k} p_i^2 \left(\frac{N_{i2}-n_{i2}}{N_{i2}-1}\right) \frac{\sigma_{iz}^2}{n_{i2}}, \tag{14}$$

where $p_i = \frac{N_i}{N}$ and $\sigma_{iz}^2 = \sum_{j=N_{i1}+1}^{N_i} \frac{f_{ij}}{q_{i2}} \left(\frac{y_{ij}}{f_{ij}/q_{i2}} - Y_{i2}\right)^2$.

Using the condition of Stratified MRHC sampling given in equation (13), the gain in efficiency is minimum when this factor is equals to zero i.e.,

$$\left[\left(\frac{N_i-n_i}{N_i-1}\right)\frac{1}{n_i} - \left(1 - \frac{n_{i2}-1}{N_{i2}-1}\right)\frac{q_{i2}}{n_{i2}}\right] = 0$$

$$n_{i2} = q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right) n_i \tag{15}$$

(i) Proportional Allocation

For proportional allocation $n_i = np_i$. So, putting the value in equation (15) we get

$$n_{i2} = q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right) np_i$$

Then, the variance under Stratified MRHC design using proportional allocation is given by

$$V\left(\hat{Y}_{sMRHC}\right)_P = \sum_{i=1}^{k} p_i^2 \left(\frac{N_{i2}-q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right)np_i}{N_{i2}-1}\right) \frac{1}{q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right)np_i} \sigma_{iz}^2$$

$$= \sum_{i=1}^{k} p_i^2 \left(\frac{N_{i2}(N_{i2}-1)-q_{i2}(N_i-1)np_i}{(N_{i2}-1)^2}\right) \frac{(N_{i2}-1)}{q_{i2}(N_i-1)np_i} \sigma_{iz}^2$$

$$= \sum_{i=1}^{k} \frac{p_i\sigma_{iz}^2}{nq_{i2}(N_i-1)} \left(\frac{N_{i2}(N_{i2}-1)-q_{i2}(N_i-1)np_i}{(N_{i2}-1)}\right) \tag{16}$$

(ii) Neyman Allocation

For Neyman allocation, $n_i = \frac{np_i\sigma_{iz}}{\sum_{i=1}^{N} p_i\sigma_{iz}}$, So, putting the value in equation (15) we get

$$n_{i2} = q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right)\frac{np_i\sigma_{iz}}{\sigma_w},$$

where $\sigma_w = \sum_{i=1}^{N} p_i\sigma_{iz}$. Then, the variance under Stratified MRHC design using Neyman allocation is given by

$$V\left(\hat{Y}_{sMRHC}\right)_N = \sum_{i=1}^{k} \frac{p_i^2\sigma_{iz}^2}{np_i\sigma_{iz}} \sigma_w \left(\frac{N_{i2}-q_{i2}\left(\frac{N_i-1}{N_{i2}-1}\right)\frac{np_i\sigma_{iz}}{\sigma_w}}{N_{i2}-1}\right) \frac{N_{i2}-1}{q_{i2}(N_i-1)}$$

$$= \sum_{i=1}^{k} \frac{p_i^2\sigma_{iz}^2}{np_i\sigma_{iz}} \sigma_w \left(\frac{N_{i2}(N_{i2}-1)\sigma_w-q_{i2}(N_i-1)np_i\sigma_{iz}}{(N_{i2}-1)^2\sigma_w}\right) \frac{N_{i2}-1}{q_{i2}(N_i-1)}$$

$$= \sum_{i=1}^{k} \frac{p_i\sigma_{iz}}{nq_{i2}(N_i-1)} \left(\frac{N_{i2}(N_{i2}-1)\sigma_w-q_{i2}(N_i-1)np_i\sigma_{iz}}{(N_{i2}-1)}\right), \tag{17}$$

where $\sigma_w = \sum_{i=1}^{k} p_i\sigma_{iz}$.

3.2 *Comparison of Stratified MRHC Sampling under Proportional and Neyman Allocation*

For fixed n, the difference of variances under proportional and Neyman allocation can be written using equation (16) and (17) as

$$V(\hat{y}_{sMRHC})_P - V(\hat{y}_{sMRHC})_N$$

$$= \sum_{i=1}^{k} \frac{p_i\sigma_{iz}^2}{nq_{i2}(N_i-1)}\left(\frac{N_{i2}(N_{i2}-1)-q_{i2}(N_i-1)np_i}{(N_{i2}-1)}\right) - \sum_{i=1}^{k} \frac{p_i\sigma_{iz}}{nq_{i2}(N_i-1)}\left(\frac{N_{i2}(N_{i2}-1)\sigma_w-q_{i2}(N_i-1)np_i\sigma_{iz}}{(N_{i2}-1)}\right)$$

$$= \sum_{i=1}^{k} \frac{p_i\sigma_{iz}}{nq_{i2}(N_i-1)}\left(\frac{N_{i2}(N_{i2}-1)(\sigma_{iz}-\sigma_w)}{(N_{i2}-1)}\right)$$

$$= \sum_{i=1}^{k} \frac{p_i\sigma_{iz}}{nq_{i2}(N_i-1)} N_{i2}(\sigma_{iz}-\sigma_w) \tag{18}$$

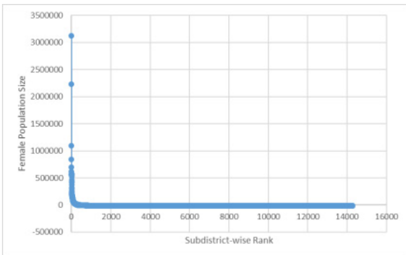
For Neyman allocation to be better than proportional allocation the factor $(\sigma_{iz} - \sigma_w) \geq 0$. But we observe in practical situations if for one stratum it is positive then for the other it will be negative. But overall, the factor $\sum_{i=1}^{k} \frac{p_i\sigma_{iz}}{nq_{i2}(N_i-1)} N_{i2}(\sigma_{iz}-\sigma_w)$ comes out to be positive as it is the weighted sum of the factor $(\sigma_{iz} - \sigma_w)$ in equation (18).

Remark: If variances of all strata i.e., σ_{iz} are equal then proportional and Neyman allocation are equally efficient as in case of Stratified MPPS sampling.

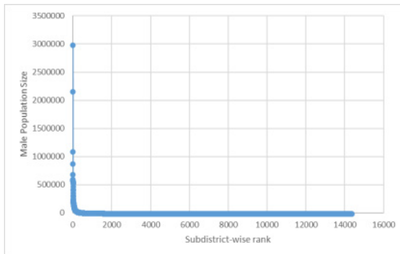
4. Application of Stratified MPPS sampling under Proportional and Neyman Allocation for Estimating Christian Literate Population in India

A real-life situation has been carried out in order to observe the behaviour of proportional and Neyman allocation in case of Stratified MPPS sampling. We have considered data from census 2011 of literate population [18]. We aim to estimate the total literate population in India and a subdistrict-wise Christian population of size $N = 28625$ is considered for analysis. The complete population of subdistricts is split into two strata i.e., female with size $N_1 = 14271$ and male with size $N_2 = 14353$ respectively. A sample approximately 10% of the population is selected from the sampling units i.e., $n = 2862$. The two strata follow Zipf's law as the subdistrict-wise population as shown in Figure 1. In the case of with and without-replacement sampling, under Stratified MPPSWR and Stratified MRHC design respectively, the sample size is split into two portions using proportional and Neyman allocation methods.

Using R software, under proportional allocation the sample of size $n_1 = 1427$ and $n_2 = 1435$ is selected from the female and male strata. In Stratified MPPSWR sampling, the optimal values are $n_{11} = 569$ for the female stratum and $n_{21} = 570$ for male stratum, determined using the condition in equation (5) and presented in Table 1. After sorting them in descending order based on the subdistrict population size for female and male strata, the population is divided into two parts. The first part is fully included in the sample consisting high frequency units. The remaining population for female stratum of size $n_{12} = 858$ and for male stratum is $n_{22} = 865$ is selected from the remaining population using PPSWR sampling.



(a) Female Population



(b) Male Population

Figure 1
Graph of Female and Male Population Size According to Census 2011

Similarly for Stratified MRHC design, the optimal value $n_{11} = 512$ and $n_{21} = 512$ for female and male strata respectively as shown in Table 2 using condition given in equation (13). The remaining units are included in the sample using RHC design from both strata. Using RHC design, the remaining units $n_{21} = 915$ and $n_{22} = 923$ are included in the sample.

According to Neyman allocation, the sample size of $n_1 = 1640$ from female and $n_2 = 1222$ from male stratum is selected using the population variances. For Stratified MPPSWR sampling, the optimal values are $n_{11} = 697$ and $n_{21} = 442$ for female and male strata respectively as given in Table 1 using equation (5). These sampling units are selected in the sample with probability 1, while the remaining units $n_{12} = 943$ and $n_{22} = 780$ are chosen using PPSWR sampling.

In case of Stratified MRHC design the optimal values $n_{11} = 668$ and $n_{21} = 441$ for female and male strata respectively using condition given in equation (13) given in Table 2. The remaining units $n_{21} = 972$ and $n_{22} = 781$ are included in the sample using RHC design from both strata.

Table 1
Optimal Values of n_{i1} under Stratified MPPSWR Sampling

		f_{ij}	n_{i1}	q_{i2}	$\left(\frac{1}{n_i} - \frac{q_{i2}}{n_{i2}}\right)$
Proportional Allocation	**Female**	0.00023962	568	0.1966326	0.000471862
		0.00023906	**569**	**0.1963930**	**0.000471874**
		0.00022425	570	0.1961539	0.000471869
	Male	0.00023005	569	0.1979912	0.000468237
		0.00022963	**570**	**0.1977612**	**0.000468238**
		0.00022854	571	0.1975315	0.000468238

Neyman Allocation	Female	0.00019783	696	0.1706381	0.000428995
		0.00019437	**697**	**0.1704403**	**0.000429010**
		0.00017219	698	0.1702493	0.000429001
	Male	0.00030211	441	0.2310590	0.000522481
		0.00030030	**442**	**0.2307565**	**0.000522487**
		0.00029431	443	0.2304580	0.000522485

Table 2
Optimal values of n_{i1} under Stratified MRHC sampling

		f_{ij}	n_{i1}	q_{i2}	N_2	$\left(1-\frac{n_{i2}-1}{N_{i2}-1}\right)\frac{q_{i2}}{n_{i2}}$
Proportional Allocation	Female	0.000270099	511	0.2105084	13760	0.000214551
		0.000269804	**512**	**0.2102383**	**13759**	**0.000214541**
		0.000243680	513	0.2099685	13758	0.000214556
	Male	0.000253551	511	0.2119638	13843	0.000214120
		0.000251203	**512**	**0.2117103**	**13842**	**0.000214098**
		0.000250889	513	0.2114591	13841	0.000214101
Neyman Allocation	Female	0.000191766	667	0.1760567	13604	0.000168030
		0.000191676	**668**	**0.1758649**	**13603**	**0.000168020**
		0.000191615	669	0.1756733	13602	0.000168024
	Male	0.000304332	440	0.2313607	13914	0.000279403
		0.000302110	**441**	**0.2310564**	**13913**	**0.000279403**
		0.000300308	442	0.2307543	13912	0.000279405

The variance using proportional allocation is 157960634.2 and using Neyman allocation is 157268319.4 in case of Stratified MPPSWR sampling. This can easily be seen that Neyman allocation gives lesser variance than proportional allocation. For Stratified MRHC design, the variance is 151872617.2 for proportional allocation and 150779968.9 for Neyman allocation. The Neyman allocation is efficient in case of without replacement also.

We calculated the efficiency by taking the ratio of variance of proportional allocation by the variance of Neyman allocation. In Stratified MPPSWR sampling, the efficiency is 1.004402125 and for Stratified MRHC design, the efficiency is 1.007246641 as given in Table 3. From this we can observe that Neyman allocation gives more efficient results than proportional allocation in both the cases of with and without replacement.

Table 3
Comparison of Stratified MPPS Sampling in Case of Proportional and Neyman Allocation

			n_{i1}	Variance	Efficiency
Stratified MPPSWR	**Proportional Allocation**	**Female**	569	406439440	**1.0044021251**
		Male	570	221613979	
		Total		**157960634.2**	
	Neyman Allocation	**Female**	697	308900376	
		Male	442	320102650	
		Total		**157268319.4**	
Stratified MRHC	**Proportional Allocation**	**Female**	512	386829480	**1.0072466411**
		Male	512	226441606	
		Total		**151872617.2**	
	Neyman Allocation	**Female**	668	293592744	
		Male	441	309430218	
		Total		**150779968.9**	

5. Conclusion

In this paper, variances of Stratified Modified Probability Proportional to Size sampling has been discussed in case of proportional allocation and Neyman allocation in case of with and without replacement design. The theoretical results of comparison of the two allocation procedures are given and under certain condition the proportional allocation gives less efficient results than Neyman allocation. In this sampling design, the sample size is distributed across different strata where proportional allocation proves to be less efficient in practical scenarios for both with-replacement and without-replacement cases.

Acknowledgement

The authors express their gratitude to the DST for providing funding through the INSPIRE fellowship program which supported the completion of this research.

References

[1] M.S. Avadhani and B.V. Sukhatme, "Controlled sampling with varying probabilities with and without replacement," *Australian Journal of Statistics*, vol. 9, pp. 8-15 (1967).

[2] W.G. Cochran, Sampling theory. New York: Wiley (1977).

[3] W.D. Evans, "On stratification and optimum allocations". *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 95–104 (1951), https://doi.org/10.1080/01621459.1951.10500772

[4] X. Gabaix, "Zipf's law and the growth of cities," *The American Economic Review*, vol. 89, no. 2, pp. 129–132 (1999), http://www.jstor.org/stable/117093.

[5] R. Goodman and L. Kish, "Controlled selection- a technique in probability sampling," *Journal of the American Statistical Association*, vol. 45, no. 251, pp. 350-372 (1950), https://doi.org/10.1080/01621459.1950.10501130.

[6] A. Goyal, S. Arora and A. Goyal, "A stratified modified probability proportional to size sampling technique," *Communications in Statistics - Theory and Methods*, vol. 53, no. 23, pp. 8525–8542 (2024). https://doi.org/10.1080/03610926.2023.2292969

[7] A. Goyal, P. Choraria, V. Dwivedi and P.C. Gupta, "Application of Modified PPSWR to estimate the actual proportion of indeclinable words of hitopadesha," *Aligarh Journal of Statistics*, vol. 43, pp. 139-148 (2023).

[8] B.K. Gupt and T.J. Rao, "Stratified PPS sampling and allocation of sample size," *Journal of the Indian Society of Agricultural Statistics*, vol. 50, no. 2, pp. 199-208 (1997).

[9] M.H. Hansen and W.N. Hurwitz, "On the theory of sampling from finite populations," *The Annals of Mathematical Statistics,* vol. 14, no. 4, pp. 333–362 (1943).

[10] K. Joshi and M.B. Rajarshi, "Estimation of actual proportion of loanwords in a language," *Sankhya-B*, vol. 79, no. I, pp. 60-75 (2017), https://doi.org/10.1007/s13571-016-0127-5.

[11] K. Joshi and M.B. Rajarshi, "Modified probability proportional to size sampling," *Communications in Statistics - Theory and Methods*, vol. 47, no. 4, pp. 805–815 (2018), https://doi.org/10.1080/03610926.2016.1139131

[12] J. Neyman, "On two different aspects of the representative method, the method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society,* vol. 97, no. 4, pp. 558-625 (1934), https://doi.org/10.1007/978-1-4612-4380-9_12

[13] J. N. Rao, H. O. Hartley, and W. G. Cochran, "On a simple procedure of unequal probability sampling without replacement," *Journal of the*

Royal Statistical Society: Series B (Methodological), vol. 24, no. 2, pp. 482-491 (1962), https://doi.org/10.1111/j.2517-6161.1962.tb00475.x

[14] P.V. Sukhatme, "Contribution to the theory of the representative method," *Supplement to the Journal of the Royal Statistical Society*, vol. 2, no. 2, pp. 253–268 (1935), https://doi.org/10.2307/2983640.

[15] B.V. Sukhatme and V.K.T. Tang, "Allocation in stratified sampling subsequent to preliminary test of significance," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 175–179 (1975). https://doi.org/10.1080/01621459.1975.10480282

[16] P.V. Sukhatme, B.V. Sukhatme and C. Asok, Theory of sample surveys with applications. New Delhi: Indian Society of Agriculture Statistics (1984).

[17] G.K. Zipf, The psycho-biology of language. Houghton, Mifflin (1935).

[18] Census tables. Government of India (censusindia.gov.in).

Received May, 2024