CHAPTER 5

Unequal Probability Sampling

5.1 INTRODUCTION

In an unequal (or varying) probability sampling (VPS) design, all possible samples need not have the same selection probabilities. In this design, samples are generally selected by using an auxiliary variable, x, which is also known as a measure of size variable or simply as a measure of size. The values of the auxiliary variable are assumed to be positive and known before the survey. It is found that if the auxiliary variable x is well related to the study variable y, then one can use auxiliary information in an appropriate manner for selection of a sample to get an efficient estimator for the population parameter of interest. For instance, to estimate the total number of inpatients treated in hospitals, one may use the number of beds in a hospital as measure of size variable. The hospitals that have a large number of beds are expected to have treated a higher number of inpatients than those with a fewer number of beds. Hence the hospitals having a large number of beds should be given a higher chance of selection in a sample than those with fewer beds to ensure an efficient estimator for the total number of inpatients. In estimating the total production of a crop, the farm size may be used as an auxiliary variable because the production of a crop in a farm is expected to be proportional to the size of the farm. In this chapter we consider various types of VPS schemes such as probability proportional to size with and without replacement sampling schemes, inclusion probability proportional to size sampling schemes, Lahiri—Midzuno—Sen (LMS), and Rao—Hartley—Cochran (RHC) sampling schemes. Methods of unbiased estimation of population characteristics, expressions of the variances of the proposed estimators, and unbiased estimators of the variances for various sampling schemes are also studied in detail.

We will denote x_i as the value of the auxiliary variable x for the ith unit, which is assumed to be known and strictly positive for every $i \in U$, and $X = \sum_{i=1}^{N} x_i$ be their total. The normed size measure for the ith unit is defined by $p_i = x_i/X(>0)$ and it is subject to $\sum_{i=1}^{N} p_i = 1$.

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00005-4

5.2 PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SAMPLING SCHEME

In the probability proportional to size with replacement (PPSWR) sampling scheme, units are selected independently in each draw so that the probability of selection of the ith unit in any draw is p_i. The probability of selection of an ordered sample $s_o = (i_1, \rightarrow, i_2, \rightarrow, \cdots, \rightarrow i_n) = (i_1, i_2, \cdots, i_n)$ where the unit i_r is selected at the rth draw, $r = 1, \ldots, n$ is $p(s_o) = p_{i_1} p_{i_2} \ldots p_{i_n}$. Here it is important to note that a unit may be selected more than once in a sample. We use cumulative total method and Lahiri's (1951) method for the selection of PPSWR samples. The methods are described as follows.

5.2.1 Cumulative Total Method

Here we first choose a constant k to make $kx_i = X_i$, an integer for each $i = 1, \ldots, N$. We associate random numbers $1 (= T_0)$ to $X_1 (= T_1)$ for the first unit. For the second unit, random numbers $X_1 + 1$ to $X_1 + X_2 (= T_2)$ and in general for the ith unit, random numbers $T_{i-1} + 1$ to T_i are associated where $T_i = X_1 + \cdots + X_i$ for $i = 1, \ldots, N$. For the selection of a sample, we select a number R at random from 1 to $T_n = kX$. The random number R selects the jth unit if $T_{j-1} < R \leq T_j$. Clearly, the probability of selection of the jth unit is $(T_j - T_{j-1})/T_n = X_j/T_n = p_j$. The procedure is repeated n times to select a sample of size n units.

Example 5.2.1

Select a sample of four households from a list of nine by the PPSWR method, using household size as an auxiliary variable (Table 5.2.1).

Here we choose $k = 10$, so that $X_i = 10x_i$ becomes an integer for $i = 1, \ldots, 9$. The values of X_i's and T_i's are given in Table 5.2.2.

From the random number table we select three-digit random numbers between 001 and 285 as given in Table 5.2.3.

Table 5.2.1

Serial number household (i)	1	2	3	4	5	6	7	8	9
Household size (x_i)	2	3.5	2	4	2.5	5	2	5	2.5

Table 5.2.2

i	1	2	3	4	5	6	7	8	9
X_i	20	35	20	40	25	50	20	50	25
T_i	20	55	75	115	140	190	210	260	285

Table 5.2.3

Random number selected	465	238	198	431	215	023
Unit selected	–	8	7	–	8	2

"–" indicates that the units are not selected because corresponding chosen random number exceeds 285.

Hence the selected ordered sample by PPSWR method is $s_o = (8, 7, 8, 2)$ where the unit 8 has been selected twice.

5.2.2 Lahiri's Method

The computation of cumulative totals becomes tedious if the population size N is large. Lahiri (1951) has proposed an alternative method where one need not compute cumulative totals. The method is described as follows.

Select two random numbers, one from 1 to N say "j" and another say "R" from 1 to M where $M \geq \underset{1 \leq i \leq N}{Max} \{x_i\}$. The unit "$j$" is selected in the sample if $R \leq x_j$. If $R > x_j$, no unit is selected and the procedure is repeated till a unit is selected. For selection of a sample size n, the method is repeated n times.

Theorem 5.2.1

Probability of selection of the ith unit for Lahiri's method is p_i for $i = 1, \ldots, N$.

Proof

Let a_i = Probability that the first pair of random numbers selects unit i

= Probability of selection of random number $i(1 \leq i \leq N)$

and a random number R less than or equal to $x_i = x_i/(NM)$

and

b = Probability that the first pair of random numbers fails to select any unit

$$= \sum_{i=1}^{N} \frac{1}{N} \frac{M - x_i}{M} = 1 - \frac{X}{NM}$$

Hence the probability of selection of ith unit using this method

= Prob (First pair of random numbers selects ith unit) + Prob (First pair of random numbers fail to select any unit) × Prob (Second pair of random number selects ith unit) + ⋯

$$= a_i + ba_i + b^2 a_i + \cdots = a_i(1 - b)^{-1} = x_i/X = p_i$$

Example 5.2.2

Select a sample of four households using Lahiri's method for the data given in Table 5.2.1. Here the population size is $N = 9$ and we may take $M = 6 \geq \underset{1 \leq i \leq 9}{Max}\{x_i\}$

Table 5.2.4

Random number between 1 and 9 (j)	4	5	3	2	5
Random number between 1 and 6 (R)	6	2	2	3	2
x_j	4	2.5	2	3.5	2.5
Unit selected	—	5	3	2	5

So, the selected ordered sample by Lahiri's method is $s_o = (5, 3, 2, 5)$.

5.2.3 Hansen—Hurwitz Estimator and its Variance

Let $y_{(r)}$ and $x_{(r)}$, respectively, be the value of the study (y) and auxiliary variable (x) of the unit that are selected at the rth draw, $r = 1,\ldots, n$ and $p_{(r)} = x_{(r)}/X$. Then $y_{(r)}/p_{(r)} = y_i/p_i$ if the rth draw produces the ith unit with probability p_i; $i = 1,\ldots, N$; $r = 1,\ldots, n$. The Hansen—Hurwitz (1943) estimator for the population total Y is defined as

$$\widehat{Y}_{hh} = \frac{1}{n}\sum_{r=1}^{n}\frac{y_{(r)}}{p_{(r)}} \tag{5.2.1}$$

Theorem 5.2.2

(i) $\widehat{Y}_{hh}$ is an unbiased estimator of the population total Y

(ii) Variance of $\widehat{Y}_{hh}$ is $V\left(\widehat{Y}_{hh}\right) = \dfrac{V_{pps}}{n}$

where

$$V_{pps} = \sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 = \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_i p_j\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

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

$$\widehat{V}\left(\widehat{Y}_{hh}\right) = \frac{1}{n(n-1)}\sum_{r=1}^{n}\left(\frac{y_{(r)}}{p_{(r)}} - \widehat{Y}_{pps}\right)^2$$

Proof

The expectation and variance of $\dfrac{y_{(r)}}{p_{(r)}}$ are given by

$$E\left(\frac{y_{(r)}}{p_{(r)}}\right) = \sum_{i=1}^{N} \frac{y_i}{p_i}\, p_i = Y \tag{5.2.2}$$

and

$$\begin{aligned} V\left(\frac{y_{(r)}}{p_{(r)}}\right) &= E\left(\frac{y_{(r)}}{p_{(r)}} - Y\right)^2 \\ &= \sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 \\ &= V_{pps}. \end{aligned} \tag{5.2.3}$$

Furthermore,

$$\begin{aligned} \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_ip_j\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 &= \sum_{i\neq}^{N}\sum_{j=1}^{N} p_j\,\frac{y_i^2}{p_i} - \sum_{i\neq}^{N}\sum_{j=1}^{N} y_iy_j \\ &= \sum_{i=1}^{N}\frac{y_i^2}{p_i}(1-p_i) - \left(Y^2 - \sum_{i=1}^{N} y_i^2\right) \\ &= V_{pps}. \end{aligned} \tag{5.2.4}$$

Now using Eqs. (5.2.2–5.2.4), we get

(i) $E\left(\widehat{Y}_{hh}\right) = \dfrac{1}{n}\sum_{r=1}^{n} E\left(\dfrac{y_{(r)}}{p_{(r)}}\right) = Y$

(ii) $$\begin{aligned} V\left(\widehat{Y}_{hh}\right) &= \frac{1}{n^2}\,V\left(\sum_{r=1}^{n}\frac{y_{(r)}}{p_{(r)}}\right) \\ &= \frac{1}{n^2}\left[\sum_{n=1}^{n} V\left(\frac{y_{(r)}}{p_{(r)}}\right) + \sum_{r\neq}^{n}\sum_{t=1}^{n} Cov\left(\frac{y_{(r)}}{p_{(r)}}, \frac{y_{(t)}}{p_{(t)}}\right)\right] \\ &= \frac{V_{pps}}{n} \end{aligned}$$

(because $V\left(\frac{y_{(r)}}{p_{(r)}}\right) = V_{pps}$ and $Cov\left(\frac{y_{(r)}}{p_{(r)}}, \frac{y_{(t)}}{p_{(t)}}\right) = 0$ for $r \neq t$ as the draws are independent)

and

(iii) $$E\left[\widehat{V}\left(\widehat{Y}_{hh}\right)\right] = \frac{1}{n(n-1)} E\left[\sum_{r=1}^{n}\left(\frac{y_{(r)}}{p_{(r)}}\right)^2 - n\left(\widehat{Y}_{pps}\right)^2\right]$$

$$= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} E\left(\frac{y_{(r)}}{p_{(r)}}\right)^2 - nE\left(\widehat{Y}_{pps}\right)^2\right]$$

$$= \frac{1}{n(n-1)}\left[n\left(V_{pps} + Y^2\right) - n\left(V\left(\widehat{Y}_{pps}\right) + Y^2\right)\right]$$

$$= \frac{V_{pps}}{n}$$

$$= V\left(\widehat{Y}_{hh}\right)$$

Remark 5.2.1

The variance of $\widehat{Y}_{hh}$ is zero when y_i is exactly proportional to x_i for every $i \in U$. This condition of the exact proportionality may not be realized in practice, but situations of approximate proportionality are not rare, such as the yield of a crop is approximately proportional to an area under cultivation and the income of a person is approximately proportional to his/her tax return. Hence, the Hansen–Hurwitz (HH) estimator becomes efficient when the study variable y is approximately proportional to the auxiliary variable x.

Remark 5.2.2

The estimator $\widehat{Y}_{hh}$ can be written as

$$\widehat{Y}_{hh} = \sum_{i=1}^{N} n_i(s_o)\frac{y_i}{p_i} \tag{5.2.5}$$

where $n_i(s_o)$ denotes the number of times the ith unit occurs in the ordered sample s_o. Now noting $n_i(s_o)$ follows a multinomial distribution with $E[n_i(s_o)] = np_i$, $V[n_i(s_o)] = np_i(1-p_i)$ and $Cov[n_i(s_o), n_j(s_o)] = -np_ip_j$ for $i \neq j$, one can prove the Theorem 5.2.2 in an alternative way.

5.2.4 Rao-Blackwellization

The HH estimator $\widehat{Y}_{hh}$ defined earlier is an ordered estimator because the estimator depends on the multiplicity of the units in the sample s_o. Hence $\widehat{Y}_{hh}$ is inadmissible as explained in Section 2.7. Let s be the unordered sample obtained by taking distinct units and arranging their labels in ascending orders from the ordered sample s_o, then by applying the Rao–Blackwell theorem (Theorem 2.7.3), one can find an improved estimator of the population total Y as follows:

$$t^* = E\left(\widehat{Y}_{hh}|s\right) = E\left[\frac{1}{n}\sum_{r=1}^{n}\frac{y_{(r)}}{p_{(r)}}\bigg| s\right] \tag{5.2.6}$$

But the estimator t^* cannot be used in practice because the expression (5.2.6) does not provide any elegant form. Let us consider the following example.

Example 5.2.3

Let $s_{o1} = (i, i, j)$ be an ordered sample of size $n = 3$ selected by the PPSWR method with $i < j$. The HH estimator based on s_{o1} is given by $\widehat{Y}_{hh}(s_{o1}) = \frac{1}{3}\left(2\frac{y_i}{p_i} + \frac{y_j}{p_j}\right)$. The ordered sample s_{o1} yields the unordered sample $s = (i, j)$. The unordered sample s would have been realized from any of the following ordered samples $s_{o1} = (i, i, j)$, $s_{o2} = (i, j, i)$, $s_{o3} = (j, i, i)$, $s_{o4} = (i, j, j)$, $s_{o5} = (j, i, j)$, and $s_{o6} = (j, j, i)$. Now noting that $\widehat{Y}_{hh}(s_{ok}) = \frac{1}{3}\left(2\frac{y_i}{p_i} + \frac{y_j}{p_j}\right)$ for $k = 1, 2, 3$; $\widehat{Y}_{hh}(s_{ok}) = \frac{1}{3}\left(\frac{y_i}{p_i} + 2\frac{y_j}{p_j}\right)$ for $k = 4, 5, 6$; $p(s_{o1}) = p(s_{o2}) = p(s_{o3}) = p_i^2 p_j$; and $p(s_{o4}) = p(s_{o5}) = p(s_{o6}) = p_i p_j^2$, we get the following improved unordered estimator:

$$\begin{aligned} t^* &= E\left(\widehat{Y}_{hh}|s\right) \\ &= \sum_{k=1}^{6}\widehat{Y}_{hh}(s_{ok})p(s_{ok}) \Big/ \sum_{k=1}^{6} p(s_{ok}) \\ &= \frac{1}{3}\left(\frac{y_i}{p_i} + \frac{y_j}{p_j} + \frac{y_i + y_j}{p_i + p_j}\right) \end{aligned} \tag{5.2.7}$$

Example 5.2.4

The Horvitz–Thompson estimator

$$\widehat{Y}_{ht} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{5.2.8}$$

with $\pi_i = 1 - (1 - p_i)^n$ as the inclusion probability of the ith unit is an unordered estimator and unbiased for the population total Y. However, this estimator is quite different from Eq. (5.2.7).

Remark 5.2.3

It is very difficult to make a theoretical comparison among unordered estimators in general. However, if y_i is proportional to x_i, the unordered estimator (5.2.7) becomes constant and its variance becomes zero. The Horvitz–Thompson estimator does not possess such properties. Hence the unordered estimator (5.2.7) is expected to fare better than $\widehat{Y}_{ht}$ when y_i is approximately proportional to x_i. Furthermore, Chaudhuri and Arnab (1978) have shown that the estimator $\widehat{Y}_{ht}$ is not consistent for the population total Y because the variance of $\widehat{Y}_{ht}$ does not necessarily decrease with the increase of the sample size n.

Remark 5.2.4

Although estimators (5.2.7) and (5.2.8) based on unordered data are more efficient than $\widehat{Y}_{hh}$, the estimators are seldom used in practice because of their complexity. Conversely, the inadmissible HH estimator $\widehat{Y}_{hh}$ is used extensively in practice mainly because of its simple expression and it possesses an elegant expression of its variance and unbiased estimator of its variance.

5.3 PROBABILITY PROPORTIONAL TO SIZE WITHOUT REPLACEMENT SAMPLING SCHEME

In a probability proportional to size without replacement (PPSWOR) sampling scheme, on the first draw, a unit i_1 (say) is selected with probability $p_{i_1}(1)$, which is proportional to its measure of size x_{i_1}, i.e., $p_{i_1}(1) = x_{i_1}/X = p_{i_1}$. On the second draw, another unit $i_2(\neq i_1)$ is selected from the remaining $N - 1$ units, which were not selected in the first draw with probability $p_{i_2|i_1}(2)$, proportional to its measure of size x_{i_2}. Hence the probability of selection of the unit i_2 in the second draw given that the unit i_1 is selected in the first draw is $p_{i_2|i_1}(2) = x_{i_2}/(X - x_{i_1}) = p_{i_2}\big/\big(1 - p_{i_1}\big)$. In general, at the $r(=1,\ldots,\ n)$th draw, the probability of selection of unit i_r given that the units $i_1,\ldots,\ i_{r-1}$ are selected in earlier $r-1$ draws is $p_{i_r|i_1\ldots i_{r-1}}(r) = p_{i_r}\big/\big(1 - p_{i_1} - \ldots - p_{i_{r-1}}\big)$. So, for the PPSWOR sampling, all units are distinct and the probability of selecting an ordered sample $s_o = (i_1,\ldots,\ i_n)$ where the unit i_r is selected at the rth draw is

$$\begin{aligned} p(i_1,\ldots,i_n) = p_{i_1} \times \frac{p_{i_2}}{1-p_{i_1}} \times \cdots \times \frac{p_{i_r}}{1-p_{i_1}-\cdots-p_{i_{r-1}}} \times \cdots \\ \times \frac{p_{i_n}}{1-p_{i_1}-\cdots-p_{i_{n-1}}} \end{aligned} \tag{5.3.1}$$

For all practical purposes, PPSWOR sample can be selected in the same manner as selection of a PPSWR sample by the cumulative method or Lahiri's (1951) method, keeping in mind that if a unit is selected once, it cannot be selected again. In other words, the PPSWR method is used until a set of n (desired sample size) distinct units are selected.

Example 5.3.1

Consider the data given in Example 5.2.1. Let us select a sample of size $n = 4$ units by the PPSWOR method using x_i's as the size measure for the ith unit. Here, we also have to start with Table 5.2.2, which was constructed in Example 5.2.1.

From the random number table we selected three-digit random numbers from 001 to 285 as given in Table 5.3.1:

Table 5.3.1

Random number selected	165	240	211	215	023	076
Unit selected	6	8	—	—	2	4

Here we note that we cannot select any unit corresponding to the random number 211 and 215 because the unit 8 has already been selected in the second draw. So the selected PPSWOR sample of size four is $s_o = (6, 8, 2, 4)$.

5.3.1 Raj's Estimator and its Variance

Let us now consider the following ordered estimators for the total Y based on the ordered sample $s_o = (i_1,\ldots,\ i_n)$.

$$t(1) = \frac{y_{i_1}}{p_{i_1}},\ t(2) = y_{i_1} + \frac{y_{i_2}}{p_{i_2}}(1 - p_{i_1})$$

and in general

$$t(r) = y_{i_1} + \cdots + y_{i_{r-1}} + \frac{y_{i_r}}{p_{i_r}}\left(1 - p_{i_1} - \cdots - p_{i_{r-1}}\right) \text{ for } r = 2, \ldots, n \tag{5.3.2}$$

The properties of ordered estimators are stated in the following theorem.

Theorem 5.3.1

(i) $E[t(r)] = Y$

(ii) $Var[t(r)] \leq Var[t(r-1)]$ for $r = 2,\ldots, n$

(iii) $Cov[t(r),\ t(k)] = 0$ for $r \neq k$

Proof

Let $E_{i_1,\ldots,i_k}$ and $V_{i_1,\ldots,i_k}$ denote unconditional expectation and variance for the selection of units $i_1,\ldots,\ i_k$. The conditional expectation and variance given that the units $i_1,\ldots,\ i_k$ are selected in earlier draws are denoted by $E(\bullet|i_1,\ldots,\ i_k)$ and $V(\bullet|i_1,\ldots,\ i_k)$, respectively.

(i) $$E[t(r)] = E_{i_1,\ldots,i_{r-1}}\left[E\left(t(r)|i_1, \ldots, i_{r-1}\right)\right] \tag{5.3.3}$$

Now

$$\begin{aligned} E\left(t(r)|i_1, \ldots, i_{r-1}\right) &= y_{i_1} + \cdots + y_{i_{r-1}} + E\left(\frac{y_{i_r}}{p_{i_r}}\left(1 - p_{i_1} - \cdots - p_{i_{r-1}}\right)|i_1, \ldots, i_{r-1}\right) \\ &= y_{i_1} + \cdots + y_{i_{r-1}} + \sum_{i_r \in U/i_1,\ldots i_{r-1}} \frac{y_{i_r}}{p_{i_r|i_1,\ldots i_{r-1}(r)}} \cdot p_{i_r|i_1,\ldots i_{r-1}} \\ &= Y \end{aligned} \tag{5.3.4}$$

where $U/i_1,\ldots,\ i_{r-1}$ = the set of $N - (r-1)$ units that were not selected in the first $r-1$ draws and $p_{i_r|i_1,\ldots,i_{r-1}} = p_{i_r}\big/\left(1 - p_{i_1} - \ldots - p_{i_{r-1}}\right)$.

Now substituting Eq. (5.3.4) in Eq. (5.3.3), we find $E[t(r)] = Y$.

(ii) $$\begin{aligned} Var[t(1)] &= Var\left(\frac{y_{i_1}}{p_{i_1}}\right) \\ &= \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \\ &= V_{pps} \text{ (Using Eq.5.2.4)} \end{aligned}$$

$$Var[t(r)] = E_{i_1,\dots,i_{r-1}}[V(t(r)|i_1,\dots,i_{r-1})] + V_{i_1,\dots,i_{r-1}}[E(t(r)|i_1,\dots,i_{r-1})] \tag{5.3.5}$$

Now using Eq. (5.3.4), we get

$$V_{i_1,\dots,i_{r-1}}[E\{t(r)|i_1,\dots,i_{r-1}\}] = V_{i_1,\dots,i_{r-1}}(Y) = 0 \tag{5.3.6}$$

and

$$\begin{aligned}V\{t(r)|i_1,\dots,i_{r-1}\} &= V\left(\frac{y_{i_r}}{p_{i_r}(r)}|i_1,\dots,i_{r-1}\right)\\ &= \frac{1}{2}\sum_{i\neq}\sum_{j\in U/i_1,\dots,i_{r-1}} p_{i|i_1,\dots,i_{r-1}}p_{j|i_1,\dots,i_{r-1}}\left(\frac{y_i}{p_{i|i_1,\dots,i_{r-1}}}-\frac{y_j}{p_{j|i_1,\dots,i_{r-1}}}\right)^2\\ &= \frac{1}{2}\sum_{i\neq}\sum_{j\in U/i_1,\dots,i_{r-1}} p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\end{aligned} \tag{5.3.7}$$

Substituting Eqs. (5.3.6) and (5.3.7) in Eq. (5.3.5), we get

$$\begin{aligned}Var[t(r)] &= E_{i_1\cdots i_{r-1}}\left[\frac{1}{2}\sum_{i\neq}\sum_{j\in U/i_1,\dots,i_{r-1}} p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\right]\\ &\leq E_{i_1\cdots i_{r-1}}\left[\frac{1}{2}\sum_{i\neq}\sum_{j\in U/i_1,\dots,i_{r-2}} p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\right]\\ &= E_{i_1\cdots i_{r-2}}\left[\frac{1}{2}\sum_{i\neq}\sum_{j\in U/i_1,\dots,i_{r-2}} p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\right]\\ &= Var[t(r-1)]\end{aligned}$$

(iii)
$$\begin{aligned}Cov[t(r),t(k)] &= E[t(r)t(k)] - [E\{t(r)\}][E\{t(k)\}]\\ &= E[t(r)t(k)] - Y^2\end{aligned} \tag{5.3.8}$$

Let $r < k$, then $E[t(r)t(k)] = E[t(r)E\{t(k)|i_1,\dots,i_r\}]$

$$\begin{aligned}&= Y\,E[t(r)] \quad \text{(Using Eq.5.3.4)}\\ &= Y^2\end{aligned} \tag{5.3.9}$$

Inserting Eq. (5.3.9) in Eq. (5.3.8), we get $Cov[t(r), t(k)] = 0$.

Raj's (1956) estimator for PPSWOR sampling is defined as

$$\widehat{Y}_{RA} = \frac{1}{n}\sum_{i=1}^{n} t(r) \tag{5.3.10}$$

The properties of Raj's estimator are given in the following theorem.

Theorem 5.3.2

(i) $\widehat{Y}_{RA}$ is unbiased for the total Y

(ii) Variance of $\widehat{Y}_{RA}$ is

$$V\left(\widehat{Y}_{RA}\right) \le \frac{V_{pps}}{n} = V\left(\widehat{Y}_{hh}\right)$$

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

$$\widehat{V}\left(\widehat{Y}_{RA}\right) = \frac{1}{n(n-1)}\sum_{r=1}^{n}\left(t(r) - \widehat{Y}_{RA}\right)^2$$

Proof

(i) $$E\left(\widehat{Y}_{RA}\right) = \frac{1}{n}\sum_{r=1}^{n} E[t(r)] = Y$$

(ii) $$V\left(\widehat{Y}_{RA}\right) = \frac{1}{n^2}\sum_{r=1}^{n} V[t(r)] \le \frac{V_{pps}}{n} = V\left(\widehat{Y}_{hh}\right)$$

(using the Theorem 5.3.1)

(iii) $$E\left[\widehat{V}\left(\widehat{Y}_{RA}\right)\right] = \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} E\{t(r)\}^2 - nE\left(\widehat{Y}_{RA}\right)^2\right]$$

$$= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n}\left(V\left(t(r)\right) + Y^2\right) - n\left\{V\left(\widehat{Y}_{RA}\right) + Y^2\right\}\right]$$

$$= V\left(\widehat{Y}_{RA}\right)$$

Remark 5.3.1

Theorem 5.3.2 indicates that for a given sample size, n, the Raj's estimator based on PPSWOR sampling is more efficient than the HH estimator based on PPSWR sampling of the same sample size.

Corollary 5.3.1

Let an ordered sample $s_o = (i, j)$ of size $n = 2$ be selected by the PPSWOR method, then

(i) $\widehat{Y}_{RA}(s_o) = \dfrac{1}{2}\left[\dfrac{y_i}{p_i}(1+p_i) + \dfrac{y_j}{p_j}(1-p_i)\right]$

$$\text{(ii) } V\left(\widehat{Y}_{RA}(s_o)\right) = \frac{1}{4}\sum_{i\neq}\sum_{j\in U}\left[\frac{y_i}{p_i}(1+p_i)+\frac{y_j}{p_j}(1-p_i)\right]^2\frac{p_ip_j}{1-p_i} - Y^2$$
$$= \frac{1}{2}\left(1-\frac{1}{2}\sum_{i\in U}p_i^2\right)\left\{\sum_{i\in U}p_i\left(\frac{y_i}{p_i}-Y\right)^2\right\}$$
$$-\frac{1}{4}\sum_{i\in U}p_i^2\left(\frac{y_i}{p_i}-Y\right)^2$$

$$\text{(iii) } \widehat{V}\left(\widehat{Y}_{RA}(s_o)\right) = \frac{(1-p_i)^2}{4}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

5.3.2 Rao-Blackwellization

It should be noted that Raj's (1956) estimator $\widehat{Y}_{RA}$ is an ordered estimator because it depends on the order of selection of units in the ordered sample. Raj's estimator based on the ordered sample $s_o = (i, j)$ is $\widehat{Y}_{RA}(s_o) = \dfrac{1}{2}\left[\dfrac{y_i}{p_i}(1+p_i) + \dfrac{y_j}{p_j}(1-p_i)\right]$, which is quite different from Raj's estimator $\widehat{Y}_{RA}(s_o^*) = \dfrac{1}{2}\left[\dfrac{y_j}{p_j}(1+p_j) + \dfrac{y_i}{p_i}(1-p_j)\right]$, which is based on the ordered sample $s_o^* = (j, i)$. Murthy (1957) improved Raj's estimator by using Rao-Blackwellization as follows.

5.3.2.1 Murthy's Estimator

Murthy (1957) symmetrized Raj's estimator by taking the weighted average of Raj's estimators using weights proportional to the probability of selection of the ordered samples. So, Murthy's estimator derived from the ordered samples $s_o = (i, j)$ or $s_o^* = (j, i)$ is given by

$$\widehat{Y}_M = \frac{\widehat{Y}_{RA}(s_o)p(s_o) + \widehat{Y}_{RA}(s_o^*)p(s_o^*)}{p(s_o)+p(s_o^*)}$$
$$= \frac{\dfrac{y_i}{p_i}(1-p_j)+\dfrac{y_j}{p_j}(1-p_i)}{(2-p_i-p_j)} \tag{5.3.11}$$

Let s be the unordered sample obtained from the ordered sample s_o and s_o^*, then we can write Eq. (5.3.11) as follows:

$$\begin{aligned}\widehat{Y}_M &= \frac{\widehat{Y}_{RA}(s_o)p(s_o) + \widehat{Y}_{RA}\left(s_o^*\right)p\left(s_o^*\right)}{p(s)}\\ &= \frac{1}{2}\frac{1}{p(s)}\left[\left\{\frac{y_i}{p_i}(1+p_i) + \frac{y_j}{p_j}(1-p_i)\right\}\frac{p_ip_j}{1-p_i}\right.\\ &= \frac{1}{p(s)}\left(y_i\frac{p_j}{1-p_i} + y_j\frac{p_i}{1-p_j}\right)\\ &= \frac{1}{p(s)}\left(y_ip(s|i) + y_jp(s|i)\right)\end{aligned} \tag{5.3.12}$$

where $p(s|k)$ denotes the conditional probability of obtaining the unordered sample s given that the unit $k(=i, j)$ is selected at the first draw.

Let $s = (j_1, \ldots, j_n)$ be an unordered sample of size n with labels $j_1 <, \ldots, < j_n$ obtained by the PPSWOR method of sampling, then Murthy's (1957) estimator can be written as

$$\widehat{Y}_M = \frac{1}{p(s)}\sum_{i\in s} y_ip(s|i) \tag{5.3.13}$$

where $p(s|i)$ = conditional probability of obtaining the unordered sample s when the ith unit is selected at the first draw.

The conditional probabilities satisfy the following relationships:

(i) $\sum_{i\in s} p_ip(s|i) = \sum_i I_{si}p_ip(s|i) = p(s)$,

(ii) $\sum_{s\supset i} p(s|i) = \sum_s I_{si}p(s|i) = 1 \text{ for } i = 1, \ldots, N$

and

(iii) $$\sum_{s\supset i,j} p(s|i,j) = \sum_s I_{si}I_{sj}p(s|i,j) = 1 \tag{5.3.14}$$

where $I_{si} = 1$ if $i \in s$, $I_{si} = 0$ if $i \notin s$ and $p(s|i,j)$ = conditional probability of s given that the units i and j have been selected in the first two draws.

Theorem 5.3.3

(i) $\widehat{Y}_M = \dfrac{1}{p(s)} \sum_{i\in s} y_i p(s|i)$ is unbiased of Y

(ii) Variance of $\widehat{Y}_M$ is

$$\begin{aligned} V_M &= \sum_{i=1}^{N} \beta_{ii} y_i^2 - \sum_{i\neq}^{N} \sum_{j=1}^{N} \beta_{ij} y_i y_j \\ &= \frac{1}{2} \sum_{i\neq}^{N} \sum_{j=1}^{N} \beta_{ij} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \end{aligned} \tag{5.3.15}$$

(iii) An unbiased estimator of V_M is

$$\widehat{V}_M = \frac{1}{2\{p(s)\}^2} \sum_{i\neq} \sum_{j\in s} \{p(s|i,j)p(s) - p(s|i)p(s|j)\} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2$$

where $\beta_{ii} = \sum_{s\supset i} \dfrac{\{p(s|i)\}^2}{p(s)} - 1$ and $\beta_{ij} = 1 - \sum_{s\supset i,j} \dfrac{p(s|i)p(s|j)}{p(s)}$.

Proof

$$\begin{aligned} \text{(i)}\quad E\left(\widehat{Y}_M\right) &= \sum_{s} \left\{ \frac{1}{p(s)} \sum_{i\in s} y_i p(s|i) \right\} p(s) \\ &= \sum_{i=1}^{N} y_i \sum_{s\supset i} p(s|i) \\ &= Y \text{ (using (ii) of Eq. 5.3.14)} \end{aligned}$$

$$\begin{aligned} \text{(ii)}\quad V_M = V(\widehat{Y}_M) &= E\left[\frac{1}{(p(s))^2} \left(\sum_{i\in s} y_i^2 (p(s|i))^2 + \sum_{i} \sum_{\neq j\in s} y_i y_j p(s|i) p(s|j) \right) \right] - Y^2 \\ &= \sum_{i=1}^{N} y_i^2 \left\{ \sum_{s\supset i} \frac{(p(s|i))^2}{p(s)} - 1 \right\} + \sum_{i\neq}^{N} \sum_{j=1}^{N} y_i y_j \left\{ \sum_{s\supset i,j} \frac{p(s|i)p(s|j)}{p(s)} - 1 \right\} \\ &= \sum_{i=1}^{N} \beta_{ii} y_i^2 - \sum_{i\neq}^{N} \sum_{j=1}^{N} \beta_{ij} y_i y_j \\ &= V_M \end{aligned}$$

Now

$$\frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N}\beta_{ij}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2=\sum_{i=1}^{N}\frac{y_i^2}{p_i}\sum_{j(\neq i)=1}^{N}p_j\beta_{ij}-\sum_{i\neq}^{N}\sum_{j=1}^{N}\beta_{ij}\,y_iy_j \tag{5.3.16}$$

and

$$\begin{aligned}\sum_{j(\neq i)=1}^{N}p_j\beta_{ij}&=1-p_i-\sum_{j(\neq i)=1}^{N}p_j\sum_s I_{si}I_{sj}p(s|i)p(s|j)/p(s)\\&=1-p_i-\sum_s I_{si}p(s|i)\sum_{j(\neq i)=1}^{N}I_{sj}p_jp(s|j)/p(s)\\&=1-p_i-\sum_s I_{si}p(s|i)\{p(s)-I_{si}p_ip(s|i)\}/p(s)\end{aligned} \tag{5.3.17}$$

[noting (i) of Eq. 5.3.14]

$$=p_i\left[\sum_{s\supset i}\frac{(p(s|i))^2}{p(s)}-1\right]\quad\text{[using (ii) of Eq. 5.3.14]}$$

$$=p_i\beta_{ii}$$

From Eqs. (5.3.16) and (5.3.17) we get

$$V_M=\frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N}\beta_{ij}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

(iii) $$E\left(\widehat{V}_M\right)=\frac{1}{2}\sum_s\frac{1}{p(s)}\sum_{i\neq}^{N}\sum_{j=1}^{N}I_{si}I_{sj}\{p(s|i,j)p(s)-p(s|i)p(s|j)\}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

$$=\frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\sum_s\frac{1}{p(s)}I_{si}I_{sj}\{p(s|i,j)p(s)-p(s|i)p(s|j)\}$$

Now noting $\sum_s\frac{1}{p(s)}I_{si}I_{sj}\{p(s|i,j)p(s)-p(s|i)p(s|j)\}=\beta_{ij}$ [from (iii) of Eq. 5.3.14], we find

$$E\left(\widehat{V}_M\right)=V_M$$

Remark 5.3.2

Pathak and Shukla (1966) proved that the variance estimator $\widehat{V}_M$ is nonnegative because $p(s|i,j)p(s) - p(s|i)p(s|j) \geq 0$ for $\forall\, i \neq j$.

Corollary 5.3.2

The expression of the Murthy's estimator $\widehat{Y}_M$, variance of $\widehat{Y}_M$, and an unbiased estimator of the variance of $\widehat{Y}_M$ based on an unordered sample $s = (i,\ j)$ with $i < j$ are, respectively, given as follows:

$$\text{(i)}\ \widehat{Y}_M = \frac{\dfrac{y_i}{p_i}\left(1 - p_j\right) + \dfrac{y_j}{p_j}\left(1 - p_i\right)}{2 - p_i - p_j}$$

$$\text{(ii)}\ V_M = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}\frac{p_i p_j\left(1 - p_i - p_j\right)}{2 - p_i - p_j}\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

$$\text{(iii)}\ \widehat{V}_M = \frac{\left(1 - p_i\right)\left(1 - p_j\right)\left(1 - p_i - p_j\right)}{\left(2 - p_i - p_j\right)^2}\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

Example 5.3.2

The number of publications and length of service of 10 academics in a University are given in Table 5.3.2.

Select a sample of size 3 academics by the PPSWOR method, using length of service as a measure of size variable. Estimate the average number of publications per academic by Raj's and Murthy's estimators. Give unbiased estimates of the variances of each of the estimators.

Table 5.3.2

Academics	1	2	3	4	5	6	7	8	9	10
Years of service (x)	10	5	15	10	25	30	10	10	20	15
No. of publications (y)	20	15	20	10	30	25	10	10	40	20

Here we make Table 5.3.3 for selection of sample.

From a random number table we select a five-digit random number 91050. This random number selects unit 10. The next five-digit random number 17898 selects the unit 3. The next five-digit number 19070 selects the unit 3 again. This selection is discarded because unit 3 has already been selected. The next five-digit number 99225 is also discarded because this number selects the unit 10, which has already been selected. The next five-digit number 32589 selects unit 5. Hence, the selected ordered sample of size 3 by PPSWOR method is $s_o = (10,\ 3,\ 5)$. Raj's estimator for total

Table 5.3.3

Unit (i)	x_i	y_i	$p_i = x_i/X$	$T_i =$ (cumulative p_i) $\times$ 100,000
1	10	20	0.06667	06,667
2	5	15	0.03333	10,000
3	15	20	0.10000	20,000
4	10	10	0.06667	26,667
5	25	30	0.16667	43,334
6	30	25	0.20000	63,334
7	10	10	0.06667	70,001
8	10	10	0.06667	76,668
9	20	40	0.13333	90,001
10	15	20	0.10000	100,000
Total	$150 = X$	–	1	

Y based on $s_o = (10, 3, 5)$ is $\widehat{Y}_{RA} = \bar{t} = \frac{1}{3}(t_1 + t_2 + t_3) = 194.66$ where $t_1 = \dfrac{y_{10}}{p_{10}} = 20/0.1 = 200$; $t_2 = \dfrac{y_3}{p_3}\,(1 - p_{10}) + y_{10} = (20/0.1) \times 0.9 + 20 = 200$; and $t_3 = \dfrac{y_5}{p_5} \times (1 - p_{10} - p_3) + y_{10} + y_3 = (30/0.16667) \times (1 - 0.1 - 0.1) + 20 + 20 \; = 183.99$ Raj's estimator for the population mean $\overline{Y}$ is $\widehat{\overline{Y}}_{RA} = \widehat{Y}_{RA}/10 = 19.466$. An unbiased estimator for variance of $\widehat{\overline{Y}}_{RA}$ is $\widehat{V}\left(\widehat{\overline{Y}}_{RA}\right) = \dfrac{1}{10^2}\,\dfrac{1}{3 \times 2}\sum_i (t_i - \bar{t})^2 = 0.284$.

Arranging the labels in ascending order of the ordered sample $s_o = (10, 3, 5)$, we get the unordered sample as $s = (3, 5, 10)$. This unordered sample would have been realized from any of the following ordered samples $s_{o1} = (3, 5, 10)$, $s_{o2} = (3, 10, 5)$, $s_{o3} = (5, 3, 10)$, $s_{o4} = (5, 10, 3)$, $s_{o5} = (10, 3, 5)$, and $s_{o6} = (10, 5, 3)$. The selection probabilities and Raj's estimator $\widehat{\overline{Y}}_{RA}$ based on the ordered samples are given in Table 5.3.4.

The conditional probabilities $p(s|i)$ are computed using the following formula:

$$p(s|3) = \frac{p_5}{1 - p_3}\,\frac{p_{10}}{1 - p_3 - p_5} + \frac{p_{10}}{1 - p_3}\,\frac{p_5}{1 - p_3 - p_{10}};$$

$$p(s|5) = \frac{p_3}{1 - p_5}\,\frac{p_{10}}{1 - p_5 - p_3} + \frac{p_{10}}{1 - p_5}\,\frac{p_3}{1 - p_5 - p_{10}} \quad \text{and}$$

$$p(s|10) = \frac{p_3}{1 - p_{10}}\,\frac{p_5}{1 - p_{10} - p_3} + \frac{p_5}{1 - p_{10}}\,\frac{p_3}{1 - p_{10} - p_5}$$

Table 5.3.4

Ordered sample (s_o)	$p(s_o)$	$\widehat{\overline{Y}}_{RA}(s_o)$	$p(s_o)\times\widehat{\overline{\mathbf{Y}}}_{\mathbf{RA}}(s_o)$
3, 5, 10	0.002525	19.28876	0.048710
3, 10, 5	0.002315	19.46657	0.045062
5, 3, 10	0.002727	19.11095	0.052122
5, 10, 3	0.002727	19.11095	0.052122
10, 3, 5	0.002315	19.46657	0.045062
10, 5, 3	0.002525	19.28876	0.048710
Total	$p(s)=0.015135$		0.291789

Conditional probabilities and the corresponding y_i values are given in Table 5.3.5:

Table 5.3.5

i	$p(s\|i)$	y_i	$y_ip(s\|i)$
3	0.048402	20	0.968035
5	0.032728	30	0.981827
10	0.048402	20	0.968035
Total	—	—	2.917897

Murthy's estimator for the population mean $\overline{Y}$ is

$$\widehat{\overline{Y}}_M=\frac{1}{Np(s)}\sum_{i\in s}y_ip(s|i)=2.917897/(10\times 0.015135)=19.278.$$

(Here we can check that $\sum p(s_o)\times\widehat{\overline{Y}}_{RA}(s_o)/p(s)=19.278.$)

An unbiased estimate of $V\left(\widehat{\overline{Y}}_M\right)$ is given by

$$\widehat{V}\left(\widehat{\overline{Y}}_M\right)=\frac{1}{100}\frac{1}{2\{p(s)\}^2}\sum_{i\neq}\sum_{j\in s}\{p(s|i,j)p(s)-p(s|i)P(s|j)\}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

$$=\frac{1}{100}\frac{1}{\{p(s)\}^2}\left[\{p(s|3,5)p(s)-p(s|3)P(s|5)\}p_3p_5\left(\frac{y_3}{p_3}-\frac{y_5}{p_5}\right)^2\right.$$

$$+\{p(s|3,10)p(s)-p(s|3)P(s|10)\}p_3p_{10}\left(\frac{y_3}{p_3}-\frac{y_{10}}{p_{10}}\right)^2$$

$$\left.+\{p(s|5,10)p(s)-p(s|5)P(s|10)\}p_3p_5\left(\frac{y_5}{p_5}-\frac{y_{10}}{p_{10}}\right)^2\right]$$

Now noting $p(s|3,\ 5)=p(s|5,\ 3)=p_{10}/(1-p_3-p_5)=0.136364$; $p(s|3,\ 10)=p(s|10,\ 3)=p_5/(1-p_3-p_{10})=0.208338$; and $p(s|5,\ 10)=p(s|10,\ 5)=p_3/(1-p_5-p_{10})=0.136364$, we prepare Table 5.3.6.

Table 5.3.6

i, j	$p(s\|i, j)$	$\delta_{ij} = p(s\|i, j)p(s) - p(s\|i)p(s\|j)$	$\left(\dfrac{y_i}{p_i} - \dfrac{y_j}{p_j}\right)^2$	$\delta_{ij}p_ip_j\left(\dfrac{y_i}{p_i} - \dfrac{y_j}{p_j}\right)^2$
3, 5	0.136364	0.00048	400.14401	0.003199698
3, 10	0.208338	0.00081	0	0
5, 10	0.136364	0.00048	400.14401	0.003199698
Total				0.006399306

$$\widehat{V}\left(\widehat{\overline{Y}}_M\right) = 0.006399396/\{10 \times \ (0.015135)\}^2 = 0.2793.$$

5.4 INCLUSION PROBABILITY PROPORTIONAL TO MEASURE OF SIZE SAMPLING SCHEME

The Horvitz–Thompson estimator $\widehat{Y}_{ht} = \sum_{i \in s} \dfrac{y_i}{\pi_i}$, based on a fixed sample size design, becomes constant if y_i's are proportional to the inclusion probabilities π_i's, and in this case the variance of $\widehat{Y}_{ht}$ becomes zero. Because the values of y_i's are unknown before the survey, one cannot construct a sampling design whose inclusion probabilities are proportional to the y_i values. In this situation, if an auxiliary variable x is available whose values are known, positive, and approximately proportional to the study variable y, then the variance of $\widehat{Y}_{ht}$ is expected to be small for a sampling design whose inclusion probabilities are proportional to the measure of size x variable, i.e., $\pi_i = nx_i/X = np_i \, (p_i = x_i/X)$ where n is the sample size. A sampling design is said to be inclusion probability proportional to measure of size (IPPS or πps) sampling scheme if $\pi_i = np_i$ for every $i \in U$. For an IPPS sampling design we must have

(i) $\pi_i = np_i \leq 1$, i.e., $p_i \leq 1/n$

Apart from the condition (i), it is desirable to stipulate the following conditions:

(ii) $\pi_{ij} > 0$ for $i \neq j \in U$

and

(iii) $\pi_i\pi_j - \pi_{ij} \geq 0$ for $i \neq j \in U$

The condition (ii) is required for unbiased estimation of variance while the condition (iii) is a requirement for obtaining nonnegative Yates-Grundy's (1953) variance estimator. To control the variance of the Horvitz–Thompson estimator, Hanurav (1966) set an additional restriction

(iv) $\pi_i\pi_j/\pi_{ij}$ should not be close to zero.

The IPPS sampling designs have been constructed by various authors viz. Narain (1951), Brewer (1963a,b), Fellegi (1963), Rao (1965), Durbin (1967), Hanurav (1966), and Sampford (1967), among others. A detailed review is given by Brewer and Hanif (1983). In this section, we will describe a few of them in detail. In general, IPPS sampling schemes for $n > 2$ are very complex. The expressions for π_{ij}'s are also not simple and hence the expressions of variances are very complex in general.

5.4.1 Inclusion Probability Proportional to Measure of Size Sampling With $n = 2$

5.4.1.1 Brewer's Sampling Scheme

In Brewer's (1963a,b) method, at the first draw, the ith unit is selected with probability

$$p_i(1) = \frac{2p_i(1-p_i)}{A(1-2p_i)}$$

where

$$\begin{aligned} A &= \sum_{i\in U} \frac{2p_i(1-p_i)}{1-2p_i} \\ &= \sum_{i\in U} \frac{p_i(1+1-2p_i)}{1-2p_i} \\ &= 1 + \sum_{i\in U} \frac{p_i}{1-2p_i} \end{aligned} \tag{5.4.1}$$

The conditional probability of selecting the ith unit in the second draw when the jth unit is selected in the first draw is

$$p_{i|j}(2) = \begin{cases} p_i/(1-p_j) & \text{for} \quad i \neq j \\ 0 & \text{for} \quad i = j \end{cases}$$

So, for this method, the inclusion probability of the ith unit is

$$\begin{aligned} \pi_i &= p_i(1) + \sum_{j\neq i} p_j(1) p_{i|j}(2) \\ &= \frac{2p_i}{A}\left(\frac{1-p_i}{1-2p_i} + \sum_{j\neq i} \frac{p_j}{1-2p_j}\right) \\ &= 2p_i. \end{aligned}$$

Inclusion probability for the ith and jth unit ($i \neq j$) is

$$\begin{aligned} \pi_{ij} &= p_i(1) p_{j|i}(2) + p_j(1) p_{i|j}(2) \\ &= \frac{2p_ip_j}{A}\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right) \end{aligned} \tag{5.4.2}$$

and the difference

$$\begin{aligned}\pi_i\pi_j - \pi_{ij} &= \frac{2p_ip_j}{A}\left\{2A - \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\right\}\\ &= \frac{4p_ip_j}{A}\sum_{k\neq(i,j)}\frac{p_k}{1-2p_k}\\ &> 0\end{aligned}$$

5.4.1.2 Durbin's Sampling Scheme

In Durbin's (1967) sampling scheme, the probability of selection of the ith unit at the first draw is $p_i(1) = p_i$, $i \in U$ and the conditional probability of selection of the ith unit in the second draw given that the jth unit selected at the first draw is

$$p_{i|j}(2) = \begin{cases} p_i\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\Big/ A & \text{for} \quad i \neq j\\ 0 & \text{for} \quad i = j\end{cases}$$

where A is given in Eq. (5.4.1) and clearly $\sum\limits_{i\in U} p_{i|j}(2) = 1$.

The inclusion probability for the ith unit is

$$\begin{aligned}\pi_i &= p_i(1) + \sum_{j\neq i} p_j(1)p_{i|j}(2)\\ &= p_i + \sum_{j\neq i} p_jp_i\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\Big/ A\\ &= 2p_i\end{aligned}$$

Inclusion probability for the ith and jth unit ($i \neq j$) is

$$\begin{aligned}\pi_{ij} &= p_i(1)p_{j|i}(2) + p_j(1)p_{i|j}(2)\\ &= 2p_ip_j\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\Big/ A\end{aligned} \qquad (5.4.3)$$

Remark 5.4.1

Brewer (1963a,b) and Durbin's (1967) schemes are identical in the sense of having the same expressions of the second order inclusion probabilities. In the same sense, Rao's (1965) sampling scheme is also similar to them.

Furthermore, all these three sampling schemes satisfy $\pi_{ij} > 0$ and $\pi_i\pi_j - \pi_{ij} \geq 0$ for $i \neq j \in U$.

5.4.1.3 Hanurav's Sampling Scheme

Hanurav's (1967) IPPS sampling scheme is defined as follows. Select two units with replacement with probability $p_i(1) = p_i$ attached to the ith unit. If the two units are distinct, accept them as a sample. Otherwise cancel this selection and select two fresh units with replacement with probability $p_i(2) = p_i^2 \Big/ \left(\sum\limits_1^N p_k^2\right) = \{p_i(1)\}^2 \Big/ \sum\limits_1^N \{p_k(1)\}^2$ for the ith unit. If the two units are different, select them as a sample, otherwise select two fresh units with replacement with probability $p_i(3) = p_i^4 \Big/ \left(\sum\limits_1^N p_k^4\right) = \{p_i(2)\}^2 \Big/ \sum\limits_1^N \{p_k(2)\}^2$ for the ith unit. If the units are distinct accept them as a sample, otherwise continue the procedure till the two units are distinct.

The inclusion probability of ith unit $\pi_i =$ probability of selection of the ith unit in first attempt + probability of selection of the ith unit in second attempt + probability of selection of the ith unit in third attempt + $\cdots$

$$
\begin{aligned}
&= \left[\sum_{j\neq i}\{p_i(1)p_j(1)+p_j(1)p_i(1)\}\right] + \left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\sum_{j\neq i}\{p_i(2)p_j(2)+p_j(2)p_i(2)\}\right]\\
&\quad + \left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\left(\sum_{j=1}^N \{p_{j(2)}\}^2\right)\sum_{j\neq i}\{p_i(3)p_j(3)+p_j(3)p_i(3)\}\right] + \cdots\\
&= [2p_i(1)\{1-p_i(1)\}] + 2\left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\Big(p_i(2)\{1-p_i(2)\}\Big)\right]\\
&\quad + 2\left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\left(\sum_{j=1}^N \{p_{j(2)}\}^2\right)\right]\big(p_i(3)\{1-p_i(3)\}\big) + \cdots\\
&= 2[p_i(1)-\{p_i(1)\}^2] + 2\left[\{p_i(1)\}^2 - \left[\sum_{j=1}^N \{p_{j(1)}\}^2\right]\{p_i(2)\}^2\right]\\
&\quad + 2\left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\left(\left(\sum_{j=1}^N \{p_j(2)\}^2\right)p_i(3)\right)\right]\\
&\quad - 2\left[\left(\sum_{j=1}^N \{p_{j(1)}\}^2\right)\left(\sum_{j=1}^N \{p_{j(2)}\}^2\right)\{p_j(3)\}^2\right] + \cdots\\
&= 2p_i(1)\\
&= 2p_i
\end{aligned}
$$

The inclusion probability of ith and jth unit $(i \neq j)$

$$\begin{aligned}&= \pi_{ij}\\ &= 2p_i(1)p_j(1) + \left[\sum_{k=1}^{N} \{p_k(1)\}^2\right] 2p_i(2)p_j(2) + \left[\sum_{k=1}^{N} \{p_k(1)\}^2\right]\\ &\quad \times \left[\sum_{k=1}^{N} \{p_k(2)\}^2\right] 2p_i(3)p_j(3) + \cdots\end{aligned}$$

It can be shown that for this sampling scheme $\pi_i\pi_j - \pi_{ij} \geq 0$.

5.4.2 Inclusion Probability Proportional to Measure of Size Sampling with *n* > 2

5.4.2.1 Lahiri–Midzuno–Sen Sampling Design

The simplest IPPS sampling scheme for an arbitrary sample size was proposed independently by Lahiri (1951), Midzuno (1952), and Sen (1953) but is applicable under the restrictive condition $\frac{n-1}{n(N-1)} < p_i < \frac{1}{n}$. For the LMS sampling design, at the first draw, the ith unit is selected with probability θ_i (>0, to be determined and subject to $\sum_{i \in U} \theta_i = 1$), called the revised normed size measure. Then the remaining $n-1$ units are selected from those $N-1$ units, which are not selected in the first draw by the SRSWOR method. Here, θ_i's are chosen to make the inclusion probability

$$\begin{aligned}\pi_i &= \theta_i + \sum_{j(\neq i)} \theta_j \frac{n-1}{N-1}\\ &= \frac{N-n}{N-1}\theta_i + \frac{n-1}{N-1}\end{aligned} \tag{5.4.4}$$

equals to np_i.

Now equating $\pi_i = np_i$, we get from Eq. (5.4.4)

$$\theta_i = \frac{(N-1)n}{N-n} p_i - \frac{n-1}{N-n} \tag{5.4.5}$$

Now putting a restriction on θ_i to be positive, we get a restriction on p_i as

$$p_i > \frac{n-1}{n(N-1)}$$

Thus the LMS–IPPS sampling scheme is applicable, if and only if,

$$\frac{n-1}{n(N-1)} < p_i < \frac{1}{n} \tag{5.4.6}$$

The inclusion probability of the *i*th and *j*th unit ($i \neq j$) for this sampling scheme is π_{ij} = probability of selection of the *i*th unit at the first draw and the *j*th unit in any of the remaining $n-1$ draws + probability of selection of the *j*th unit at the first draw and the *i*th unit in any of the remaining $n-1$ draws + probability that none of the *i*th and *j*th units were selected in the first draw but are selected in the remaining $n-1$ draws

$$= \theta_i \frac{n-1}{N-1} + \theta_j \frac{n-1}{N-1} + (1-\theta_i-\theta_j)\frac{(n-1)(n-2)}{(N-1)(N-2)}$$

$$= \frac{(n-1)(N-n)}{(N-1)(N-2)}(\theta_i+\theta_j) + \frac{(n-1)(n-2)}{(N-1)(N-2)} \tag{5.4.7}$$

$$= \frac{n(n-1)}{(N-2)}\left[p_i + p_j - \frac{1}{(N-1)}\right] \quad \text{(using Eq. 5.4.5)}$$

It can be easily checked that for this sampling scheme $\pi_i\pi_j - \pi_{ij}$ is positive.

Remark 5.4.2

For the sample size $n=2$, Rao (1963) proved that the Horvitz–Thompson estimator $\widehat{Y}_{ht} = \sum_{i \in s} \frac{y_i}{\pi_i} = \frac{1}{n}\sum_{i \in s} \frac{y_i}{p_i}$ based on the LMS–IPPS sampling scheme possesses lower variance than that of the HH estimator based on the PPSWR sampling scheme of the same sample size. Asok (1974) and Asok and Sukhatme (1978) proved that the Rao's result is valid for $n \geq 2$, i.e.,

$$V\left(\widehat{Y}_{ht}\right) \leq V\left(\widehat{Y}_{hh}\right) = \sum_{i \in U} p_i \left(\frac{y_i}{p_i} - Y\right)^2 \Big/ n. \tag{5.4.8}$$

5.4.2.2 Probability Proportionate to Size Systematic Sampling Scheme

Madaw (1949) and Goodman and Kish (1950) proposed the PPS systematic sampling procedure, which ensures that $\pi_i = np_i$. The proposed IPPS sampling scheme is very simple to execute and is applicable to any value of n as long as $\pi_i = np_j \leq 1$. But the main drawback of this procedure is that the expressions for the second-order inclusion probabilities are highly complex. The PPS systematic sampling procedure is described as follows.

Let $T_i = n\sum_{j=1}^{i} p_j$ for $i = 1,\ldots, N$ and $T_0 = 0$. Select a random sample (called random start) d from a uniform distribution with the range (0, 1). This random start d selects a sample with those units whose index, "j," satisfies $T_{j-1} \leq d + k < T_j$ for $k = 0, 1,\ldots, n - 1$.

It can easily be noted that each value of k results in the selection of only one unit because $np_i \leq 1$. For this sampling scheme, the inclusion probability of the ith unit is $\pi_i = T_i - T_{i-1} = np_i$. Assuming that the labels of the units are attached at random, Hartley and Rao (1962) gave the following approximate expression of π_{ij} to the order $O(N^{-4})$ when p_i is of $O(N^{-1})$, n is relatively small to N, and N is moderately large.

$$\pi_{ij} = n(n-1)p_ip_j\left[\begin{array}{l} 1 + \left\{(p_i + p_j) - \sum_j p_j^2\right\} + \left\{2\left(p_i^2 + p_j^2\right) - 2\sum_j p_j^3\right. \\ \left. + 2p_ip_j - 3(p_i + p_j)\sum_j p_j^2 + 3\left(\sum_j p_j^2\right)^2\right\} \end{array}\right] \tag{5.4.9}$$

An approximate expression the variance of the $\widehat{Y}_{ht}$ was obtained by Harley and Rao (1962) correct to $O(N^{-2})$ as follows:

$$\begin{aligned} V\left(\widehat{Y}_{ht}\right) &= \frac{1}{n}\left[\sum_i p_i z_i^2 - (n-1)\sum_i p_i^2 z_i^2\right] \\ &\quad - \frac{(n-1)}{n}\left[2\sum_i p_i^3 z_i^2 - \left(\sum_i p_i^2\right)\left(\sum_i p_i^2 z_i^2\right) - 2\left(\sum_i p_i^2 z_i\right)^2\right] \\ &= V_{GK} \end{aligned} \tag{5.4.10}$$

where $\quad z_i = \dfrac{y_i}{p_i} - Y.$

The approximate expression for V_{GK} correct to $O(N^{-1})$ is given by

$$\begin{aligned} V_{GK} &= \frac{1}{n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 - \frac{n-1}{n}\sum_{i\in U} p_i^2\left(\frac{y_i}{p_i} - Y\right)^2 \\ &< \frac{1}{n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 = V\left(\widehat{Y}_{hh}\right) \end{aligned} \tag{5.4.11}$$

The expression (5.4.11) indicates that the variance of the Horvitz–Thompson estimator based on the PPS systematic sampling design provides a smaller variance than that of the HH estimator based on PPWR sampling of the same size.

An approximate expression for an unbiased estimator of V_{GK} correct to $O(N)$ was given by Hartley and Rao (1962) as follows:

$$\widehat{V}_{GK} = \left(\frac{1}{n^2(n-1)}\right)\frac{1}{2}\left[\sum_{i\neq}\sum_{j\in s}\left\{1-n(p_i+p_j)+n\left(\sum_{i=1}^{N}p_i^2\right)\right\}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\right] \tag{5.4.12}$$

Example 5.4.1

Consider a population of size $N = 10$ units from which an IPPS sample of size $n = 4$ is to be selected. The measures of size (x_j) are given in Table 5.4.1.

Table 5.4.1

Unit	1	2	3	4	5	6	7	8	9	10
Measure of size (x_j)	10	25	24	15	12	18	35	32	18	11
np_j	0.20	0.50	0.48	0.30	0.24	0.36	0.70	0.64	0.36	0.22
T_j	0.20	0.70	1.18	1.48	1.72	2.08	2.78	3.42	3.78	4.00

Let a random start $d = 0.382$ be selected from a uniform distribution (0, 1). For $k = 0$, unit 2 is selected because $T_1 \leq d = 0.382 < T_2$; similarly, corresponding to $k = 1$, 2, and 3 units 4, 7, and 8 are selected, respectively, because $T_3 \leq d + 1 = 1.382 < T_4$, $T_6 \leq d + 2 = 2.382 < T_7$, and $T_7 \leq d + 3 = 3.382 < T_8$.

5.4.2.3 Sampford's Sampling Scheme

In Sampford's (1967) IPPS sampling scheme, on the first draw the ith ($i = 1,\ldots, N$) unit is selected with probability p_i, i.e., $p_i(1) = p_i$. Then the remaining $(n - 1)$ units are drawn with replacement from the entire population with probability proportional to $\lambda_i = p_i/(1 - np_i)$, i.e., the probability of selecting the ith unit at the kth draw is $p_i(k) = \lambda_i \Big/ \sum\limits_{j=1}^{N} \lambda_j$ $k = 2,\ldots, n$; $i = 1,\ldots N$. The selected units are accepted as a sample if all the n units happened to be different, otherwise the entire selection is discarded, and this process is repeated unless a set of n distinct units is obtained.

Sampford (1967) has shown that the inclusion probability for the selection of ith unit is $\pi_i = np_i$ and $\pi_i\pi_j - \pi_{ij} \geq 0$. The expression for the second-order inclusion probabilities is not simple. However, approximate expression of π_{ij} correct to $O(N^{-4})$, derived by Asok and Sukhatme (1976), is given for $n \geq 3$ as follows:

$$\pi_{ij} = n(n-1)p_i p_j \left[\begin{array}{l} 1 + \left(p_i + p_j - \sum_j p_j^2\right) \\ + \left(2\left(p_i^2 + p_j^2\right) - 2\sum_j p_j^3 - (n-2)p_i p_j \right. \\ \left. + (n-3)(p_i + p_j)\sum_j p_j^2 - (n-3)\left(\sum_j p_j^2\right)^2\right) \end{array} \right] \tag{5.4.13}$$

For Sampford's sampling, the variance of the Horvitz–Thompson estimator correct to $O(N^{-2})$ is given by Asok and Sukhatme (1976) as

$$\begin{aligned} V(\widehat{Y}_{ht}) &= \frac{1}{n}\left(\sum_i p_i z_i^2 - (n-1)\sum_i p_i^2 z_i^2\right) \\ &\quad - \frac{(n-1)}{n}\left(2\sum_i p_i^3 z_i^2 - \left(\sum_i p_i^2\right)\left(\sum_i p_i^2 z_i^2\right)\right. \\ &\quad \left. + (n-2)\left(\sum_i p_i^2 z_i\right)^2\right) \\ &= V_{SAM} \end{aligned}$$

where z_i is given in Eq. (5.4.10).

5.4.2.3.1 Comparison of Efficiency

Asok and Sukhatme (1976) showed that the variances of the Horvitz–Thompson estimator for a finite population total Y based on the PPS systematic sampling scheme and Sampford's (1967) sampling scheme correct to $O(N^{-1})$ are exactly equal to

$$V_{GK} = V_{SAM} = \frac{1}{n}\left[\sum_i p_i z_i^2 - (n-1)\sum_i p_i^2 z_i^2\right] \tag{5.4.14}$$

The above expression (5.4.14) indicates that the Horvitz–Thompson estimator based on the PPS systematic as well as Sampford procedures possesses a uniformly smaller variance than that of the HH estimator based on a PPSWR sampling design of the same sample size n. Furthermore, when the variance is considered to $O(N^{-2})$, the Horvitz-Thompson estimator based on Sampford sampling has a uniformly smaller variance than that of the PPS systematic sampling procedure. Their difference

$$V_{GK} - V_{SAM} = (n-1)\left(\sum_i p_i^2 z_i\right)^2$$

is nonnegative and increases with the sample size n.

Remark 5.4.3

For $n = 2$, Sampford (1967), Durbin (1967), and Brewer's (1963a,b) sampling designs are identical in the sense that the probability of getting an unordered sample $s = (i,\ j)$, $i < j$ is the same and is equal to $2p_ip_j\left(\frac{1}{1-2p_i}+\frac{1}{1-2p_j}\right)\Big/A$ where A is given in Eq. (5.4.1).

Example 5.4.2

Table 5.4.2 relates to the annual income tax paid by 10 employees of a certain university. Select a sample of size 4 by Sampford's IPPS sampling procedure using the amount of tax paid as measure of size variable (x).

In Sampford sampling, the first unit is selected with probability proportional to a measure of size x. So, we select the first unit by the cumulative total method and prepare Table 5.4.3.

Here we select a six-digit random number between 000001 and 360000. From the random number table we select random number 029092. This random number selects unit 1. The remaining three units needed to be selected with replacement from the units 1 to 10 with probability $q_i = \frac{p_i/(1-4p_i)}{\sum_{i=1}^{10} p_i/(1-4p_i)}$ attached to the ith unit, $i = 1,\ldots,\ 10$. The values of q_i's and cumulative totals Q_i are given in Table 5.4.4.

Now we select three four-digit random numbers from a random number table, and putting a decimal point to the left of each number, we get the numbers as follows: 0.9356, 0.1892, 0.4598. Now looking at the cumulative total Q_i values, we select units 9, 3, and 6. Because the selected units (1, 9, 3, 6) are all distinct, we accept $s = (1, 3, 6, 9)$ as a sample according to Sampford's IPPS sampling scheme. It should be noted that if all the units were not distinct, we would have to repeat the procedure until all the four selected units were distinct.

Table 5.4.2

Serial no. of employees (*i*)	1	2	3	4	5	6	7	8	9	10
Tax paid in dollars (x_i)	30,000	35,000	45,000	25,000	20,000	60,000	40,000	50,000	30,000	25,000

Table 5.4.3

Serial no. of employees (*i*)	1	2	3	4	5	6	7	8	9	10
Tax paid in dollars (x_i)	30,000	35,000	45,000	25,000	20,000	60,000	40,000	50,000	30,000	25,000
Cumulative total T_i	30,000	65,000	110,000	135,000	155,000	215,000	255,000	305,000	335,000	360,000

Table 5.4.4

Unit	1	2	3	4	5	6	7	8	9	10
$\frac{p_i}{(1-4p_i)}$	0.125	0.1591	0.25	0.0962	0.0714	0.5	0.2	0.3125	0.125	0.0962
q_i	0.0646	0.0822	0.1292	0.0497	0.0369	0.2583	0.1033	0.1615	0.0646	0.0497
Q_i	0.0646	0.1468	0.276	0.3257	0.3626	0.6209	0.7242	0.8857	0.9503	1

5.4.2.4 Poisson (or Bernoulli) Sampling

In the Poisson or Bernoulli sampling scheme, the units are selected by performing N (population size) Bernoulli trials independently. The ith ($i = 1,\ldots, N$) unit is selected by performing a Bernoulli trial with a success probability π_i. If the trial produces a success, the ith unit is selected in the sample, otherwise it is not included in the sample. Clearly, in a Poisson sampling scheme, the sample size is not fixed. It is a random variable. But the expected sample size $\sum\limits_{i=1}^{N}\pi_i$ becomes fixed as n if $\pi_i = np_i$. For the Poisson sampling scheme, the inclusion probability for the ith unit is π_i and inclusion probability of a pair of units i, $j(i \neq j)$ is $\pi_{ij} = \pi_i\pi_j$ (because the draws are independent). Hence for a Poisson sampling $\pi_i\pi_j - \pi_{ij} = 0$. The expression for the variance of $\widehat{Y}_{ht} = \sum\limits_{i\in s} y_i/\pi_i$ and its unbiased estimator are very simple and they are given, respectively, as follows:

$$V\left(\widehat{Y}_{ht}\right) = \sum_{i=1}^{N} y_i^2\left(\frac{1}{\pi_i} - 1\right) \tag{5.4.15}$$

and

$$\widehat{V}\left(\widehat{Y}_{ht}\right) = \sum_{i\in s}\frac{y_i^2}{\pi_i}\left(\frac{1}{\pi_i} - 1\right) \tag{5.4.16}$$

5.4.2.5 Use of Combinatorics

Let $\mathfrak{I} = (s_1,\ldots, s_j,\ldots, s_b)$ be the collection of b possible samples each of effective size n and $p(s_j)$ be the probability of selection of the sample s_j with $p(s_j) \geq 0$, $\sum\limits_{j=1}^{b} p(s_j) = 1$. Then the inclusion probability of the ith unit is $\pi_i = \sum\limits_{j=1}^{b} n_{ij}p(s_j)$, where $n_{ij} = 1$, if the sample s_j contains the ith unit and $n_{ij} = 0$ otherwise; $i = 1,\ldots, N$; $j = 1,\ldots, b$. So, we can write the inclusion probability matrix as

$$\mathbf{\Pi} = \mathbf{N}\cdot\mathbf{P} \tag{5.4.17}$$

where

$$\mathbf{\Pi} = \begin{pmatrix} \pi_1 \\ \bullet \\ \pi_i \\ \bullet \\ \pi_N \end{pmatrix}, \quad \mathbf{N} = \begin{pmatrix} n_{11} & \bullet & n_{1j} & \bullet & n_{1b} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ n_{i1} & \bullet & n_{ij} & \bullet & n_{ib} \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ n_{N1} & \bullet & n_{Nj} & \bullet & n_{Nb} \end{pmatrix} \text{ and } \mathbf{P} = \begin{pmatrix} p(s_1) \\ \bullet \\ p(s_j) \\ \bullet \\ p(s_b) \end{pmatrix}$$

Thus **P** will be an IPPS sampling design if a nonnegative solution of **P** for Eq. (5.4.17) exists for $\pi_i(=np_i)$ for $i=1,\ldots,N$. The solution of Eq. (5.4.17) was derived by Gupta et al. (1982) using the properties of a balanced incomplete block design (BIBD). If a unit is considered as a treatment and a sample as a block, then the matrix **N** may be considered as an incidence matrix of an incomplete block design with parameter $v=$ number of treatments $=$ total numbers of units $=N$, $b=$ total number of blocks of a design $=$ total number of samples, and $k=$ total number of treatments in a block $=$ sample size $=n$. Furthermore, the incidence matrix **N** corresponds to a BIBD if every treatment (unit) is exactly repeated in r blocks (samples) and any two treatments (units) occur together in λ blocks (samples). The parameters of a BIBD design are denoted by v, b, r, k, and λ. The parameters satisfy (i) $bk=vr$ and (ii) $\lambda(v-1)=r(k-1)$ [vide Raghavarao, 1971]. Here, we will suppose that the b samples (blocks) are so chosen that **N** corresponds to an incidence matrix of a BIBD.

Theorem 5.4.1

Let the samples s_j's are selected with probability

$$p(s_j)=v\left(r\sum_{i\in s_j}p_i-\lambda\right)\Big/\{b(r-\lambda)\}\quad\text{for}\quad j=1,\ldots b$$

with $\sum_{i\in s_j}p_i\geq\lambda/r=(n-1)/(N-1)$.

Then, (i) $\pi_i=np_i$

$$\text{(ii)}\pi_{ij}=\frac{v}{b(r-\lambda)}\left[r\left\{\lambda(p_i+p_j)+\sum_{l(\neq i,j)=1}^{N}p_l\,\mu_{ijl}\right\}-\lambda^2\right]\qquad(5.4.18)$$

where μ_{ijl} is the number of times the ith, jth, and lth $(i\neq j\neq l)$ treatments (units) occur together in the same block (sample).

Proof

(i) $$\pi_i=\sum_{j=1}^{b}n_{ij}p(s_j)=v\sum_{j=1}^{b}n_{ij}\left(r\sum_{t=1}^{N}n_{tj}p_t-\lambda\right)\Big/\{b(r-\lambda)\}$$

$$=v\left[r\left(p_i\sum_{j=1}^{b}n_{ij}+\sum_{t(\neq i)=1}^{N}p_t\sum_{j=1}^{b}n_{ij}n_{tj}\right)-\lambda\sum_{j=1}^{b}n_{ij}\right]\Big/[b(r-\lambda)]$$

$$=v\left[r\left(r\,p_i+\lambda\sum_{t(\neq i)=1}^{N}p_t\right)-\lambda r\right]\Big/[b(r-\lambda)]$$

$$=vrp_i/b=np_i\,(\text{since for a BIDB, }bk=vr\ \text{ and }\ k=n)$$

$$\text{(ii)}\quad \pi_{ij}=\sum_{t=1}^{b}n_{it}\,n_{jt}p(s_t)=v\sum_{t=1}^{b}n_{it}\,n_{jt}\left(r\sum_{l=1}^{N}n_{lt}p_l-\lambda\right)\Big/[b(r-\lambda)]$$

$$=v\left[\sum_{t=1}^{b}n_{it}\,n_{jt}\,r\left(\left(n_{it}p_i+n_{jt}p_j\right)+\sum_{l(\neq i,j)=1}^{N}n_{lt}p_l\right)-\lambda^2\right]\Big/[b(r-\lambda)]$$

$$=v\left[r\left(\lambda\left(p_i+p_j\right)+\sum_{l(\neq i,j)=1}^{N}\mu_{ijl}p_l\right)-\lambda^2\right]\Big/[b(r-\lambda)]$$

The previous expression of π_{ij} is complex and one is not sure about the sign of $\pi_i\pi_j-\pi_{ij}$. However, instead of considering BIBD, if we consider doubly balanced incomplete block design (DBIBD) introduced by Calvin (1954), where every triplet of treatment (units) appears together in the same block (sample), an equal number of times (μ, say), we may construct an IPPS sampling design with π_{ij} identical to the LMS—IPPS sampling scheme satisfying $\pi_i\pi_j-\pi_{ij}\geq 0\ \forall\, i\neq j$. For a DBIBD, $\mu(v-2)=\lambda(k-2)$ [vide Hedayat and Kageyama, 1980].

Theorem 5.4.2

For a DBIBD, the expression (5.4.18) reduces to $\pi_{ij}=\dfrac{n(n-1)}{N-2}\times\left(p_i+p_j\right)-\dfrac{n(n-1)}{(N-1)(N-2)}$.

Proof

From Eq. (5.4.18), we get

$$\pi_{ij}=\frac{v}{b(r-\lambda)}\left[r\left(\lambda(p_i+p_j)+\mu\sum_{k(\neq i,j)=1}^{N}p_k\right)-\lambda^2\right]$$

$$=\frac{v}{b(r-\lambda)}\left[r\left((\lambda-\mu)(p_i+p_j)+\mu\right)-\lambda^2\right]$$

Now using the relations (i) $bk = vr$, (ii) $\lambda(v-1) = r(k-1)$, and (iii) $\mu(v-2) = \lambda(k-2)$, we get

$$\begin{aligned}\pi_{ij} &= \frac{k(k-1)}{(v-2)}(p_i+p_j) - \frac{k(k-1)}{(v-1)(v-2)} \\ &= \frac{n(n-1)}{(N-2)}(p_i+p_j) - \frac{n(n-1)}{(N-1)(N-2)} \\ &\quad (\text{noting } k = n \text{ and } v = N)\end{aligned}$$

Remark 5.4.4

Gupta et al. (1982) IPPS design requires $\sum_{i\in s_j} p_i \geq \lambda/r = (n-1)/(N-1)$, which is less restrictive than that of the LMS—IPPS sampling scheme, which requires $p_i \geq (n-1)/\{n(N-1)\}$ for every $i = 1,\ldots, N$.

Remark 5.4.5

The construction of sampling designs realizing preassigned sets of values of inclusion probabilities of the first and second order were provided by Sinha (1973). Arnab and Roy (1990) have given methods of construction of such sampling designs by using of BIB designs.

5.4.2.6 The Nearest Proportional to Size Sampling

Suppose a sampling design p_0 is preferable because of practical consideration, such as cost. But on the other hand, an IPPS sampling design p^* with inclusion probability $\pi_i = np_i$ is desirable for theoretical considerations, such as efficiency. Gabler (1987a,b) provided a method of constructing of a p^* sampling design, which is nearest to the sampling design p_0 in the sense of minimizing the distance

$$D\left(p_0, p^*\right) = \sum_{s\in\mathfrak{I}_0} \frac{\left\{p^*(s) - p_0(s)\right\}^2}{p_0(s)}$$

where $\mathfrak{I}_0$ is the support of the sampling design p_0, which is the collection of all possible samples s of n distinct units with $p_0(s) > 0$, $\sum_{s\in\mathfrak{I}_0} p_0(s) = 1$. Here we assume that $\mathfrak{I}^*$, the support of the design p^* based on n distinct units, is a subset of $\mathfrak{I}_0$.

Now minimizing distance $D(p_0, p^*)$ subject to $\sum_{s\supset i} p^*(s) = \pi_i = np_i$, we get

$$p^*(s) = p_0(s)\sum_{i\in s}\lambda_i \tag{5.4.19}$$

where λ_i's are the solution of the equation

$$\mathbf{\Pi}_0\boldsymbol{\lambda} = \boldsymbol{\pi} \tag{5.4.20}$$

with

$$\mathbf{\Pi}_0 = \begin{pmatrix} \pi_1^0 & \pi_{12}^0 & \dots & \pi_{1N}^0 \\ \pi_{12}^0 & \pi_2^0 & \dots & \pi_{2N}^0 \\ \dots & \dots & \dots & .. \\ \pi_{N1}^0 & \pi_{N2}^0 & \dots & \pi_N^0 \end{pmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \dots \\ \pi_N \end{pmatrix}, \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_N \end{pmatrix},$$

$$\pi_i^0 = \sum_{s\supset i} p_0(s) \text{ and } \pi_{ij}^0 = \sum_{s\supset i,j} p_0(s) \text{ for } i\neq j$$

Eq. (5.4.19) provides the solution of sampling design p^* if $\sum_{i\in s}\lambda_i \geq 0$. Clearly, $p^*(s) \geq 0$ whenever all λ_i's are positive.

Example 5.4.3

Let p_0 be an SRSWOR design of size n with $p_0(s) = 1\Big/\binom{N}{n}$.

Here $\pi_i^0 = n\big/N = \alpha$, $\pi_{ij}^0 = n(n-1)\big/\{N(N-1)\} = \beta$,

$$\mathbf{\Pi}_0 = \begin{pmatrix} \alpha & \beta & \dots & \beta \\ \beta & \alpha & \dots & \beta \\ \dots & \dots & \dots & \dots \\ \beta & \beta & \dots & \alpha \end{pmatrix} \text{ and } \mathbf{\Pi}_0^{-1} = \begin{pmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \dots & \dots & \dots & \dots \\ b & b & \dots & a \end{pmatrix}$$

with $a = \dfrac{N}{n} - (n-1)b$ and $b = -\dfrac{(n-1)N}{n^2(N-n)}$.

From Eq. (5.4.20), we get $\lambda_i = \dfrac{N(N-1)}{n(N-n)}\pi_i - \dfrac{N(n-1)}{n(N-n)}$ for $i = 1,\dots, N$. Finally Eq. (5.4.19) yields

$$\begin{aligned} p^*(s) &= \left[\frac{N-1}{N-n}\sum_{i\in s}\pi_i - \frac{n(n-1)}{N-n}\right]\Big/\binom{N-1}{n-1} \\ &= \left[\frac{N-1}{N-n}n\sum_{i\in s}p_i - \frac{n(n-1)}{N-n}\right]\Big/\binom{N-1}{n-1} \end{aligned} \tag{5.4.21}$$

Eq. (5.4.21) is the selection probability of the LMS–IPPS sampling procedure. Hence the LMS–IPPS sampling scheme is the nearest to the SRSWOR sampling scheme of size n.

Remark 5.4.6

Arnab (2004a,b) provided the conditions of existence of the solution to Eq. (5.4.21).

5.5 PROBABILITY PROPORTIONAL TO AGGREGATE SIZE WITHOUT REPLACEMENT

Lahiri–Midzuno–Sen (1951, 1952, 1953) considered the probability proportional to aggregate size (PPAS) sampling scheme. Here, at the first draw, the ith unit is selected with probability p_i and the remaining $n-1$ units are selected by the SRSWOR method from those units that were not selected in the first draw. So, the probability of selection of an unordered sample, $s=(i_1,\ldots,i_k,\ldots,i_n)$ with $i_1<\cdots<i_k<\cdots<i_n$ is

$$p(s)=\sum_{i\in s}p_i\bigg/\binom{N-1}{n-1}=x_s/(M_1X)$$

where $x_s=\sum_{i\in s}x_i$, $X=\sum_{i=1}^{N}x_i$ and $M_1=\binom{N-1}{n-1}$.

Hence, the probability of selection of an unordered sample is proportional to the aggregate measure of size x_s.

The inclusion probabilities for the ith, and ith and jth units ($i\neq j$) are, respectively, given as follows:

$\pi_i=$ probability of selection of the ith unit at the first draw + probability that the ith unit is not selected at the first draw and it is selected in any of the remaining $(n-1)$ draws

$$=p_i+(1-p_i)\binom{N-1}{n-1}\bigg/\binom{N}{n}$$

$$=\frac{N-n}{N}p_i+\frac{n}{N}$$

and

$\pi_{ij}=$ probability of selection of the ith unit at the first draw and the jth unit in any of the remaining $(n-1)$ draws + probability of selection of the jth unit at the first draw and the ith unit in any of the remaining $(n-1)$ draws + probability that none of the ith and jth units are selected at the first draw and they are selected in the remaining $(n-1)$ draws

$$=p_i\frac{n}{N}+p_j\frac{n}{N}+(1-p_i-p_j)\frac{n(n-1)}{N(N-1)}$$

$$=\frac{n(N-n)}{N(N-1)}(p_i+p_j)+\frac{n(n-1)}{N(N-1)}$$

Theorem 5.5.1

(i) $\widehat{Y}_{lms} = \frac{y_s}{x_s} X$ is unbiased for the total Y

(ii) The variance of $\widehat{Y}_{lms}$ is

$$V_{lms} = V\left(\widehat{Y}_{lms}\right) = \sum_{i=1}^{N} y_i^2 (T_i - 1) + \sum_{i \neq}^{N} \sum_{j=1}^{N} y_i y_j (T_{ij} - 1)$$

where $y_s = \sum_{i \in s} y_i, \quad T_i = \frac{X}{M_1} \sum_s I_{si}/x_s = \frac{1}{M_1^2} \sum_s I_{si}/p(s),$

$T_{ij} = \frac{X}{M_1} \sum_s I_{si} I_{sj}/x_s = \frac{1}{M_1^2} \sum_s I_{si} I_{sj}/p(s),\ I_{si} = 1$ if $i \in s$, and $I_{si} = 0$ if $i \notin s$.

(iii) $V\left(\widehat{Y}_{lms}\right)$ can be unbiasedly estimated by any of the following:

(a) $\widehat{V}_{lms} = \sum_{i \in s} y_i^2 (T_i - 1)/\pi_i + \sum_{i \neq} \sum_{j \in s} y_i y_j (T_{ij} - 1)/\pi_{ij}$

or

(b) $\widehat{V}^{*}_{lms} = \left(\widehat{Y}_{lms}\right)^2 - \frac{X}{x_s} \left[\sum_{i \in s} y_i^2 + \frac{(N - 1)}{(n - 1)} \sum_{i \neq} \sum_{j \in s} y_i y_j \right]$

Proof

(i)
$$\begin{aligned} E(\widehat{Y}_{lms}) &= E\left(\frac{y_s}{x_s} X\right) \\ &= E\left(\frac{y_s}{M_1 p(s)}\right) \\ &= \sum_s \frac{y_s}{M_1} \\ &= \frac{1}{M_1} \sum_s \sum_{i=1}^{N} I_{si} y_i \\ &= \frac{1}{M_1} \sum_{i=1}^{N} y_i \sum_s I_{si} \\ &= Y \end{aligned}$$

$$\text{(ii)}\quad V\left(\widehat{Y}_{lms}\right) = E\left(\widehat{Y}_{lms}\right)^2 - Y^2$$

$$= E\left[\frac{1}{M_1}\frac{1}{p(s)}\sum_{i=1}^{N} I_{si}y_i\right]^2 - Y^2$$

$$= \sum_{s}\frac{1}{(M_1)^2}\frac{1}{p(s)}\left(\sum_{i=1}^{N} I_{si}y_i^2 + \sum_{i\neq}^{N}\sum_{j=1}^{N} I_{si}I_{sj}y_iy_j\right) - Y^2$$

$$= \sum_{i=1}^{N} y_i^2\left\{\frac{1}{(M_1)^2}\sum_{s} I_{si}\frac{1}{p(s)} - 1\right\}$$

$$+ \sum_{i\neq}^{N}\sum_{j=1}^{N} y_iy_j\left\{\frac{1}{(M_1)^2}\sum_{s} I_{si}I_{sj}\frac{1}{p(s)} - 1\right\}$$

$$= \sum_{i=1}^{N} y_i^2(T_i - 1) + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_iy_j(T_{ij} - 1)$$

$$\text{(iii)}\,\text{(a)}\,E\left(\widehat{V}_{lms}\right) = E\left[\sum_{i\in s} y_i^2(T_i - 1)/\pi_i + \sum_{i\neq}\sum_{j\in s} y_iy_j(T_{ij} - 1)/\pi_{ij}\right]$$

$$= \sum_{i=1}^{N} y_i^2(T_i - 1)E(I_{si})/\pi_i + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_iy_j(T_{ij} - 1)E(I_{si}I_{sj})/\pi_{ij}$$

$$= V\left(\widehat{Y}_{lms}\right)\ (\text{since } E(I_{si}) = \pi_i \ \text{ and } \ E(I_{si}I_{sj}) = \pi_{ij})$$

$$\text{(b)}\,E\left(\widehat{V}^{*}_{lms}\right) = \left(V\left(\widehat{Y}_{lms}\right) + Y^2\right) - \sum_{s}\left\{\frac{X}{x_s}\left(\sum_{i\in s} y_i^2 + \frac{(N-1)}{(n-1)}\sum_{i\neq}\sum_{j\in s} y_iy_j\right)\right\}p(s)$$

$$= \left(V\left(\widehat{Y}_{lms}\right) + Y^2\right) - \frac{1}{M_1}\sum_{s}\left\{\sum_{i\in s} y_i^2 + \frac{(N-1)}{(n-1)}\sum_{i\neq}\sum_{j\in s} y_iy_j\right\}$$

$$= \left(V\left(\widehat{Y}_{lms}\right) + Y^2\right) - \frac{1}{M_1}\left\{\sum_{i=1}^{N} y_i^2\sum_{s\supset i} + \frac{(N-1)}{(n-1)}\sum_{i\neq}^{N}\sum_{j\in s}^{N} y_iy_j\sum_{s\supset i,j}\right\}$$

$$= V\left(\widehat{Y}_{lms}\right)$$

Remark 5.5.1

The variance estimator $\widehat{V}_{lms}$ was proposed by Rao T.J. (1967). Although the estimator $\widehat{V}_{lms}$ seems to be elegant, it is seldom used in practice because

T_i and T_{ij}'s are very tedious to compute. Conversely, the estimator $\widehat{V}^*_{lms}$ is used very often in practice because it is very easy to compute. However, both the estimators can take negative values. Chaudhuri and Arnab (1981) studied nonnegativity properties of the unbiased estimators of $\widehat{V}_{lms}$ in detail.

5.6 RAO–HARTLEY–COCHRAN SAMPLING SCHEME

In the Rao–Hartley–Cochran (1962) sampling scheme, the population is divided at random into n disjoint groups $G_1,\ldots, G_n$. The number of units that belong to the jth group G_j is N_j, a preassigned number with $N = \sum_{j=1}^{n} N_j$. From each of the groups, one unit is selected independently with its probability proportional to its measure of size viz. if the unit i_j belongs to the jth group G_j, it is selected with probability $q_{i_j} = x_{i_j} \big/ \sum_{k\in G_j} x_k = p_{i_j}\big/P_j$, where $p_i = x_i/X$ and $P_j = \sum_{k\in G_j} p_k =$ sum of p_k's for the group G_j.

Suppose that the units $i_1, \ldots, i_j, \ldots, i_n$ are selected from the groups $G_1, \ldots, G_j, \ldots, G_n$, respectively. Then, we have the following theorem.

Theorem 5.6.1

(i) $\widehat{Y}_{rhc} = \sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j$ is an unbiased estimator for the population total Y

(ii) The variance of $\widehat{Y}_{rhc}$ is

$$V_{rhc} = V\left(\widehat{Y}_{rhc}\right) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2$$

(iii) An unbiased estimator for $V\left(\widehat{Y}_{rhc}\right)$ is

$$\widehat{V}\left(\widehat{Y}_{rhc}\right) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N^2 - \sum_{j=1}^{n} N_j^2} \sum_{j=1}^{n} P_j \left(\frac{y_{i_j}}{p_{i_j}} - \widehat{Y}_{rhc}\right)^2$$

Proof

Let E_G and V_G denote the unconditional expectation and the variance over G, respectively. The conditional expectation and variance for a given $G = (G_1,\ldots, G_j,\ldots, G_n)$ are denoted by $E(\cdot|G)$ and $V(\cdot|G)$ respectively. Then

$$\text{(i) } E\left(\widehat{Y}_{rhc}\right) = E_G\left\{ \sum_{j=1}^{n} E\left(\frac{y_{i_j}}{p_{i_j}} P_j \middle| G \right) \right\} = E_G\left(\sum_{j=1}^{n} Y_j \right) = Y$$

where $Y_j = \sum_{k\in G_j} y_k$

(ii) $$V\left(\widehat{Y}_{rhc}\right) = E_G\left\{\sum_{j=1}^{n} V\left(\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right)\right\} + V_G\left\{E\left(\sum_{j=1}^{n}\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right)\right\} \tag{5.6.1}$$

Now

$$V_G\left\{E\left(\sum_{j=1}^{n}\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right)\right\} = 0 \quad \left(\text{since } E\left(\sum_{j=1}^{n}\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right) = Y\right) \tag{5.6.2}$$

and

$$\begin{aligned}
E_G\left\{\sum_{j=1}^{n} V\left(\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right)\right\} &= E_G\left[\sum_{j=1}^{n}\left\{\sum_{k\in G_j}\frac{y_k^2}{p_k}P_j - \left(\sum_{k\in G_j} y_k\right)^2\right\}\right] \\
&= E_G\left[\sum_{j=1}^{n}\left\{\left(\sum_{k\in G_j} y_k^2 + \sum_{k\neq}\sum_{t\in G_j} p_t\frac{y_k^2}{p_k}\right)\right.\right. \\
&\qquad \left.\left. - \left(\sum_{k\in G_j} y_k^2 + \sum_{k\neq}\sum_{t\in G_j} y_k y_t\right)\right\}\right] \\
&= \sum_{j=1}^{n} E_G\left\{\sum_{k\neq}\sum_{t\in G_j}\left(p_t\frac{y_k^2}{p_k} - y_k y_t\right)\right\}
\end{aligned}$$

Now noting that G_j is a random sample of size N_j selected from the population of N units by SRSWOR we get

$$\begin{aligned}
E_G\sum_{k\neq}\sum_{t\in G_j}\left(p_t\frac{y_k^2}{p_k} - y_k y_t\right) &= \frac{N_j(N_j-1)}{N(N-1)}\sum_{t\neq}^{N}\sum_{k=1}^{N}\left(p_t\frac{y_k^2}{p_k} - y_k y_t\right) \\
&= \frac{N_j(N_j-1)}{N(N-1)}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2
\end{aligned}$$

and

$$E_G\left\{\sum_{j=1}^{n} V\left(\frac{y_{i_j}}{p_{i_j}}P_j\middle|G\right)\right\} = \frac{\sum_{j=1}^{n} N_j(N_j-1)}{N(N-1)}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 \tag{5.6.3}$$

Finally, substituting Eqs. (5.6.2) and (5.6.3) in Eq. (5.6.1), we prove (ii) of the theorem.

$$\begin{aligned}\text{(iii) } E\left[\sum_{j=1}^{n} P_j\left(\frac{y_{i_j}}{p_{i_j}} - \widehat{Y}_{rhc}\right)^2\right] &= E\left(\sum_{j=1}^{n} P_j\frac{y_{i_j}^2}{p_{i_j}^2}\right) - E\left(\widehat{Y}_{rhc}^2\right) \\ &= E\left(\sum_{j=1}^{n} P_j\frac{y_{i_j}^2}{p_{i_j}^2}\right) - V\left(\widehat{Y}_{rhc}\right) - Y^2\end{aligned} \tag{5.6.4}$$

Now

$$\begin{aligned}E\left(\sum_{j=1}^{n} P_j\frac{y_{i_j}^2}{p_{i_j}^2}\right) &= E_G\left[\sum_{j=1}^{n} E\left(P_j\frac{y_{i_j}^2}{p_{i_j}^2}\middle| G\right)\right] \\ &= \sum_{j=1}^{n} E_G\left(\sum_{i\in G_j}\frac{y_i^2}{p_i}\right) \\ &= \sum_{j=1}^{n}\frac{N_j}{N}\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i}\right) \\ &= \sum_{i=1}^{N}\frac{y_i^2}{p_i}\end{aligned} \tag{5.6.5}$$

Substituting Eq. (5.6.5) in Eq. (5.6.4), we get

$$\begin{aligned}E\left[\sum_{j=1}^{n} P_j\left(\frac{y_{i_j}}{p_{i_j}} - \widehat{Y}_{rhc}\right)^2\right] &= \sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 - V\left(\widehat{Y}_{rhc}\right) \\ &= \left(\frac{N(N-1)}{\sum\limits_j N_j^2 - N} - 1\right) V\left(\widehat{Y}_{rhc}\right)\end{aligned} \tag{5.6.6}$$

From Eq. (5.6.6), we get $E\left[\widehat{V}\left(\widehat{Y}_{rhc}\right)\right] = V\left(\widehat{Y}_{rhc}\right)$.

Remark 5.6.1

$\sum_{j=1}^{n} N_j^2$ attains a minimum when $N_j = N/n$. Hence if N/n is an integer, we should choose $N_j = N/n$ for every $j = 1,\ldots, n$ to minimize $V\left(\widehat{Y}_{rhc}\right)$ for a given value of n. In this case

$$V\left(\widehat{Y}_{rhc}\right) = \frac{N-n}{n(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 \tag{5.6.7}$$

and its unbiased estimator becomes

$$\widehat{V}\left(\widehat{Y}_{rhc}\right) = \frac{N-n}{N(n-1)} \sum_{j=1}^{n} P_j \left(\frac{y_{i_j}}{p_{i_j}} - \widehat{Y}_{rhc}\right)^2 \tag{5.6.8}$$

In case N/n is not an integer and $N = kn + r$ where k is an integer and $1 \le r < n$, we choose $N_1 = N_2 = \cdots = N_r = k + 1$ and $N_{r+1} = N_{r+2} = \cdots = N_n = k$ to minimize $V\left(\widehat{Y}_{rhc}\right)$.

Remark 5.6.2

The estimator $\widehat{Y}_{rhc}$ based on an RHC sampling scheme with N/n as an integer is more efficient than the HH estimator, $\widehat{Y}_{hh}$, based on a PPSWR sampling scheme because $V\left(\widehat{Y}_{rhc}\right) \le V\left(\widehat{Y}_{hh}\right)$. The relative efficiency of $\widehat{Y}_{rhc}$ with respect to $\widehat{Y}_{hh}$ is $R = \frac{V\left(\hat{Y}_{hh}\right)}{V\left(\hat{Y}_{rhc}\right)} 100 = \frac{N-1}{N-n} 100\%$, which increases with the sample size n.

Remark 5.6.3

Although the estimator $\widehat{Y}_{rhc}$ is simple and possesses elegant expressions for its variance and unbiased estimator of the variance, it suffers from a serious drawback that the estimator $\widehat{Y}_{rhc}$ is inadmissible. This is because $\widehat{Y}_{rhc}$ is an ordered estimator because it depends on the formation of the initial group $G = \left(G_1, \ldots, G_n\right)$. The technique of Rao-Blackwellization does not yield any elegant expression of $\widehat{Y}_{rhc}$. For example, let $N = 6$ and $n = 2$, then we can select a sample $s = (1, 2)$ by forming the following six distinct set of groups, each with probability 1/6.

Set 1: $G_1 = (1, 3, 4)$, $G_2 = (2, 5, 6)$; Set 2: $G_1 = (1, 3, 5)$, $G_2 = (2, 4, 6)$;
Set 3: $G_1 = (1, 3, 6)$, $G_2 = (2, 4, 5)$; Set 4: $G_1 = (1, 4, 5)$, $G_2 = (2, 3, 6)$;
Set 5: $G_1 = (1, 4, 6)$, $G_2 = (2, 3, 5)$; Set 6: $G_1 = (1, 5, 6)$, $G_2 = (2, 3, 4)$.

The RHC estimator based on the sets and the probabilities of selection of the respective samples are given in the following table with $P_{ijk} = p_i + p_j + p_k$:

Set (j)	1	2	3	4	5	6
$\widehat{Y}_{rhc}(j)$	$\frac{y_1}{p_1}P_{134} + \frac{y_2}{p_2}P_{256}$	$\frac{y_1}{p_1}P_{135} + \frac{y_2}{p_2}P_{246}$	$\frac{y_1}{p_1}P_{136} + \frac{y_2}{p_2}P_{245}$	$\frac{y_1}{p_1}P_{145} + \frac{y_2}{p_2}P_{236}$	$\frac{y_1}{p_1}P_{146} + \frac{y_2}{p_2}P_{235}$	$\frac{y_1}{p_1}P_{156} + \frac{y_2}{p_2}P_{234}$
Probability $P(j)$	$\frac{1}{6}\frac{p_1p_2}{P_{134}P_{256}}$	$\frac{1}{6}\frac{p_1p_2}{P_{135}P_{246}}$	$\frac{1}{6}\frac{p_1p_2}{P_{136}P_{245}}$	$\frac{1}{6}\frac{p_1p_2}{P_{145}P_{236}}$	$\frac{1}{6}\frac{p_1p_2}{P_{146}P_{235}}$	$\frac{1}{6}\frac{p_1p_2}{P_{156}P_{234}}$

The Rao-Blackwellization of $\widehat{Y}_{rhc}(j)$ gives a uniformly better estimator

$$\widehat{Y}^{*}_{rhc} = \frac{\sum_j \widehat{Y}(j)P(j)}{\sum_j P(j)}$$

which is obviously not a simple expression.

Example 5.6.1

Table 5.6.1 gives the yield of fish from nine ponds along with their areas. Select a sample of three ponds by the (i) RHC and (ii) PPAS method of sampling taking the area of the pond as measure of size. Estimate the average yield of fish and 90% confidence interval of the average yield for each of the samples separately.

(i) RHC method: For selection of a sample of size $n = 3$, we divide the population of size $N = 9$ at random into three groups each of size 3. For making groups, we select six one-digit random numbers from 1 to 9

Table 5.6.1

Ponds	1	2	3	4	5	6	7	8	9
Area (x, in acre)	1	3	1.5	2.5	1.5	2	2	1.5	2
Yield (y, in kg)	200	500	250	600	300	600	700	250	300

without replacement. The selected random numbers are 6, 1, 9, 3, 5, and 4. So, we allocate units 6, 1, and 9 to group G_1; units 3, 5, and 4 for G_2, and the remaining units 2, 7, and 8 to the group G_3. Now from each of the groups, we need to select one unit by the PPSWR method. For the selection of the sample, we prepare Table 5.6.2:

Table 5.6.2

G_1			G_2			G_3		
Unit	***x***	**Cumulative total × 10**	**Unit**	***x***	**Cumulative total × 10**	**Unit**	***x***	**Cumulative total × 10**
6	2	20	3	1.5	15	2	3	30
1	1	30	5	1.5	30	7	2	50
9	2	50	4	2.5	55	8	1.5	65

For selection of a unit from group G_1, we select one two-digit random number from 01 to 50; for G_2, one two-digit random number from 01 to 55; and for group G_3, one two-digit random number from 01 to 65. The selected random numbers are 19, 42, and 56. The random numbers select unit 6 from G_1, 4 from G_2, and 8 from G_3.

So, the selected sample is $s = (i_1 = 6,\ i_2 = 4,\ i_3 = 8)$. The estimator for the population mean $\overline{Y}$ is

$$\widehat{\overline{Y}}_{rhc} = \frac{\hat{Y}_{rhc}}{N} = \frac{1}{N}\sum_{j=1}^{n}\frac{y_{i_j}}{p_{i_j}}P_j = \frac{1}{9}\left[\frac{600}{2}\times 5 + \frac{600}{2.5}\times 5.5 + \frac{250}{1.5}\times 6.5\right] = 3903.333/9 = 433.704\ \text{kg}$$

Estimated standard error of $\widehat{\overline{Y}}_{rhc} = SE\left(\widehat{\overline{Y}}_{rhc}\right) = \sqrt{\widehat{V}\left(\widehat{\overline{Y}}_{rhc}\right)}$

$$= \frac{1}{N}\sqrt{\frac{N-n}{N(n-1)}\sum_{j=1}^{n}P_j\left(\frac{y_{i_j}}{p_{i_j}} - \widehat{\overline{Y}}_{rhc}\right)^2}$$

$$= \frac{1}{9}\sqrt{\frac{9-3}{9\times 2}\left\{\frac{5}{17}\left(\frac{600}{2}\times 17 - 3903.333\right)^2 + \frac{5.5}{17}\left(\frac{600}{2.5}\times 17 - 3903.333\right)^2 + \frac{6.5}{17}\left(\frac{250}{1.5}\times 17 - 3903.333\right)^2\right\}}$$

$$= 59.8091\ \text{kg}$$

90% confidence interval for $\overline{Y}$ is

$$\widehat{\overline{Y}}_{rhc} \pm t_{.05,2}SE\left(\widehat{\overline{Y}}_{rhc}\right) = 433.703 \pm 2.92\times 81.653$$

$$= (259.081\ \text{kg}, 608.325\ \text{kg}).$$

(ii) LMS method: Here the first unit should be selected with the probability proportional to the measure of size. For this selection, we prepare Table 5.6.3:

Table 5.6.3

Units	1	2	3	4	5	6	7	8	9
Measure of size (x)	1	3	1.5	2.5	1.5	2	2	1.5	2
Cumulative total (T) × 10	10	40	55	80	95	115	135	150	170

Here we select one three-digit random number from 001 to 170. The selected random number is 154 and the corresponding unit selected is 9. Now we need to select the remaining two units from 1 to 8 by SRSWOR. For this, we select two one-digit random numbers from 1 to 8. The selected random numbers are 5 and 2. So, the selected sample is $s = (2, 5, 9)$. The estimator for the population mean $\overline{Y}$ is

$$\widehat{\overline{Y}}_{lms} = \widehat{Y}_{lms}/N = \frac{y_s}{x_s}X/N = \left(\frac{500+300+300}{3+1.5+2}\times 17\right)\Big/9 = 2876.92/9$$
$$= 319.65\ \text{kg}$$

Estimated standard error of $\widehat{\overline{Y}}_{lms} = SE\left(\widehat{\overline{Y}}_{lms}\right) = \sqrt{\widehat{V}^{*}_{lms}}\Big/N.$
Now

$$\widehat{V}^{*}_{lms} = \left(\widehat{Y}_{lms}\right)^2 - \frac{X}{x_s}\left\{\sum_{i\in s} y_i^2 + \frac{(N-1)}{(n-1)}\sum_{i\neq}\sum_{j\in s} y_i y_j\right\}$$
$$= (2876.92)^2 - \frac{17}{6.5}(3550000)$$
$$= -1007929$$

Because the estimated variance $\widehat{V}^{*}_{lms} = -1007929$ is negative, we cannot find the confidence interval using $\widehat{V}^{*}_{lms}$ from this selected sample.

5.7 COMPARISON OF UNEQUAL (VARYING) PROBABILITY SAMPLING DESIGNS

VPS provides an efficient estimator for the population total Y if the measure of size variable x is well related to the study variable y. Various VPS designs are available in the literature. The most popular is the PPSWR sampling design. The HH estimator for the population total based on PPSWR sampling becomes efficient if the study variable is proportional to the measure of size variable. However, this estimator is inadmissible because it is based on the repetition of units in the samples. One can improve the HH estimator by employing the Rao-Blackwellization technique. The improved estimator, however, cannot be used in practice because of its complexity. In addition to this difficulty, it does not possess any simple expression of its unbiased variance estimator. Raj's estimator based on PPSWOR sampling is more efficient than the HH estimator based on PPSWR sampling of the same size. Murthy's estimator is obtained by applying Rao-Blackwellization techniques on Raj's estimator. Although Murthy's estimator is more efficient than Raj's estimator, it is seldom used in practice because the estimator itself and also its unbiased variance estimators are very tedious to compute. HTE based on an IPPS sampling design becomes very efficient if the study variable is proportional to the inclusion probability. The construction of an IPPS sampling design is not simple when sample size exceeds 2. The LMS—IPPS sampling scheme requires $p_i > (n-1)/\{n(N-1)\}$, which is seldom satisfied in practice. The PPS systematic sampling scheme is very easy to apply for any sample size, but the unbiased variance estimator is difficult to compute because of the complexity of the second-order inclusion probabilities. Sampford's (1967) IPPS sampling scheme is slightly more difficult to execute than the PPS systematic sampling scheme. The unbiased estimators of the variances of the estimators of the population total are equally complex. The variances of the HTE based on both the PPS systematic and Sampford sampling schemes are smaller than that of the HH estimator based on a PPSWR sampling scheme of the same sample size. The variance of the HTE based on Sampord's sampling is smaller than that of the PPS systematic sampling scheme. The LMS—PPAS sampling scheme is very easy to execute and possesses a smaller variance than that of the HH estimator of the same size, but it again suffers from the drawback that it does not always possess a nonnegative unbiased variance estimator. The RHC sample scheme is easy to execute and possesses elegant expressions for the population total, variance and nonnegative unbiased variance estimator, but it also suffers from the drawback of being inadmissible.

5.8 EXERCISES

5.8.1 Let a sample of size n be selected from a finite population of N units by the PPSWR method of sampling with p_i as the normed size measure for the ith unit. Find the expected number of distinct units in a sample and its variance. Obtain the expression of the Horvitz-Thompson estimator for the population total.

5.8.2 Show that the variance of the Hansen—Hurwitz estimator $\widehat{Y}_{hh}$ can be written as follows:

$$V\left(\widehat{Y}_{hh}\right)=\frac{1}{n}\sum_{i=1}^{N}p_i\left(\frac{y_i}{p_i}-Y\right)^2=\frac{1}{2n}\sum_{i\neq}^{N}\sum_{j=1}^{N}p_ip_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

5.8.3 Let $s=(i,j)$, $i<j$ be the set of distinct units in a sample of size 3 obtained by the PPSWR sampling method. List all the possible ordered samples s_o of size 3 that produces the unordered sample $s=(i,j)$. Obtain the expression of the unordered estimator that could be obtained by applying the Rao-Blackwellization technique on an HH estimator. Suggest an unbiased estimator of the variance of the unordered estimator.

5.8.4 (a) Let a sample size n be selected by the PPSWR sampling scheme. Derive the bias and mean square error of the sample mean, $\overline{y}_n$, based on n units with repetition, as an estimator for the population mean (Rao, 1966a,b).

(b) Let a sample of size n be selected by the RHC method with normed size measure p_i attached to the ith unit. Derive the expression of bias and mean square error of the estimator $T=N\sum_{i\in s}y_iP_i$ as an estimator of the population total Y where P_i is the sum of the p_j values for the group containing the ith unit (Rao, 1966a,b).

5.8.5 Let s be the unordered sample of size n obtained from ordered samples $s_1,\ldots,s_M$, each of size n and selected with probability $p(s_1),\ldots,$ $p(s_M)$, respectively. Let $\widehat{\theta}(s_1),\ldots,\widehat{\theta}(s_M)$ be the unbiased estimators of θ based on $s_1,\ldots,s_M$, respectively. Then show that

(i) $\widehat{\theta}=\sum_{j=1}^{M}\widehat{\theta}(s_j)p(s_j)/p(s)$ is unbiased for θ where $p(s)=\sum_{j=1}^{M}p(s_j)$

(ii) $V\left(\widehat{\theta}\right)\leq V\left\{\widehat{\theta}(s_j)\right\}$

(iii) $V\left\{\widehat{\theta}(s_j)\right\}-V\left(\widehat{\theta}\right)=\sum_{s}\sum_{i}\left\{\widehat{\theta}(s_i)-\widehat{\theta}\right\}^2p(s_i)$

(iv) An unbiased estimator of $V\left(\widehat{\theta}\right)$ is

$$\widehat{V}\left(\widehat{\theta}\right)=\widehat{\theta}^{2}-\left[\sum_{i}\left\{\widehat{\theta}(s_j)\right\}^{2}-\widehat{V}\left\{\widehat{\theta}(s_j)\right\}\right]p(s_j)/p(s)$$

where $\widehat{V}\left\{\widehat{\theta}(s_j)\right\}$ is an unbiased estimator for $V\left\{\widehat{\theta}(s_j)\right\}$ (Murthy, 1957).

5.8.6 Determine the expressions of the first two order inclusion probabilities of the PPSWOR sampling design of size $n = 2$. Derive the expressions of $\widehat{Y}_{ht}$ (Horvitz–Thompson estimator) and $\widehat{Y}_{RA}$ (the Raj's estimator) of the population total Y. Derive the expressions of variances of $\widehat{Y}_{RA}$ and $\widehat{Y}_{ht}$.

5.8.7 Let an ordered sample $s = (i, j)$ of size 2 be selected by the PPSWOR method.

(i) Show that the Raj's estimator and Murthy's estimator are $\widehat{Y}_{RA}=\frac{1}{2}\left\{(1+p_i)\frac{y_i}{p_i}+\left(1-p_i\right)\frac{y_j}{p_j}\right\}$ and

$\widehat{Y}_{M}=\frac{1}{2-p_i-p_j}\left\{\left(1-p_j\right)\frac{y_i}{p_i}+(1-p_i)\frac{y_j}{p_j}\right\}$, respectively.

(ii) Unbiased estimators of the variances of $\widehat{Y}_{RA}$ and $\widehat{Y}_{M}$ are, respectively $\widehat{V}\left(\widehat{Y}_{RA}\right)=\frac{1}{4}(1-p_i)^2\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$ and $\widehat{V}\left(\widehat{Y}_{M}\right)=$

$\frac{(1-p_i)\left(1-p_j\right)\left(1-p_i-p_j\right)}{\left(2-p_i-p_j\right)^2}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2.$

5.8.8 (a) Describe the IPPS sampling scheme and explain where it is useful.

(b) Consider selecting a sample of size 2 with replacement where the probability of selection of the ith unit at the first draw is $p_i(1) = p_i$ and at the second draw is $p_i(2) \propto p_i/(1-2p_i)$ for $i = 1,\ldots, N$. If the selected units are distinct, they are accepted as a sample. If they are the same, the entire selection is discarded and two units are selected again with the earlier method. The process is continued until two distinct units are selected. Show that for this sampling scheme (i) $\pi_i = 2p_i$ and (ii) $\pi_i\pi_j - \pi_{ij} \geq 0$ (Rao, 1965).

5.8.9 Let a sample s of size 2 be selected from a population U of size N following a without replacement scheme $d(2)$ of size 2 whose first- and second-order inclusion probabilities $\pi_i(2)$ and $\pi_{ij}(2)$ satisfy $\pi_i(2)\pi_j(2) - \pi_{ij}(2) \geq 0$ for $\forall i \neq j$. Let $d(n)$ be a sampling design of size n whose first two units are selected using $d(2)$ and the remaining $n-2$ units are selected by SRSWOR from the remaining $N-2$ units. Express the first- and second-order inclusion probabilities $\pi_i(n)$ and $\pi_{ij}(n)$ of $d(n)$ in terms of $\pi_i(2)$ and $\pi_{ij}(2)$. Show that $\pi_i(n)\pi_j(n) - \pi_{ij}(n) \geq 0$ (Seth, 1966).

5.8.10 (a) Describe the PPAS sampling scheme. Derive the expression of π_i and π_{ij}'s and show that $\pi_i\pi_j - \pi_{ij} \geq 0$ for $\forall i \neq j$.

(b) Derive the expression of bias and mean square error of the sample mean $\bar{y}(s) = \sum_{i \in s} y_i/n$ as an estimator of the population total mean $\overline{Y}$ when the sample is selected by the PPAS sampling scheme (Rao, 1966).

5.8.11 Table 5.8.1 gives the floor area and rent of 20 flats in Durban. Select a sample of five flats by probability proportional to area of the flat with replacement using (i) the cumulative total method and (ii) Lahiri's method. Estimate the average rent of the flat and its standard error from each of the selected samples.

Table 5.8.1

Serial no. of flat	1	2	3	4	5	6	7	8	9	10
Area (in m^2)	50	70	90	50	60	80	75	40	80	70
Rent (in \$)	3000	4000	5000	2500	3000	4500	4000	2500	5000	4000

Serial no. of the flat	11	12	13	14	15	16	17	18	19	20
Area (in m^2)	60	40	80	50	60	80	75	50	75	80
Rent (in \$)	3000	3000	4000	2000	3500	4000	4500	2000	5500	6000

5.8.12 Consider the data given in Exercise 5.8.11. Select a sample of size two flats by the PPSWOR method using the area of the flat as the measure of size variable. Estimate the average rent by using (i) Raj's estimator, (ii) Murthy's estimator, and (iii) Horvitz–Thompson estimator. Compute standard errors of each of the estimators and comment.

5.8.13 Table 5.8.2 gives the number of cows and production of milk of 10 farms.

Table 5.8.2

Serial no. of farm	1	2	3	4	5	6	7	8	9	10
No. of cows	10	20	8	10	15	20	15	5	25	12
Milk (in L)	100	240	40	150	200	250	175	75	300	130

Select a sample of size 5 by the PPSWOR method using the number of cows as the measure of size variable. Estimate the average yield of milk per farm using (a) Raj's and (b) Murthy's estimator. Estimate 95% confidence intervals for the average yield milk per farm based on the above estimators.

5.8.14 Consider the data given in Exercise 5.8.13. Select a sample of size 2 farms by using (i) Brewer, (ii) Durbin, and (iii) Hanurav's IPPS sampling design, using the number of cows as a measure of size variable. Estimate the average production of milk per farm based on the selected sample. Compare the efficiencies of your estimators.

5.8.15 Consider the data given in Exercise 5.8.13. Select a sample of five farms using the PPS systematic sampling procedure using the number of cows as the size variable. Estimate the average production of milk per farm and estimate its variance.

5.8.16 The weekly wages and expenditure on beer of 15 people are given in Table 5.8.3.

Table 5.8.3

Persons	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Weekly wages in $	200	250	300	400	450	500	450	400	300	250	200	250	300	400	450
Expenditure on beer in $	50	20	30	40	25	40	30	30	25	30	50	20	30	40	25

Select samples of four people by the (i) PPAS and (ii) RHC methods using weekly wages as the size variable. Estimate average expenditure on beer and 90% confidence of the average expenditure from each of the samples separately.