Chapter 7
Varying Probability Sampling

The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection. Under certain circumstances, more efficient estimators are obtained by assigning unequal probabilities of selection to the units in the population. This type of sampling is known as varying probability sampling scheme.

If Y is the variable under study and X is an auxiliary variable related to Y, then in the most commonly used varying probability scheme, the units are selected with probability proportional to the value of X, called as size. This is termed as probability proportional to a given measure of size (pps) sampling. If the sampling units vary considerably in size, then SRS does not takes into account the possible importance of the larger units in the population. A large unit, i.e., a unit with a large value of Y contributes more to the population total than the units with smaller values, so it is natural to expect that a selection scheme which assigns more probability of inclusion in a sample to the larger units than to the smaller units would provide more efficient estimators than the estimators which provide equal probability to all the units. This is accomplished through pps sampling.

Note that the "size" considered is the value of auxiliary variable X and not the value of study variable Y. For example, in an agriculture survey, the yield depends on the area under cultivation. So bigger areas are likely to have a larger population, and they will contribute more towards the population total, so the value of the area can be considered as the size of the auxiliary variable. Also, the cultivated area for a previous period can also be taken as the size while estimating the yield of the crop. Similarly, in an industrial survey, the number of workers in a factory can be considered as the measure of size when studying the industrial output from the respective factory.

Difference between the methods of SRS and varying probability scheme:

In SRS, the probability of drawing a specified unit at any given draw is the same. In varying probability scheme, the probability of drawing a specified unit differs from draw to draw.

It appears in pps sampling that such procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This will happen in the case of the sample mean as an estimator of the population mean where all the units are given equal weight. Instead of giving equal weights to all the units, if the sample observations are suitably weighted at the estimation stage by taking the probabilities of selection into account, then it is possible to obtain unbiased estimators.

In pps sampling, there are two possibilities to draw the sample, i.e., with replacement and without replacement.

Selection of units with replacement:

The probability of selection of a unit will not change, and the probability of selecting a specified unit is the same at any stage. There is no redistribution of the probabilities after a draw.

Selection of units without replacement:

The probability of selection of a unit will change at any stage, and the probabilities are redistributed after each draw.

PPS without replacement (WOR) is more complex than PPS with replacement (WR). We consider both the cases separately.

PPS sampling with replacement (WR):

First, we discuss the two methods to draw a sample with PPS and WR.

1. Cumulative total method:

The procedure of selecting a simple random sample of size n consists of

- associating the natural numbers from 1 to N units in the population and
- then selecting those n units whose serial numbers correspond to a set of n numbers where each number is less than or equal to N which is drawn from a random number table.

In the selection of a sample with varying probabilities, the procedure is to associate with each unit a set of consecutive natural numbers, the size of the set being proportional to the desired probability.

If $X_1, X_2, ..., X_N$ are the positive integers proportional to the probabilities assigned to the N units in the population, then a possible way to associate the cumulative totals of the units. Then the units are selected based on the values of cumulative totals. This is illustrated in the following table:

Units	Size	Cumulative Total		
1	X_1	$T_1 = X_1$	Select a random number R between 1 and T_N by using the random number table.	• If $T_{i-1} \le R \le T_i$, then i^{th} unit is selected with probability $\frac{X_i}{T_N}$, $i = 1,2,\ldots,N$. • Repeat the procedure n times to get a sample of size n.
2	X_2	$T_2 = X_1 + X_2$		
⋮	⋮	⋮		
$i-1$	X_{i-1}	$T_{i-1} = \sum_{j=1}^{i-1} X_j$		
i	X_i	$T_i = \sum_{j=1}^{i} X_j$		
⋮	⋮	⋮		
N	$X_N = \sum_{j=1}^{N} X_j$	$T_N = \sum_{j=1}^{N} X_j$		

In this case, the probability of selection of i^{th} unit is

$$P_i = \frac{T_i - T_{i-1}}{T_N} = \frac{X_i}{T_N}$$
$$\Rightarrow P_i \propto X_i.$$

Note that T_N is the population total which remains constant.

Drawback: This procedure involves writing down the successive cumulative totals. This is time-consuming and tedious if the number of units in the population is large.

This problem is overcome in Lahiri's method.

Lahiri's method:

Let $M = \underset{i=1,2,\ldots,N}{Max} X_i$, i.e., maximum of the sizes of N units in the population or some convenient number greater than M.

The sampling procedure has the following steps:

1. Select a pair of the random number (i, j) such that $1 \le i \le N,\ 1 \le j \le M$.
2. If $j \le X_i$, then i^{th} unit is selected otherwise rejected and another pair of random number is chosen.
3. To get a sample of size n, this procedure is repeated till n units are selected.

Now we see how this method ensures that the probabilities of selection of units are varying and are proportional to size.

Probability of selection of i^{th} unit at a trial depends on two possible outcomes

- either it is selected at the first draw
- or it is selected in the subsequent draws preceded by ineffective draws. Such probability is given by

$$P(1 \le i \le N)P(1 \le j \le M \mid i)$$
$$= \frac{1}{N}.\frac{X_i}{M} = P_i^*, \text{ say.}$$

Probability that no unit is selected at a trial $= \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{X_i}{M}\right)$

$$= \frac{1}{N}\left(N - \frac{N\bar{X}}{M}\right)$$
$$= 1 - \frac{\bar{X}}{M} = Q, \text{ say.}$$

The probability that unit i is selected (all other previous draws result in the non selection of unit i)

$$= P_i^* + QP_i^* + Q^2P_i^* + \ldots$$
$$= \frac{P_i^*}{1-Q}$$
$$= \frac{X_i / NM}{\bar{X} / M} = \frac{X_i}{N\bar{X}} = \frac{X_i}{X_{total}} \propto X_i.$$

Thus the probability of selection of unit i is proportional to the size X_i. So this method generates a pps sample.

Advantage:

1. It does not require writing down all cumulative totals for each unit.
2. Sizes of all the units need not be known beforehand. We need only some number greater than the maximum size and the sizes of those units which are selected by the choice of the first set of random numbers 1 to N for drawing sample under this scheme.

Disadvantage: It results in the wastage of time and efforts if units get rejected.

A draw is ineffective if one of the ineffective random numbers is selected.

The probability of rejection of a drawn number, i.e., probability that no unit is selected at a trial $= \frac{1}{N}.\sum_{i=1}^{N}\left(1-\frac{X_i}{M}\right) = \frac{1}{N}.\left(N - \frac{N\bar{X}}{M}\right) = 1 - \frac{\bar{X}}{M}.$

The expected numbers of draws required to draw one unit $= \frac{M}{\bar{X}}$.

This number is large if M is much larger than $\bar{X}$.

Example: Consider the following data set of 10 number of workers in the factory and its output. We illustrate the selection of units using the cumulative total method.

Factory no.	Number of workers (X) (in thousands)	Industrial production (in metric tons) (Y)	Cumulative total of sizes
1	2	30	$T_1 = 2$
2	5	60	$T_2 = 2+5 = 7$
3	10	12	$T_3 = 2+5+10 = 17$
4	4	6	$T_4 = 17+4 = 21$
5	7	8	$T_5 = 21+7 = 28$
6	2	13	$T_6 = 28+2 = 30$
7	3	4	$T_7 = 30+3 = 33$
8	14	17	$T_8 = 33+14 = 47$
9	11	13	$T_9 = 47+11 = 58$
10	6	8	$T_{10} = 58+6 = 64$

Selection of sample using cumulative total method:

1. First draw: - Draw a random number between 1 and 64.

- Suppose it is 23
- $T_4 < 23 < T_5$
- Unit Y is selected and $Y_5 = 8$ enters in the sample.

2. Second draw:

- Draw a random number between 1 and 64
- Suppose it is 38
- $T_7 < 38 < T_8$
- Unit 8 is selected and $Y_8 = 17$ enters the sample
- and so on.
- This procedure is repeated until the sample of required size is obtained.

Selection of sample using Lahiri's Method

In this case

$$M = \underset{i=1,2,...,10}{Max} X_i = 14$$

So we need to select a pair of random number (i, j) such that $1 \le i \le 10, 1 \le j \le 14$.

Following table shows the sample obtained by Lahiri's scheme:

Random no $1 \le i \le 10$	Random no $1 \le j \le 14$	Observation	Selection of unit
3	7	$j = 7 < X_3 = 10$	trial accepted (y_3)
8	13	$j = 13 < X_8 = 14$	trial accepted (y_8)
4	7	$j = 7 > X_4 = 4$	trial rejected
2	9	$j = 9 > X_2 = 5$	trial rejected
9	2	$j = 2 < X_9 = 11$	trial accepted (y_9)

and so on. Here (y_3, y_8, y_9) are selected into the sample.

Varying probability scheme with replacement: Estimation of population mean

Let

Y_i : Value of study variable for the i^{th} unit of the population, $i = 1, 2,\ldots,N$.

X_i : Known value of an auxiliary variable (size) for the i^{th} unit of the population.

P_i : Probability of selection of i^{th} unit in the population at any given draw and is proportional to size X_i.

$Z_i = \frac{Y_i}{NP_i}, i = 1, 2, \ldots, N.$

Consider the varying probability scheme and with replacement for a sample of size *n*. Let y_r be the value of r^{th} observation on study variable in the sample and p_r be its initial probability of selection. Define

$$z_r = \frac{y_r}{Np_r}, \; r = 1, 2, ..., n,$$

then $\bar{z} = \frac{1}{n}\sum_{r=1}^{n} z_i$ is an unbiased estimator of the population mean $\bar{Y}$, variance of $\bar{z}$ is $\frac{\sigma_z^2}{n}$ where $\sigma_z^2 = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{NP_i} - \bar{Y} \right)^2$ and an unbiased estimate of variance of $\bar{z}$ is $\frac{s_z^2}{n} = \frac{1}{n-1}\sum_{r=1}^{n} (z_r - \bar{z})^2$.

Proof:

Note that z_r can take any one of the N values out of $Z_1, Z_2, ..., Z_N$ with corresponding initial probabilities $P_1, P_2, ..., P_N$, respectively. So

$$E(z_r) = \sum_{i=1}^{N} Z_i P_i$$

$$= \sum_{i=1}^{N} \frac{Y_i}{NP_i} P_i$$

$$= \bar{Y}.$$

Thus

$$E(\bar{z}) = \frac{1}{n} \sum_{i=1}^{n} E(z_r)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \bar{Y}$$

$$= \bar{Y}.$$

So $\bar{z}$ is an unbiased estimator of the population mean $\bar{Y}$.

The variance of $\bar{z}$ is

$$Var(\bar{z}) = \frac{1}{n^2} Var\left(\sum_{r=1}^{n} z_r \right)$$

$$= \frac{1}{n^2} \sum_{r=1}^{n} Var(z_r) \qquad (z_r's \text{ are independent in WR case}).$$

Now

$$Var(z_r) = E\left[z_r - E(z_r) \right]^2$$

$$= E\left[z_r - \bar{Y} \right]^2$$

$$= \sum_{i=1}^{N} \left(Z_i - \bar{Y} \right)^2 P_i$$

$$= \sum_{i=1}^{N} \left(\frac{Y_i}{NP_i} - \bar{Y} \right)^2 P_i$$

$$= \sigma_z^2 \quad \text{(say)}.$$

Thus

$$Var(\bar{z}) = \frac{1}{n^2} \sum_{r=1}^{n} \sigma_z^2$$

$$= \frac{\sigma_z^2}{n}.$$

To show that $\frac{s_z^2}{n}$ is an unbiased estimator of the variance of $\bar{z}$, consider

$$(n-1)E(s_z^2) = E\left[\sum_{r=1}^{n}(z_r - \bar{z})^2\right]$$

$$= E\left[\sum_{r=1}^{n} z_r^2 - n\bar{z}^2\right]$$

$$= \left[\sum_{r=1}^{n} E(z_r^2) - nE(\bar{z}^2)\right]$$

$$= \sum_{r=1}^{n}\left[Var(z_r) + \{E(z_r)\}^2\right] - n\left[Var(\bar{z}) + \{E(\bar{z})\}^2\right]$$

$$= \sum_{r=1}^{n}\left(\sigma_z^2 + \bar{Y}^2\right) - n\left(\frac{\sigma_z^2}{n} + \bar{Y}^2\right) \quad \left(\text{using } Var(z_r) = \sum_{i=1}^{N}\left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 P_i = \sigma_z^2\right)$$

$$= (n-1)\sigma_z^2$$

$$E(s_z^2) = \sigma_z^2$$

or $$E\left(\frac{s_z^2}{n}\right) = \frac{\sigma_z^2}{n} = Var(\bar{z})$$

$$\Rightarrow \widehat{Var}(\bar{z}) = \frac{s_z^2}{n} = \frac{1}{n(n-1)}\left[\sum_{r=1}^{n}\left(\frac{y_r}{Np_r}\right)^2 - n\bar{z}^2\right].$$

Note: If $P_i = \frac{1}{N}$, then $\bar{z} = \bar{y}$,

$$Var(\bar{z}) = \frac{1}{n}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{N.\frac{1}{N}} - \bar{Y}\right)^2 = \frac{\sigma_y^2}{n}$$

which is the same as in the case of SRSWR.

Estimation of population total:

An estimate of population total is

$$\hat{Y}_{tot} = \frac{1}{n}\sum_{r=1}^{n}\left(\frac{y_r}{p_r}\right) = N\,\bar{z}..$$

Taking expectation, we get

$$\begin{aligned}
E(\hat{Y}_{tot}) &= \frac{1}{n}\sum_{r=1}^{n}\left[\frac{Y_1}{P_1}P_1 + \frac{Y_2}{P_2}P_2 + ... + \frac{Y_N}{P_N}P_N\right] \\
&= \frac{1}{n}\sum_{r=1}^{n}\left[\sum_{i=1}^{N} Y_i\right] \\
&= \frac{1}{n}\sum_{r=1}^{n} Y_{tot} \\
&= Y_{tot}.
\end{aligned}$$

Thus $\hat{Y}_{tot}$ is an unbiased estimator of population total. Its variance is

$$\begin{aligned}
Var(\hat{Y}_{tot}) &= N^2 Var(\bar{z}) \\
&= N^2\,\frac{1}{n}\sum_{i=1}^{N}\frac{1}{N^2}\left(\frac{Y_i}{P_i} - N\bar{Y}\right)^2 P_i \\
&= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 P_i \\
&= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - Y_{tot}^2\right].
\end{aligned}$$

An estimate of the variance

$$\widehat{Var}(\hat{Y}_{tot}) = N^2\,\frac{s_z^2}{n}.$$

Varying probability scheme without replacement

In varying probability scheme without replacement, when the initial probabilities of selection are unequal, then the probability of drawing a specified unit of the population at a given draw changes with the draw. Generally, the sampling WOR provides a more efficient estimator than sampling WR. The estimators for population mean and variance are more complicated. So this scheme is not commonly used in practice, especially in large scale sample surveys with small sampling fractions.

Let U_i : i^{th} unit,

P_i : Probability of selection of U_i at the first draw, $i = 1, 2, ..., N$

$\sum_{i=1}^{N} P_i = 1$

$P_{i(r)}$: Probability of selecting U_i at the r^{th} draw

$P_{i(1)} = P_i$.

Consider

$P_{i(2)} =$ Probability of selection of U_i at 2nd draw.

Such an event can occur in the following possible ways:

U_i is selected at 2nd draw when

- U_1 is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_2 is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_{i-1} is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_{i+1} is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_N is selected at 1^{st} draw and U_i is selected at 2^{nd} draw

So $P_{i(2)}$ can be expressed as

$$P_{i(2)} = P_1 \frac{P_i}{1-P_1} + P_2 \frac{P_i}{1-P_2} + ... + P_{i-1} \frac{P_i}{1-P_{i-1}} + P_{i+1} \frac{P_i}{1+P_{i+1}} + ... + P_N \frac{P_i}{1-P_N}$$

$$= \sum_{j(\neq i)=1}^{N} P_j \frac{P_i}{1-P_j}$$

$$= \sum_{j(\neq i)=1}^{N} P_j \frac{P_i}{1-P_j} + P_i \frac{P_i}{1-P_i} - P_i \frac{P_i}{1-P_i}$$

$$= \sum_{j=1}^{N} P_j \frac{P_i}{1-P_j} - P_i \frac{P_i}{1-P_i}$$

$$= P_i \left[\sum_{j=1}^{N} \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i} \right]$$

$P_{i(2)} \neq P_{i(1)}$ for all i unless $P_i = \frac{1}{N}$.

$P_{i(2)}$ will, in general, be different for each $i = 1,2,\ldots, N$. So $E\left(\frac{y_i}{p_i}\right)$ will change with successive draws.

This makes the varying probability scheme WOR more complex. Only $\frac{y_1}{Np_1}$ will provide an unbiased estimator of $\bar{Y}$. In general, $\frac{y_i}{Np_i}(i \neq 1)$ will not provide an unbiased estimator of $\bar{Y}$.

Ordered estimates

To overcome the difficulty of changing expectation with each draw, associate a new variate with each draw such that its expectation is equal to the population value of the variate under study. Such estimators take into account the order of the draw. They are called the ordered estimates. The order of the value obtained at previous draw will affect the unbiasedness of population mean.

We consider the ordered estimators proposed by Des Raj, first for the case of two draws and then generalize the result.

Des Raj ordered estimator

Case 1: Case of two draws:

Let y_1 and y_2 denote the values of units $U_{i(1)}$ and $U_{i(2)}$ drawn at the first and second draws respectively. Note that anyone out of the N units can be the first unit or second unit, so we use the notations $U_{i(1)}$ and $U_{i(2)}$ instead of U_1 and U_2. Also note that y_1 and y_2 are not the values of the first two units in the population. Further, let p_1 and p_2 denote the initial probabilities of selection of $U_{i(1)}$ and $U_{i(2)}$, respectively.

Consider the estimators

$$z_1 = \frac{y_1}{Np_1}$$

$$z_2 = \frac{1}{N}\left[y_1 + \frac{y_2}{p_2/(1-p_1)} \right]$$

$$= \frac{1}{N}\left[y_1 + y_2 \frac{(1-p_1)}{p_2} \right]$$

$$\bar{z} = \frac{z_1 + z_2}{2}.$$

Note that $\frac{p_2}{1-p_1}$ is the probability $P(U_{i(2)} \mid U_{i(1)})$.

Estimation of Population Mean:

First, we show that $\bar{z}$ is an unbiased estimator of $\bar{Y}$.

$$E(\bar{z}) = \bar{Y}.$$

Note that $\sum_{i=1}^{N} P_i = 1.$

Consider

$$E(z_1) = \frac{1}{N} E\left(\frac{y_1}{p_1}\right) \quad \left(\text{Note that } \frac{y_1}{p_1} \text{ can take any one of out of the } N \text{ values } \frac{Y_1}{P_1}, \frac{Y_2}{P_2}, ..., \frac{Y_N}{P_N}\right)$$

$$= \frac{1}{N}\left[\frac{Y_1}{P_1} P_1 + \frac{Y_2}{P_2} P_2 + ... + \frac{Y_N}{P_N} P_N\right]$$

$$= \bar{Y}$$

$$E(z_2) = \frac{1}{N} E\left[y_1 + y_2 \frac{(1-p_1)}{p_2}\right]$$

$$= \frac{1}{N}\left[E(y_1) + E_1\left\{E_2\left(y_2 \frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right)\right\}\right] \quad (\text{Using } E(Y) = E_X[E_Y(Y \mid X)].$$

where E_2 is the conditional expectation after fixing the unit $U_{i(1)}$ selected in the first draw.

Since $\frac{y_2}{p_2}$ can take any one of the $(N-1)$ values (except the value selected in the first draw) $\frac{Y_j}{P_j}$ with probability $\frac{P_j}{1-P_1}$, so

$$E_2\left[y_2 \frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right] = (1-P_1) E_2\left[\frac{y_2}{p_2}\middle| U_{i(1)}\right] = (1-P_1)\sum_j^*\left[\frac{Y_j}{P_j}.\frac{P_j}{1-P_1}\right].$$

where the summation is taken over all the values of Y except the value y_1 which is selected at the first draw. So

$$E_2\left[y_2 \frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right] = \sum_j^* Y_j = Y_{tot} - y_1.$$

Substituting it in $E(z_2)$, we have

$$E(z_2) = \frac{1}{N}\left[E(y_1) + E_1(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N}\left[E(y_1) + E(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N} E(Y_{tot}) = \frac{Y_{tot}}{N} = \bar{Y}.$$

Thus

$$E(\bar{z}) = \frac{E(z_1) + E(z_2)}{2}$$

$$= \frac{\bar{Y} + \bar{Y}}{2}$$

$$= \bar{Y}.$$

Variance:

The variance of $\bar{z}$ for the case of two draws is given as

$$Var(\bar{z}) = \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2\right)\left[\frac{1}{2N^2}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2\right] - \frac{1}{4N^2}\sum_{i=1}^{N} P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

Proof: Before starting the proof, we note the following property

$$\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i\left[\sum_{j=1}^{N} b_j - b_i\right]$$

which is used in the proof.

The variance of $\bar{z}$ is

$$Var(\bar{z}) = E(\bar{z}^2) - \left[E(\bar{z})\right]^2$$

$$= E\left[\frac{1}{2N}\left\{\frac{y_1}{p_1} + y_1 + \frac{y_2(1-p_1)}{p_2}\right\}\right]^2 - \bar{Y}^2$$

$$= \frac{1}{4N^2}E\left[\frac{y_1(1+p_1)}{p_1} + \frac{y_2(1-p_1)}{p_2}\right]^2 - \bar{Y}^2$$

$\downarrow$ $\downarrow$

nature of variable depends only on 1^{st} draw	nature of variable depends upon 1^{st} and 2^{nd} draw

$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i(1+P_i)}{P_i} + \frac{Y_j(1-P_i)}{P_j}\right\}^2 \frac{P_iP_j}{1-P_i}\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i^2}\frac{P_iP_j}{1-P_i} + \frac{Y_j^2(1-P_i)^2}{P_j^2}\frac{P_iP_j}{1-P_i} + 2Y_iY_j\frac{(1-P_i^2)}{P_iP_j}\frac{P_iP_j}{1-P_i}\right\}\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i}\frac{P_j}{1-P_i} + \frac{Y_j^2(1-P_i)^2}{P_j}\frac{P_i}{1-P_i} + 2Y_iY_j(1+P_i)\right\} - \bar{Y}^2\right].$$

Using the property

$\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i \left[\sum_{j=1}^{N} b_j - b_i\right]$, we can write

$$Var(\bar{z}) = \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2(1+P_i)^2}{P_i(1-P_i)}\left\{\sum_{j=1}^{N}P_j - P_i\right\} + \sum_{i=1}^{N}P_i(1-P_i)\left\{\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i}\right\} + 2\sum_{i=1}^{N}Y_i(1+P_i)(\sum_{j=1}^{N}Y_j - Y_i)\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i}(1+P_i^2+2P_i) + \sum_{i=1}^{N}P_i(1-P_i)\left\{\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i}\right\} + 2\sum_{i=1}^{N}Y_i(1+P_i)(\sum_{j=1}^{N}Y_j - Y_i)\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} + \sum_{i=1}^{N}Y_i^2P + 2\sum_{i=1}^{N}Y_i^2 + \sum_{i=1}^{N}P_i\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \sum_{i=1}^{N}Y_i^2 - \sum_{i=1}^{N}P_i^2\sum_{j=1}^{N}\frac{Y_j^2}{P_j}\right.$$

$$\left. + \sum_i P_iY_i^2 + 2\sum_{i=1}^{N}Y_i\sum_{j=1}^{N}Y_j - 2\sum_{i=1}^{N}Y_i^2P_i + 2\sum_{i=1}^{N}Y_iP_i\sum_{j=1}^{N}Y_j - 2\sum_{i=1}^{N}Y_i^2\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[2\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - \sum_{i=1}^{N}P_i^2\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \sum_{i=1}^{N}Y_i^2 + 2Y_{tot}^2 + 2Y_{tot}\sum_{i=1}^{N}Y_iP_i\right] - \bar{Y}^2$$

$$= 2\left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{4N^2}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - Y_{tot}^2 + Y_{tot}^2\right) - \frac{1}{4N^2}\left[\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i + 4N^2\bar{Y}^2\right]$$

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}(\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i - 2Y_{tot}^2 + 4Y_{tot}^2)$$

$$+\left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}Y_{tot}^2$$

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}(\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i + 2Y_{tot}^2 - 2Y_{tot}^2 + \sum_i P_i^2Y_{tot}^2)$$

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}\left(Y_i^2 - 2Y_{tot}Y_iP_i + P_i^2Y_{tot}^2\right)$$

$$= \frac{1}{2N^2}\left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

$$Var(\bar{z}) = \frac{1}{2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

↓ variance of WR case for $n = 2$

↓ reduction of variance in WR with varying probability

Estimation of $Var(\bar{z})$

$$\begin{aligned} Var(\bar{z}) &= E(\bar{z}^2) - (E(\bar{z}))^2 \\ &= E(\bar{z}^2) - \bar{Y}^2 \end{aligned}$$

Since

$$\begin{aligned} E(z_1 z_2) &= E\left[z_1 E(z_2 \mid u_1)\right] \\ &= E\left[z_1 \bar{Y}\right] \\ &= \bar{Y} E(z_1) \\ &= \bar{Y}^2. \end{aligned}$$

Consider

$$\begin{aligned} E\left[\bar{z}^2 - z_1 z_2\right] &= E(\bar{z}^2) - E(z_1 z_2) \\ &= E(\bar{z}^2) - \bar{Y}^2 \\ &= Var(\bar{z}) \end{aligned}$$

$\Rightarrow \widehat{Var}(\bar{z}) = \bar{z}^2 - z_1 z_2$ is an unbiased estimator of $Var(\bar{z})$

Alternative form

$$\begin{aligned} \widehat{Var}(\bar{z}) &= \bar{z}^2 - z_1 z_2 \\ &= \left(\frac{z_1 + z_2}{2}\right)^2 - z_1 z_2 \\ &= \frac{(z_1 - z_2)^2}{4} \\ &= \frac{1}{4}\left[\frac{y_1}{Np_1} - \frac{y_1}{N} - \frac{y_2}{N}\frac{1-p_1}{p_2}\right]^2 \\ &= \frac{1}{4N^2}\left[(1-p_1)\frac{y_1}{p_1} - \frac{y_2(1-p_1)}{p_2}\right]^2 \\ &= \frac{(1-p_1)^2}{4N^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2. \end{aligned}$$

Case 2: General Case

Let $(U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)})$ be the units selected in the order in which they are drawn in n draws where $U_{i(r)}$ denotes that the i^{th} unit is drawn at the r^{th} draw. Let $(y_1, y_2, .., y_r, ..., y_n)$ and $(p_1, p_2, ..., p_r, ..., p_n)$ be the values of study variable and corresponding initial probabilities of selection, respectively. Further, let $P_{i(1)}, P_{i(2)}, ..., P_{i(r)}, ..., P_{i(n)}$ be the initial probabilities of $U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)}$, respectively.

Further, let

$$z_1 = \frac{y_1}{Np_1}$$

$$z_r = \frac{1}{N}\left[y_1 + y_2 + ... + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right] \text{ for } r = 2,3,...,n.$$

Consider $\bar{z} = \frac{1}{n}\sum_{r=1}^{n} z_r$ as an estimator of population mean $\bar{Y}$.

We already have shown in case 1 that $E(z_1) = \bar{Y}$.

Now we consider $E(z_r), r = 2,3,...,n.$ We can write

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[z_r \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right]$$

where E_2 is the conditional expectation after fixing the units $U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)}$ drawn in the first (r - 1) draws.

Consider

$$E\left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right] = E_1 E_2 \left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right]$$

$$= E_1 \left[(1 - P_{i(1)} - P_{i(2)} \cdots - P_{i(r-1)}) E_2 \left(\frac{y_r}{p_r} \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right) \right].$$

Since conditionally $\frac{y_r}{p_r}$ can take any one of the N - (r -1) values $\frac{Y_j}{P_j}, j = 1,2,...,N$ with probabilities

$\frac{P_j}{1 - P_{i(1)} - P_{i(2)} \cdots - P_{i(r-1)}}$, so

$$E\left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right] = E_1 \left[(1 - P_{i(1)} - P_{i(2)} \cdots - P_{i(r-1)}) \sum_{j=1}^{N}{}^{*} \frac{Y_j}{P_j} \cdot \frac{P_j}{(1 - P_{i(1)} - P_{i(2)} \cdots - P_{i(r-1)})} \right]$$

$$= E_1 \left[\sum_{j=1}^{N}{}^{*} Y_j \right]$$

where $\sum_{j=1}^{N}{}^{*}$ denotes that the summation is taken over all the values of y except the y values selected in the first (r -1) draws

like as $\sum_{j=1(\neq i(1), i(2), ..., i(r-1))}^{N}$, i.e., except the values $y_1, y_2, ..., y_{r-1}$ which are selected in the first (r -1) draws.

Thus now we can express

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[y_1 + y_2 + ... + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \sum_{j=1}^{N} {}^{*} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \sum_{j=1(\neq i(1), i(2), ..., i(r-1))}^{N} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \left\{ Y_{tot} - \left(Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} \right) \right\} \right]$$

$$= \frac{1}{N} E_1 \left[Y_{tot} \right]$$

$$= \frac{Y_{tot}}{N}$$

$$= \bar{Y} \quad \text{for all} \quad r = 1, 2, ..., n.$$

Then

$$E(\bar{z}) = \frac{1}{n} \sum_{r=1}^{n} E(z_r)$$

$$= \frac{1}{n} \sum_{r=1}^{n} \bar{Y}$$

$$= \bar{Y}.$$

Thus $\bar{z}$ is an unbiased estimator of population mean $\bar{Y}$.

The expression for variance of $\bar{z}$ in general case is complex but its estimate is simple.

Estimate of variance:

$Var(\bar{z}) = E(\bar{z}^2) - \bar{Y}^2$.

Consider for $r < s$,

$$E(z_r z_s) = E\left[z_r E(z_s \mid U_1, U_2, ..., U_{s-1}) \right]$$

$$= E\left[z_r \bar{Y} \right]$$

$$= \bar{Y} E(z_r)$$

$$= \bar{Y}^2$$

because for $r < s$, z_r will not contribute

and similarly for $s < r, z_s$ will not contribute in the expectation.

Further, for $s < r$,

$$\begin{aligned} E(z_r z_s) &= E\left[z_s E(z_r \mid U_1, U_2, ..., U_{r-1})\right] \\ &= E\left[z_s \bar{Y}\right] \\ &= \bar{Y} E(z_s) \\ &= \bar{Y}^2. \end{aligned}$$

Consider

$$\begin{aligned} E\left[\frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s\right] &= \frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} E(z_r z_s) \\ &= \frac{1}{n(n-1)} n(n-1) \bar{Y}^2 \\ &= \bar{Y}^2. \end{aligned}$$

Substituting $\bar{Y}^2$ in $Var(\bar{z})$, we get

$$\begin{aligned} Var(\bar{z}) &= E(\bar{z}^2) - \bar{Y}^2 \\ &= E(\bar{z}^2) - E\left[\frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s\right] \\ \Rightarrow \widehat{Var(\bar{z})} &= \bar{z}^2 - \frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s \end{aligned}$$

Using $\left(\sum_{r=1}^{n} z_r\right)^2 = \sum_{r=1}^{n} z_r^2 + \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s$

$$\Rightarrow \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s = n^2 \bar{z}^2 - \sum_{r=1}^{n} z_r^2,$$

The expression of $\widehat{Var}(\bar{z})$ can be further simplified as

$$\begin{aligned} \widehat{Var}(\bar{z}) &= \bar{z}^2 - \frac{1}{n(n-1)}\left[n^2 \bar{z}^2 - \sum_{r=1}^{n} z_r^2\right] \\ &= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} z_r^2 - n\bar{z}^2\right] \\ &= \frac{1}{n(n-1)} \sum_{r=1}^{n} (z_r - \bar{z})^2. \end{aligned}$$

Unordered estimator:

In ordered estimator, the order in which the units are drawn is considered. Corresponding to any ordered estimator, there exist an unordered estimator which does not depend on the order in which the units are drawn and has smaller variance than the ordered estimator.

In case of sampling WOR from a population of size N, there are $\binom{N}{n}$ unordered sample(s) of size n. Corresponding to any unordered sample(s) of size n units, there are $n!$ ordered samples.
For example, for $n=2$ if the units are u_1 and u_2, then

- there are 2! ordered samples - (u_1,u_2) and (u_2,u_1)
- there is one unordered sample (u_1,u_2).

Moreover,

$$\begin{pmatrix}\text{Probability of unordered}\\ \text{sample } (u_1,u_2)\end{pmatrix}=\begin{pmatrix}\text{Probability of ordered}\\ \text{sample } (u_1,u_2)\end{pmatrix}+\begin{pmatrix}\text{Probability of ordered}\\ \text{sample } (u_2,u_1)\end{pmatrix}$$

For $n=3,$ there are three units u_1,u_2,u_3 and

-there are following 3! = 6 ordered samples:

$$(u_1,u_2,u_3),(u_1,u_3,u_2),(u_2,u_1,u_3),(u_2,u_3,u_1),(u_3,u_1,u_2),(u_3,u_2,u_1)$$

- there is one unordered sample (u_1,u_2,u_3).

Moreover,

Probability of unordered sample
= Sum of probability of ordered sample, i.e.

$$P(u_1,u_2,u_3)+P(u_1,u_3,u_2)+P(u_2,u_1,u_3)+P(u_2,u_3,u_1)+P(u_3,u_1,u_2)+P(u_3,u_2,u_1),$$

Let z_{si}, $s=1,2,..,\binom{N}{n}$, $i=1,2,...,n!(=M)$ be an estimator of population parameter θ based on ordered sample s_i. Consider a scheme of selection in which the probability of selecting the ordered sample (s_i) is p_{si}. The probability of getting the unordered sample(s) is the sum of the probabilities, i.e.,

$$p_s=\sum_{i=1}^{M}p_{si}.$$

For a population of size N with units denoted as 1, 2,…, N, the samples of size n are $n-$tuples. In the n^{th} draw, the sample space will consist of $N(N-1)...(N-n+1)$ unordered sample points.

$$p_{sio} = P\left[\text{selection of any ordered sample}\right] = \frac{1}{N(N-1)...(N-n+1)}$$

$$p_{siu} = P\left[\text{selection of any unordered sample}\right] = \frac{n!}{N(N-1)...(N-n+1)} = n!P\begin{bmatrix}\text{selection of any}\\ \text{ordered sample}\end{bmatrix}$$

then $p_s = \sum_{i=1}^{M(=n!)} p_{sio} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$.

Theorem: If $\hat{\theta}_0 = z_{si},\ s = 1,2,...,\binom{N}{n};\ i = 1,2,...,M(=n!)$ and $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p'_{si}$ are the ordered and unordered estimators of θ repectively, then

(i) $E(\hat{\theta}_u) = E(\hat{\theta}_0)$

(ii) $Var(\hat{\theta}_u) \le Var(\hat{\theta}_0)$

where z_{s_i} is a function of s_i^{th} ordered sample (hence a random variable) and p_{s_i} is the probability of selection of s_i^{th} ordered sample and $p'_{si} = \frac{p_{si}}{p_s}$.

Proof: Total number of ordered sample = $n!\binom{N}{n}$

$$(i)\ E(\hat{\theta}_0) = \sum_{s=1}^{\binom{N}{n}}\sum_{i=1}^{M} z_{si}p_{si}$$

$$E(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}}\left(\sum_{i=1}^{M} z_{si}p'_{si}\right)p_s$$

$$= \sum_{s}\left(\sum_{i} z_{si}\frac{p_{si}}{p_s}\right)p_s$$

$$= \sum_{s}\sum_{i} z_{si}p_{si}$$

$$= E(\hat{\theta}_0)$$

(ii) Since $\hat{\theta}_0 = z_{si}$, so $\hat{\theta}_0^2 = z_{si}^2$ with probability $p_{si},\ i = 1,2,...,M,\ s = 1,2,...,\binom{N}{n}$.

Similarly, $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p'_{si}$, so $\hat{\theta}_u^2 = \left(\sum_{i=1}^{M} z_{si}p'_{si}\right)^2$ with probability p_s

Consider

$$Var(\hat{\theta}_0) = E(\hat{\theta}_0^2) - \left[E(\hat{\theta}_0)\right]^2$$

$$= \sum_s \sum_i z_{si}^2 \, p_{si} - \left[E(\hat{\theta}_0)\right]^2$$

$$Var(\hat{\theta}_u) = E(\hat{\theta}_u^2) - \left[E(\hat{\theta}_u)\right]^2$$

$$= \sum_s \left(\sum_i z_{si} \, p'_{si} \right)^2 p_s - \left[E(\hat{\theta}_0)\right]^2$$

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i z_{si}^2 \, p_{si} - \sum_s \left(\sum_i z_{si} p'_{si} \right)^2 p_s$$

$$= \sum_s \sum_i z_{si}^2 \, p_{si} + \sum_s \left(\sum_i z_{si} p'_{si} \right)^2 p_s$$

$$- 2 \sum_s \left(\sum_i z_{si} \, p'_{si} \right) \left(\sum_i z_{si} \, p'_{si} \right) p_s$$

$$= \sum_s \left[\sum_i z_{si}^2 \, p_{si} + \left(\sum_i z_{si} \, p'_{si} \right)^2 \left(\sum_i p_{si} \right) - 2 \left(\sum_i z_{si} \, p'_{si} \right) \left(\sum_i z_{si} p'_{si} \right) p_s \right]$$

$$= \sum_s \left[\sum_i \left\{ z_{si}^2 \, p_{si} + \left(\sum_i z_{si} \, p'_{si} \right)^2 p_{si} - 2 \left(\sum_i z_{si} \, p'_{si} \right) z_{si} p_{si} \right\} \right]$$

$$= \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p'_{si})^2 \, p_{si} \right] \geq 0$$

$$\Rightarrow Var(\hat{\theta}_0) - Var(\hat{\theta}_u) \geq 0$$

or $Var(\hat{\theta}_u) \leq Var(\hat{\theta}_0)$

Estimate of $Var(\hat{\theta}_u)$

Since

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p'_{si})^2 \, p_{si} \right]$$

$$\widehat{Var}(\hat{\theta}_u) = \widehat{Var}(\hat{\theta}_0) - \sum_s \sum_i \left[\widehat{(z_{si} - \sum_i z_{si} p'_{si})^2 \, p_{si}} \right]$$

$$= \sum_i p'_{si} \, \widehat{Var}(\hat{\theta}_0) - \sum_i p'_{si} \widehat{(z_{si} - \sum_i z_{si} p'_{si})^2}.$$

Based on this result, now we use the ordered estimators to construct an unordered estimator. It follows from this theorem that the unordered estimator will be more efficient than the corresponding ordered estimators.

Murthy's unordered estimator corresponding to Des Raj's ordered estimator for the sample size 2

Suppose y_i and y_j are the values of units U_i and U_j selected in the first and second draws respectively with varying probability and WOR in a sample of size 2 and let p_i and p_j be the corresponding initial probabilities of selection. So now we have two ordered estimates corresponding to the ordered samples s_1^* and s_2^* as follows

$s_1^* = (y_i, y_j)$ with (U_i, U_j)

$s_2^* = (y_j, y_i)$ with (U_j, U_i)

which are given as

$$\bar{z}(s_1^*) = \frac{1}{2N}\left[(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_i + \frac{y_i}{p_i} + \frac{y_j(1-p_i)}{p_j}\right]$$

and

$$\bar{z}(s_2^*) = \frac{1}{2N}\left[(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_j + \frac{y_j}{p_j} + \frac{y_i(1-p_j)}{p_i}\right].$$

The probabilities corresponding to $\bar{z}(s_1^*)$ and $\bar{z}(s_2^*)$ are

$$p(s_1^*) = \frac{p_i p_j}{1-p_i}$$

$$p(s_2^*) = \frac{p_j p_i}{1-p_j}$$

$$p(s) = p(s_1^*) + p(s_2^*)$$

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

$$p'(s_1^*) = \frac{1-p_j}{2-p_i-p_j}$$

$$p'(s_2^*) = \frac{1-p_i}{2-p_i-p_j}.$$

Murthy's unordered estimate $\overline{z}(u)$ corresponding to the Des Raj's ordered estimate is given as

$$\overline{z}(u) = \overline{z}(s_1^*)\,p'(s_1) + \overline{z}(s_2^*)\,p'(s_2)$$

$$= \frac{\overline{z}(s_1^*)\,p(s_1^*) + \overline{z}(s_2^*)\,p(s_2^*)}{p(s_1^*) + p(s_2^*)}$$

$$= \frac{\left[\frac{1}{2N}\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}\left(\frac{p_i p_j}{1-p_i}\right)\right] + \left[\frac{1}{2N}\left\{(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right\}\left(\frac{p_j p_i}{1-p_j}\right)\right]}{\frac{p_i p_j}{1-p_i} + \frac{p_j p_i}{1-p_j}}$$

$$= \frac{\frac{1}{2N}\left[\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}(1-p_j) + \left\{(1+p_j)\frac{y_j}{p_i} + (1-p_j)\frac{y_i}{p_i}\right\}(1-p_i)\right]}{(1-p_j)+(1-p_i)}$$

$$= \frac{\frac{1}{2N}\left[(1-p_j)\frac{y_i}{p_i}\{(1+p_i)+(1-p_i)\} + (1-p_i)\frac{y_j}{p_j}\{(1-p_j)+(1+p_j)\}\right]}{2-p_i-p_j}$$

$$= \frac{(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}}{N(2-p_i-p_j)}.$$

Unbiasedness:

Note that y_i and p_i can take any one of the values out of $Y_1, Y_2, ..., Y_N$ and $P_1, P_2, ..., P_N$, respectively. Then y_j and p_j can take any one of the remaining values out of $Y_1, Y_2, ..., Y_N$ and $P_1, P_2, ..., P_N$, respectively, i.e., all the values except the values taken at the first draw. Now

$$E\left[\bar{z}(u)\right]=\frac{1}{N}\sum_{i<j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_iP_j}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}2\sum_{i<j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_jP_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}\sum_{i\neq j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_jP_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}\sum_{i\neq j}\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\left\{\frac{P_iP_j}{(1-P_i)(1-P_j)}\right\}\right]$$

$$=\frac{1}{2N}\sum_{i\neq j}\left[\frac{Y_iP_j}{1-P_i}+\frac{Y_jP_i}{1-P_j}\right]$$

Using result $\sum_{i\neq j=1}^{N}a_ib_j=\sum_{i=1}^{N}a_i\left\{\sum_{j=1}^{N}b_j-b_i\right\}$, we have

$$E\left[\bar{z}(u)\right]=\frac{1}{2N}\left[\left\{\sum_{i=1}^{N}\frac{Y_i}{1-P_i}(\sum_{j=1}^{N}P_j-P_i)\right\}+\left\{\sum_{j=1}^{N}\frac{Y_j}{1-P_j}(\sum_{i=1}^{N}P_i-P_j)\right\}\right]$$

$$=\frac{1}{2N}\left[\left\{\sum_{i=1}^{N}\frac{Y_i}{1-P_i}(1-P_i)\right\}+\sum_{j=1}^{N}\frac{Y_j}{1-P_j}(1-P_j)\right]$$

$$=\frac{1}{2N}\left\{\sum_{i=1}^{N}Y_i+\sum_{j=1}^{N}Y_j\right\}$$

$$=\frac{\bar{Y}+\bar{Y}}{2}$$

$$=\bar{Y}.$$

Variance: The variance of $\bar{z}(u)$ can be found as

$$Var\left[\bar{z}(u)\right] = \frac{1}{2}\sum_{i\neq j=1}^{N} \frac{(1-P_i-P_j)(1-P_i)(1-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2 \frac{P_iP_j(2-P_i-P_j)}{(1-P_i)(1-P_j)}$$

$$= \frac{1}{2}\sum_{i\neq j=1}^{N} \frac{P_iP_j(1-P_i-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2$$

Using the theorem that $Var(\hat{\theta}_u) \leq Var(\hat{\theta}_0)$ we get

$$Var\left[\bar{z}(u)\right] \leq Var\left[\bar{z}(s_1^*)\right]$$

$$\text{and } Var\left[\bar{z}(u)\right] \leq Var\left[\bar{z}(s_2^*)\right]$$

Unbiased estimator of $V\left[\bar{z}(u)\right]$

An unbiased estimator of $Var\left(\bar{z} \mid u\right)$ is

$$\widehat{Var}\left[\bar{z}(u)\right] = \frac{(1-p_i-p_j)(1-p_i)(1-p_j)}{N^2(2-p_i-p_j)^2}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2.$$

Horvitz Thompson (HT) estimate

The unordered estimates have limited applicability as they lack simplicity and the expressions for the estimators and their variance becomes unmanageable when sample size is even moderately large. The HT estimate is simpler than other estimators. Let N be the population size and $y_i, (i=1,2,...,N)$ be the value of characteristic understudy and a sample of size n is drawn by WOR using arbitrary probability of selection at each draw.

Thus prior to each succeeding draw, there is defined a new probability distribution for the units available at that draw. The probability distribution at each draw may or may not depend upon the initial probability at the first draw.

Define a random variable $\alpha_i (i=1,2,..,N)$ as

$$\alpha_i = \begin{cases} 1 \text{ if } Y_i \text{ is included in a sample '}s\text{' of size } n \\ 0 \text{ otherwise.} \end{cases}$$

Let $z_i = \frac{ny_i}{NE(\alpha_i)}$, $i=1...N$ assuming $E(\alpha_i) > 0$ for all i

where

$$E(\alpha_i) = 1.P(Y_i \in s) + 0.P(Y_i \notin s)$$

$$= \pi_i$$

is the probability of including the unit i in the sample and is called as **inclusion probability.**

The HT estimator of $\bar{Y}$ based on $y_1, y_2, ..., y_n$ is

$$\begin{aligned}\bar{z}_n = \hat{\bar{Y}}_{HT} &= \frac{1}{n}\sum_{i=1}^{n} z_i \\ &= \frac{1}{n}\sum_{i=1}^{N} \alpha_i z_i .\end{aligned}$$

Unbiasedness

$$\begin{aligned}E(\hat{\bar{Y}}_{HT}) &= \frac{1}{n}\sum_{i=1}^{N} E(z_i\alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} z_i E(\alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{NE(\alpha_i)} E(\alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{N} = \bar{Y}\end{aligned}$$

which shows that HT estimator is an unbiased estimator of the population mean.

Variance

$$\begin{aligned}V(\hat{\bar{Y}}_{HT}) &= V(\bar{z}_n) \\ &= E(\bar{z}_n^2) - \left[E(\bar{z}_n)\right]^2 \\ &= E(\bar{z}_n^2) - \bar{Y}^2 .\end{aligned}$$

Consider

$$\begin{aligned}E(\bar{z}_n^2) &= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i z_i\right]^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j z_i z_j\right] \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_i z_j E(\alpha_i\alpha_j)\right].\end{aligned}$$

If $S = \{s\}$ is the set of all possible samples and π_i is probability of selection of i^{th} unit in the sample s then

$$\begin{aligned}E(\alpha_i) &= 1\ P(y_i \in s) + 0.P(y_i \notin s) \\ &= 1.\pi_i + 0.(1-\pi_i) = \pi_i \\ E(\alpha_i^2) &= 1^2.P(y_i \in s) + 0^2.P(y_i \notin s) \\ &= \pi_i .\end{aligned}$$

So

$$E(\alpha_i) = E(\alpha_i^2)$$

$$E(\bar{z}_n^2) = \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 \pi_i + \sum_{i(\neq j)}^{N}\sum_{i=1}^{N} \pi_{ij} z_i z_j\right]$$

where π_{ij} is the probability of inclusion of i^{th} and j^{th} unit in the sample. This is called as **second-order inclusion probability**.

Now

$$\bar{Y}^2 = \left[E(\bar{z}_n)\right]^2$$

$$= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{N} \alpha_i z_i\right)\right]^2$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 \left[E(\alpha_i)\right]^2\right] + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_i z_j E(\alpha_i)E(\alpha_j)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 \pi_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \pi_i \pi_j z_i z_j\right].$$

Thus

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} \pi_i z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \pi_{ij} z_i z_j\right]$$

$$-\frac{1}{n^2}\left[\sum_{i=1}^{N} \pi_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \pi_i \pi_j z_i z_j\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} \pi_i (1-\pi_i) z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_i) z_i z_j\right]$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} \pi_i (1-\pi_i) \frac{n^2 y_i^2}{N^2 \pi_i^2} + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_i) \frac{n^2 y_i y_j}{N^2 \pi_i \pi_j}\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right) y_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij} - \pi_i \pi_i}{\pi_i \pi_j}\right) y_i y_j\right]$$

Estimate of variance

$$\hat{V}_1 = \widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{n} \frac{y_i^2 (1-\pi_i)}{\pi_i^2} + \sum_{i(\neq j)=1}^{n}\sum_{j=1}^{n}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right)\frac{y_i y_j}{\pi_i \pi_j}\right].$$

This is an unbiased estimator of variance.

Drawback: It does not reduces to zero when all $\frac{y_i}{\pi_i}$ are same, i.e., when $y_i \propto \pi_i$.

Consequently, this may assume negative values for some samples.

A more elegant expression for the variance of $\hat{\bar{y}}_{HT}$ has been obtained by Yates and Grundy.

Yates and Grundy form of variance

Since there are exactly n values of α_i which are 1 and $(N-n)$ values which are zero, so

$$\sum_{i=1}^{N} \alpha_i = n.$$

Taking expectation on both sides

$$\sum_{i=1}^{N} E(\alpha_i) = n.$$

Also

$$E\left(\sum_{i=1}^{N} \alpha_i\right)^2 = \sum_{i=1}^{N} E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N} \sum_{j=1}^{N} E(\alpha_i \alpha_j)$$

$$E(n)^2 = \sum_{i=1}^{N} E(\alpha_i) + \sum_{i(\neq j)=1}^{N} \sum_{j=1}^{N} E(\alpha_i \alpha_J) \quad (\text{using } E(\alpha_i) = E(\alpha_i^2))$$

$$n^2 = n + \sum_{i(\neq j)=1}^{N} \sum_{j=1}^{N} E(\alpha_i \alpha_J)$$

$$\sum_{i(\neq j)=1}^{N} \sum_{j=1}^{N} E(\alpha_i \alpha_J) = n(n-1)$$

Thus $E(\alpha_i \alpha_j) = P(\alpha_i = 1, \alpha_j = 1)$

$$= P(\alpha_i = 1) P(\alpha_j = 1 \,|\, \alpha_i = 1)$$

$$= E(\alpha_i) E(\alpha_j \,|\, \alpha_i = 1)$$

Therefore

$$\sum_{j(\neq i)=1}^{N} \left[E(\alpha_i \alpha_j) - E(\alpha_i) E(\alpha_j) \right]$$

$$= \sum_{j(\neq i)=1}^{N} \left[E(\alpha_i) E(\alpha_j \,|\, \alpha_i = 1) - E(\alpha_i) E(\alpha_j) \right]$$

$$= E(\alpha_i) \sum_{j(\neq i)=1}^{N} \left[E(\alpha_j \,|\, \alpha_i = 1) - E(\alpha_j) \right]$$

$$= E(\alpha_i) \left[(n-1) - (n - E(\alpha_i)) \right]$$

$$= -E(\alpha_i) \left[1 - E(\alpha_i) \right]$$

$$= -\pi_i (1 - \pi_i) \qquad (1)$$

Similarly

$$\sum_{i(\neq j)=1}^{N} \left[E(\alpha_i \alpha_j) - E(\alpha_i) E(\alpha_j) \right] = -\pi_j (1 - \pi_j). \qquad (2)$$

We had earlier derived the variance of HT estimator as

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)z_iz_j\right]$$

Using (1) and (2) in this expression, we get

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2 + \sum_{j=1}^{N}\pi_j(1-\pi_j)z_j^2 - 2\sum_{i\neq j=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})z_iz_j\right]$$

$$= \frac{1}{2n^2}\left[-\sum_{i=1}^{N}\left\{\sum_{j(\neq i)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_i^2\right.$$

$$\left.-\sum_{j=1}^{N}\left\{\sum_{i(\neq j)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_j^2 - 2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{n}\left\{E(\alpha_i)E(\alpha_j)-E(\alpha_i\alpha_j)\right\}z_iz_j\right]$$

$$= \frac{1}{2n^2}\left[\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_j^2 + 2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_iz_j\right]\right.$$

$$= \frac{1}{2n^2}\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})(z_i^2+z_j^2-2z_iz_j)\right].$$

The expression for π_i and π_{ij} can be written for any given sample size.

For example, for $n=2$, assume that at the second draw, the probability of selecting a unit from the units available is proportional to the probability of selecting it at the first draw. Since

$E(\alpha_i) =$ Probability of selecting Y_i in a sample of two

$= P_{i1} + P_{i2}$

where P_{ir} is the probability of selecting Y_i at r^{th} draw $(r=1,2)$. If P_i is the probability of selecting the i^{th} unit at first draw $(i=1,2,...,N)$ then we had earlier derived that

$$P_{i1} = P_i$$

$$P_{i2} = P\begin{bmatrix} y_i \text{ is not selected} \\ \text{at } 1^{st} \text{ draw} \end{bmatrix} P\begin{bmatrix} y_i \text{ is selected at } 2^{nd} \text{ draw}| \\ y_i \text{ is not selected at } 1^{st} \text{ draw} \end{bmatrix}$$

$$= \sum_{j(\neq i)=1}^{N}\frac{P_jP_i}{1-P_j}$$

$$= \left[\sum_{j=1}^{N}\frac{P_j}{1-P_j} - \frac{P_i}{1-P_i}\right]P_i.$$

So

$$E(\alpha_i) = P_i\left[\sum_{j=1}^{N}\frac{P_j}{1-P_j} - \frac{P_i}{1-P_i}\right] + P_i$$

Again

$$E(\alpha_i\alpha_j) = \text{Probability of including both } y_i \text{ and } y_j \text{ in a sample of size two}$$

$$= P_{i1}\, P_{j2|i} + P_{j1}P_{i2|j}$$

$$= P_i \frac{P_j}{1-P_i} + P_j \frac{P_i}{1-P_j}$$

$$= P_iP_j\left[\frac{1}{1-P_i} + \frac{1}{1-P_j}\right] + P_i.$$

Estimate of Variance

The estimate of variance is given by

$$\widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2}\sum_{i(\neq j)}^{n}\sum_{j=1}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}(z_i - z_j)^2.$$

Midzuno system of sampling:

Under this system of selection of probabilities, the unit in the first draw is selected with unequal probabilities of selection (i.e., pps) and remaining all the units are selected with SRSWOR at all subsequent draws.

Under this system

$$E(\alpha_i) = \pi_i = P\ (\text{unit } i\ (U_i)\ \text{is included in the sample})$$

$$= P(U_i \text{ is included in } 1^{st} \text{ draw}) + P(U_i \text{ is included in any other draw})$$

$$= P_i + \begin{pmatrix} \text{Probability that } U_i \text{ is not selected at the first draw and} \\ \text{is selected at any of subsequent } (n-1) \text{ draws} \end{pmatrix}$$

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

$$= \frac{N-n}{N-1}P_i + \frac{n-1}{N-1}.$$

Similarly,

$$E(\alpha_i\alpha_j) = \text{Probability that both the units } U_i \text{ and } U_j \text{ are in the sample}$$

$$= \begin{pmatrix} \text{Probability that } U_i \text{ is selected at the first draw and} \\ U_j \text{ is selected at any of the subsequent draws } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that } U_j \text{ is selected at the first draw and} \\ U_i \text{ is selected at any of the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that neither } U_i \text{ nor } U_j \text{ is selected at the first draw but} \\ \text{both of them are selected during the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$= P_i \frac{n-1}{N-1} + P_j \frac{n-1}{N-1} + (1-P_i-P_j)\frac{(n-1)(n-2)}{(N-1)(N-2)}$$

$$= \frac{(n-1)}{(N-1)}\left[\frac{N-n}{N-2}(P_i+P_j)+\frac{n-2}{N-2}\right]$$

$$\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_i+P_j)+\frac{n-2}{N-2}\right].$$

Similarly,

$$E(\alpha_i\alpha_j\alpha_k) = \pi_{ijk} = \text{Probability of including } U_i, U_j \text{ and } U_k \text{ in the sample}$$

$$= \frac{(n-1)(n-2)}{(N-1)(N-2)}\left[\frac{N-n}{N-3}(P_i+P_j+P_k)+\frac{n-3}{N-3}\right].$$

By an extension of this argument, if $U_i, U_j, ..., U_r$ are the r units in the sample of size $n(r<n)$, the probability of including these r units in the sample is

$$E(\alpha_i\alpha_j...\alpha_r) = \pi_{ij...r} = \frac{(n-1)(n-2)...(n-r+1)}{(N-1)(N-2)...(N-r+1)}\left[\frac{N-n}{N-r}(P_i+P_j+...+P_r)+\frac{n-r}{N-r}\right]$$

Similarly, if $U_1, U_2, ..., U_q$ be the n units, the probability of including these units in the sample is

$$E(\alpha_i\alpha_j...\alpha_q) = \pi_{ij...q} = \frac{(n-1)(n-2)...1}{(N-1)(N-2)...(N-n+1)}(P_i+P_j+...+P_q)$$

$$= \frac{1}{\binom{N-1}{n-1}}(P_i+P_j+...+P_q)$$

which is obtained by substituting $r=n$.

Thus if $P_i's$ are proportional to some measure of size of units in the population then the probability of selecting a specified sample is proportional to the total measure of the size of units included in the sample.

Substituting these $\pi_i, \pi_{ij}, \pi_{ijk}$ etc. in the HT estimator, we can obtain the estimator of population's mean and variance. In particular, an unbiased estimate of variance of HT estimator given by

$$\widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2} \sum_{i \neq j=1}^{n} \sum_{j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} (z_i - z_j)^2$$

where

$$\pi_i \pi_j - \pi_{ij} = \frac{N-n}{(N-1)^2} \left[(N-n) P_i P_j + \frac{n-1}{N-2} (1 - P_i - P_j) \right].$$

The main advantage of this method of sampling is that it is possible to compute a set of revised probabilities of selection such that the inclusion probabilities resulting from the revised probabilities are proportional to the initial probabilities of selection. It is desirable to do so since the initial probabilities can be chosen proportional to some measure of size.