UNIT - II

Varying Probability Sampling:

Def: When units vary in their sizes and the variate under study is highly correlated to size (Prob. proportion to size sampling).

When samples from different sizes are grouped and used and sampling is taken with the sample probability the chances of selecting a member from a large group or less than selecting a member from a smaller group. this is known as probability proportional to size sampling. {when n is in different sizes}.

For eg:

If one sample had 20,000 members, the prob of a member being selected would be 1/20,000 or 0.005%. If another sample had 10,000 members, the chance of a member being selected would be 1/10,000 or 0.01%.

Difference between SRS and Varing PS:

In simple random sampling the probability of drawing a specified unit

at any given draw is the same. In varying prob. scheme the probability of drawing a specified unit differ from draw to draw.

Procedures of Selecting a Sample:

The procedure of selecting a sample consist in associating with each unit a number or set of number equal to its size. The selection of units is done corresponding to a no. chosen at random from the totality of numbers associated. There are two methods of selection.

1. Cumulative total method
2. Lahiri's method

* Let the size of the i^{th} unit be $(i=1,2,3\ldots,N)$ the total being $X = \sum_{i}^{N} X_i$.
* We associate the number 1 to X_1 with the first unit, the numbers (X_1+1) to (X_1+X_2) with the second unit, and so on.
* A number is chosen at random from 1 to X and the unit with which this number is associated is selected.

~~with which this~~ n

* The i^{th} unit in the population is being selected with a prob. proportional to x_i.
* If a sample of size n is required, the procedure is repeated n times with replacement of the units selected.
* This procedure of selection is known as the cumulative total method for the method needs cumulation of the unit sizes.
* The main difficulty in this procedure is the compulsion to complete successive cumulative totals, which becomes time consuming and costly when the population size is large.

Example:

A village has 10 holdings consisting of 50, 30, 45, 25, 40, 26, 24, 35, 28 and 27 fields, respectively. Select a sample of four holdings with the replacement method and with probability proportional to the number of fields in the holdings.

The first step in the selection of holdings is to form cumulative totals as per following eg:

S. No. of holdings	Size (x_i)	Cumulative size	Numbers associated
1	50	50	1-50
2	30	80	51-80
3	45	125	81-125
4	25	150	126-150
5	40	190	151-190
6	26	216	191-216
7	44	260	217-260
8	35	295	261-295
9	28	323	296-323
10	27	350	324-350.

To select a holding, a random number not exceeding 350 is drawn with the help of a random number table. Suppose the random number thus selected is 272.

It can be seen from the cumulative totals that the number is associated with the group 261-295, i.e, the 8th holding is selected corresponding to the random number 272.

Similarly, we select three more random numbers. Suppose these numbers are 346, 165 and

Lahiri's method:

Lahiri (1951) suggested an alternative procedure in which cumulations are avoid completely.

It consists in selecting a number at random between 1 and N and noting down the unit with the corresponding serial number, provisionally. Another random number is then chosen between 1 and M, where M is the maximum size of the N units of the population.

If the second random number is smaller than the size of the unit provisionally selected, the unit is selected into the sample. If not, the entire procedure is repeated until a unit is finally selected. For selecting a sample of n units, the procedure is to be repeated until n units are selected.

eg: Select a sample of four holding in the various example by Lahiri's method of PPS, wr.

In this case, N = 10, M = 50.

First we have to select a random number which is not greater than 10 and a second random ~~table~~ number which is not greater than 50. Referring to the random number table, the pair is (10, 13). Hence, the 10th unit is selected in the sample. Similarly, choosing other pairs, we can have (4, 26), (5, 35), (7, 26).

The pair (4, 26) is rejected as 26 is greater than the size value (25) and so another pair is drawn which turns out to be (8, 16).

Hence, the sample will consist of the holdings with serial numbers 10, 5, 7 and 8.

PROCEDURES OF SELECTION OF A PPS SAMPLE WITHOUT REPLACEMENT:

$P_i = Y_i / X.$

There are several procedures for selecting samples with unequal prob. A brief discussion of some important procedures are as follows.

General selection procedure:

A generalization of the sampling scheme, ~~where~~ as there would be to select a pps sample of size unity and remove the selected unit from the population.

From the remaining units, another PPS sample of size one is taken as before and the selected unit removed from the population. This process is repeated until n selections.

If n units are selected one by one, with probability proportional to size measure x_j at each draw, without replacing the units selected in the previous draws. The probability of selecting a first draw for the j^{th} unit is

$$P_j = X_j / X, \quad j = 1, \ldots N.$$

Where $X = \sum_{j}^{N} X_j$.

Similarly, the probability that the i^{th} unit is selected at the second draw, when the j^{th} unit has been selected at the first draw, is given by

$$P_{i/j} = P_i / (1 - P_j) \quad i \neq j, \text{ and so on.}$$

This set up of sampling comprises an ordered set of sample values $(y_1, y_2 \ldots y_n)$ with probabilities $(P_1, P_2, \ldots P_n)$.

Narain's Scheme of Sample Selection:

The scheme consists of constructing revised prob. of selection p_i' $(i = 1, 2 \ldots, N)$ such that the inclusion probabilities π_i are proportional to the orginal probabilities of selection $p_i (i = 1, 2, \ldots N)$. and sampling is done without replacement.

The Inclusion Probabilities π_i are given by

$$\pi_i = np_i, \quad i = 1, 2, \ldots N.$$

The Probabilities of Selection and Inclusion at the Second and Subsequent draws are proportional to revised Probabilities p_i' on lines similar to those at the first draw. Let us consider the simple case $N = 4$ and $n = 2$.

The Problem is to Evaluate π_{ij}, given p_i $(i, j = 1, 2, 3, 4)$. The relationship $\sum_{j \neq i} \pi_{ij} = \pi_i$ Provides a System of 4 linear equations with six unknowns. The Problem is to choose the values of two arbitrary parameters with the restriction that all the above values are positive. The Computations become tedious for n greater than 2.

Example for General Selection Procedure:

In a village, there are 8 Orchards with 50, 30, 25, 40, 26, 44, 20 and 35 trees, respectively,

Select a Sample of 2 orchards with Probability Proportional to the number of tree in the orchard and without replacement.

(i) For Selecting a unit by Prob. Proportional to the Number of trees, using Lahiri's method of Selection, Consider the following arrangement:

Orchard number :	1	2	3	4	5	6	7	8
Number of tree :	50	30	25	40	26	44	20	35

Selecting a pair of r. numbers $(i,j) / (i \leq 8, j \leq 50)$ using the random number table, we get the pair $(5,17)$. Since the number (~~II~~) of the trees (x_i) for orchard number 5 is greater than the second number (17) of the selected random pair, the 5th orchard is selected in the sample.

(ii) For selecting the second unit by probability proportional to the number of trees in the orchard, we prepare the following arrangement once again after deleting the 5th orchard:

Orchard number ⇒ . Same as above.
Number of trees ⇒

As in (i) a pair of random numbers (i,j) $(i \leq 7, j \leq 50)$ has to be selected, using the random number table. Referring to the table of random numbers, the pair selected is $(6,18)$. As the size of the 6th unit in the above arrangement is greater than the second number of the random pair selected, orchard 6 is selected into the sample. Thus, the sample selected consists of the units at serial numbers 5 and 7 of the original list with the number of trees being 16 and 20, respectively.

Sen - Midzuno Method:

A simple procedure of selecting a sample was suggest by midzuno and independently by sen (1952), which consists in selecting the first unit with PPs and the remaining $(n-1)$ units from $(N-1)$ units of the population by simple random sampling.

For this selection procedure, the inclusion probabilities for individual and pairwise units are given by.

$$\pi_i = p_i + (1-p_i)\frac{n-1}{N-1} \quad \text{for } i=1,2,\ldots,N.$$

$$= \frac{N-n}{N-1}\,p_i + \frac{n-1}{N-1}$$

and $$\pi_{ij} = p_i \frac{n-1}{N-1} + p_j \cdot \frac{n-1}{N-1} + (1-p_i-p_j)$$

$$\frac{(n-1)(n-2)}{(N-1)(N-2)} \quad \text{for } i \neq j = 1,2,\ldots,N.$$

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

By extension of the above argument, we can have $y_1, y_2, \ldots, y_n$, a sample of n units.

The probability of including these n units in the sample is given by

$$\pi_{ij\ldots q} = \frac{1}{\binom{N-1}{n-1}} (p_i + p_j + \ldots + p_q).$$

ESTIMATION IN PPS SAMPLING WITHOUT REPLACEMENT: Total and its Sampling Variance:

ORDERED ESTIMATORS:-

Those estimators which are based on the order of units selected in the sample and do not require calculations of inclusion probabilities.

Des (1951) and Des Raj (1956) have proposed such estimators these make use of conditional probabilities without calculating π_i and π_{ij} which are generally difficult to compute for may sampling schemes.

We shall consider the estimator proposed by Des Raj, for the case when $n = 2$.

Desraj's order estimator:

Consider, Case (i) $n=2$ and the result is generalized.

* Let the initial probability of units U_i and P_i; $i = 1, 2 \ldots N$ where $P_i = X_i / X$.

* The first draw is made with prob. P_i, the second draw is taken with conditional prob. $P_j/(1-P_i)$

* Suppose y_1 and y_2 are the value of units draw at the first and 2nd draw respectively and P_1 and P_2 are their corresponding initial probability.

Define the estimators

$$z_1 = y_1 / P_1$$

$$z_2 = y_2 / y_2(1-P_1)/P_2.$$

Then,

$$\hat{Y}_D = (z_1 + z_2)/2$$

$$= \left[y_1(1+P_1)/P_1 + y_2(1-P_1)/P_2 \right].$$

$$= \frac{\frac{y_1}{P_1}\left[1+P_1\right] + \frac{y_2}{P_2}\left[1-P_1\right]}{2}.$$

Theorem:

In PPS SWOR, the estimator $\hat{\bar{Y}}_D$ is an unbiased estimator & sampling variance:

$$V(\hat{\bar{Y}}_D) = \left(1 - \frac{1}{n}\sum_i^N P_i^2\right)\left[\frac{1}{2}\sum_i^N \left(y_i/P_i - \bar{Y}\right)^2 P_i\right] - \frac{1}{4}\sum_i^N \left(y_i/P_i - \bar{Y}\right)^2 P_i^2.$$

Proof:

Where $z_i = y_i/P_i$.

(ie) $E(z_i) = \sum\left(y_i/P_i\right) = \sum y_i = \bar{Y}$.

Similarly for given y_1 drawn at the first draw we have, $E\left[y_2\left(1-P_1\right)/P_2 \,/\, y_1\right] = \sum \frac{y_j(1-P_i)}{P_j}\,\frac{P_j}{(1-P_i)}$.

where $\sum$ is taken over all the values of y except y_1.

Hence $E\left[y_2\left(\frac{1-P_1}{P_2}\right)/y_1\right] = \bar{Y} - y_1$

$\Rightarrow E(z_2) = E\left[E[z_2/y_1]\right]$

$= y_1 + \bar{Y} - y_1$

$= \bar{Y}$.

In General, $E[\bar{y}_D] = \bar{Y}$ thus we can be show that

To find out variance.

$$V[Z_1] = \sum_i \sum_j p_i p_j \left(y_i/p_i - y_j/p_j\right)^2$$

also $V(Z_2) = E_1 V_2(Z_2) + V_1 E_2(Z_2)$

Since,

$E_2(Z_2) = \bar{Y}$, $V_1 E_2(Z_2) = 0$.

hence we have

$$V(Z_2) = \sum_i \sum_j p_i p_j \left(y_i/p_i - y_j/p_j\right) \cdot (1 - p_i - p_j)$$

Thus,

$$V(\bar{y}_D) = \frac{1}{4} \sum_i \sum_j p_i p_j \left(\frac{y_i}{p_i} - \frac{y_i}{p_j}\right)(2 - p_i - p_j)$$

Contd....

Theorem (2).

In PPS SWOR the estimator $\hat{Y}_D$ is an unbiased estimator of Population total Y, and its Sampling Variance is give by

$$V(\hat{Y}_D) = \frac{1}{n^2} \sum_i \sum_j p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \left[1 + r_{ij}(1) + \cdots + r_{ij}^{(k)} + \cdots r_{ij}(n-1) \right]$$

where $r_{ij}^{(k)}$ is the prob. that y_i & y_j are not included in the Sequence.

Proof:

$E(Z_1) = Y$ and $E[Z_i / y_1, y_2 \cdots y_{i-1}) =$;

where $i = 2, 3 \cdots n$

hence

$E(Z_i) = Y$ for $i = 2, 3 \cdots n$

It follows that

$\hat{Y}_D = \bar{Z} = \sum_i Z_i / n$. is an unbiased estimator.

To derive Variance : -

$E(Z_i Z_j) = Y^2$; its shows that Z_i and Z_j are uncorrelated.

As the expansion for $V(\hat{Y}_D)$ is some what Complex, it can be modified to a Simpler form as $V(\hat{Y}_D) = V\left[\sum_{i=1}^{n} \frac{Z_i}{n} \right] = \frac{1}{n^2} \sum_i V(Z_i)$.

Unordered estimator:

The unordered estimator which do not depends on a order in which the units of drawn within the sample.

Horvitz and Thompson (1952), Murthy (1957) and Basu (1958), have shown that these unordered estimators are more efficient than ordered estimators.

Horvitz and Thompson estimator:-

Horvitz and Thompson (1952) suggested an estimator which is an unbiased estimator of population total. Let the initial probability of selection of the unit U_i is p_i, where

$$p_i = \frac{x_i}{X} \quad \text{for } i = 1, 2 \ldots N.$$

The prob. that unit U_i is included in the sample would be given by

$$\pi_i = p_i + \sum_{j \neq i} p_j p_i / (1 - p_j).$$

$$= p_i \left[1 + \sum_{j \neq 1} p_j / (1 - p_j) \right].$$

From that the prob. that both a units U_i and U_j are included in the sample is given by

$$\pi_{ij} = p_i p_j / (1-p_i) + p_i p_j / (1-p_j)$$

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

Suppose that y_i be the value of the i^{th} unit with π_i, the prob. of inclusion in the sample. Horvitz Thompson estimator is defined by

$$\hat{Y}_{HT} = \sum_i y_i / \pi_i$$

Theorem:

In PPS SWOR $\hat{Y}_{HT}$ is unbiased and its sampling variance is given by.

$$V_{HT}(\hat{Y}_{HT}) = \sum_i^N \frac{(1-\pi_i)}{\pi_i} y_i^2 + \sum_i^N \sum_{j \neq i}^N \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j$$

where, π_{ij} is the prob of inclusion of both the i^{th} and j^{th} units in the sample.

Proof:

The most general form of linear estimator

$$\hat{Y} = \sum_i^N a_i c_i y_i$$

where, a_i is a random variate that takes the values 1 if the i^{th} unit is drawn and 0 otherwise.

C_i are constants attached to the units U_i, $i = 1, 2 \ldots N$.

Obviously a_i follows a binomial distribution for a sample of size 1 with prob P_i.

Hence, $E(a_i) = \pi_i$

$V(a_i) = \pi_i \times (1 - \pi_i)$

Since $a_i \cdot a_j = 1$ only if both the units are distinct and appear in the sample, Covariance of

$Cov(a_i, a_j) = \pi_{ij} - \pi_i \pi_j$

Now, $E(\bar{y}) = E\left(\sum_1^N a_i c_i y_i\right)$

$$E(\hat{\bar{y}}) = E\left[\sum_i^N a_i c_i y_i\right)$$

$$E(\hat{\bar{y}}) = \bar{Y}.$$

It is unbiased

$$\therefore c_i = 1/\pi_i .$$

Hence, $\hat{\bar{y}}_{HT} = \sum_i^N y_i/\pi_i$ is an unbiased estimator.

$$V_{HT}\left(\hat{\bar{y}}_{HT}\right) = V_{HT}\left(\sum_i^N a_i c_i y_i\right)$$

$$V_{HT}\left(\hat{\bar{y}}_{HT}\right) = \sum_i^N \frac{y_i^2}{\pi_i^2} V(a_i) + \sum_i^N \sum_{j \neq i}^N \frac{y_i y_j}{\pi_i \pi_j} \operatorname{cov}(a_i, a_j).$$

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

Note:

① The $V\left(\hat{\bar{y}}_{HT}\right)$ depends on π_{ij} and π_j which are calculated from sampling procedure.

② $\pi_i = n y_i/y$, $V(\hat{\bar{y}})$ becomes zero.

③ If there is some way of choosing values of π_i so that they are very close $n y_i/y$, $\hat{\bar{y}}$ will have a very small variance.

Result:

1) An unbiased estimator of $V\left(\hat{\bar{y}}_{HT}\right)$ is given by

$$V\left(\hat{\bar{y}}_{HT}\right) = \sum_i^n (1-\pi_i) \frac{y_i^2}{\pi_i^2} + \sum_i^n \sum_{j>i}^n \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) \frac{y_i y_j}{\pi_{ij}}.$$

Provided that none of the π_{ij} in the formula is zero.

(2) Another expression for $V(\bar{y}_{HT})$ derived by Yates and Grundy (1923) is given by,

$$V_{YG}(\bar{y}_{HT}) = \sum\sum_{i<j} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \text{ and}$$

$\bar{y}_{HT}$ is unbiased estimator is given by

$$V_{YG}(\bar{y}_{HT}) = \sum_{i<j}\sum \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Moorthy's unordered estimator :-

Moorthy (1957) suggested that an unordered estimator can be obtained by weighting all poridble ordered estimator with their respective prob's. In sampling n units WOR from a finite popln, there will be $\binom{N}{n}$ unordered samples. Each unordered sample of size, n can be ordered in M (= n!) ways: (ie) an unordered sample correspond to M ordered samples. Consider a scheme of selection in which the prob of selecting the sample so is $p(s_i)$. Then the prob getting unordered sample the sum of prob of getting the ordered samples (s) corresponding to (s), (ie) $p_o = \sum_{i}^{M} p_{si}$. Let y_{si} be an estimator of popln parameter Q based on the ordered sample s_i an unordered

estimate Q is given by $\hat{Y}_M = \sum_{i}^{M} y_{si} P'_{si}$ where $P'_{si} = P_{ri}/P_s$.

Theorem (5):-

In pps SWOR the unordered estimator $\bar{y}'_m$ is an unbiased estimator of Q. and it's sampling variance is given by

$$V(\bar{y}'_m) = \sum_s P_s \left[\sum_i y_{si} P'_{si}\right]^2 - \left(\sum_s \sum_i y_{si} P_{si}\right)^2$$

Result :-

1) In pps SWOR the ordered estimator y_{si} is an unbiased estimator of Q and its sampling variance is given by

$$V(y_{si}) = \sum_s \sum_i P_{s_i} y_{s_i}^2 - \left(\sum_s \sum_i y_{si} P_{si}\right)^2$$

The variance of an unordered estimator $\bar{y}'_N$ is less than or equal to that of the ordered estimator, because

$$V(y_{si}) - V(\bar{y}'_m) = \sum_s \sum_i P_{si} y_{si}^2 - \sum_s \sum_i P \left[\sum_i y_{si} P'_{si}\right]^2$$

2) An unbiased estimator of the sampling variance of $\bar{y}'_m$ is given by

$$v(\bar{y}'_m) = \frac{1}{(P_s)^2} \sum_i^n \sum_{i>j}^n \left[\frac{P_s}{P} P (y_i - \bar{y}) - P_{si} P_{sj}\right] - P_i P_j \left(\frac{y_i}{P_i} - \frac{y_j}{P_j}\right)^2$$

where, $p(s|i,j)$ is a conditional prob. of getting the s^{th} sample given that the units i & j are already selected in the first two draws.

PPS Systematic Sampling :-

This PPS Systematic Sampling is always better than PPS SWR. Madow (1949) Suggested that generalization can be made out sys. sampling with equal probs.

The method consists of arranging the units at random and frequency cumulative totals

$$T_i = \sum_{j=1}^{i} X_j \quad (i = 1, 2 \cdots N)$$

where X_j is size of j^{th} unit. In selecting n units by PPS Systematic Sampling, a random number R is choosen from 1 to k, where $k = \frac{T_N}{n}$. The units corresponding to the nos $R + jk$ $(j = 0, 1, 2 \cdots (n-1))$ ~~constitute~~ consitute the sample.

(Ex),

Consider the example give in cumulative total

here, $T_N = \sum_{j=1}^{N} X_j = 330$, and $k = \frac{330}{4} = 82.5$

$k \simeq 82$

Suppose the r.n choosen below 1 to 82 is 51.

Then the units corresponding to the sys sample numbers.

$R + jk$

$R = 51$, $j = 0, 1, 2 \ldots (n-1)$

$R + k = 51 + 82 = 133$

$R + 2k = 51 + 2 \times 82 = 215$

$R + 3k = 51 + 3 \times 82 = 297.$

51, 133, 215, 297 are 2, 4, 6 and 9. It is clear from the procedure that i^{th} unit is selected in the sample if

$T_{i-1} < R + jk \leq T_i$ for $j = 0, 1, 2 \ldots (n-1)$

and the prob. of inclusion of i^{th} unit is np_i.

This method can be applied even when $\frac{T_N}{n}$ is not an integer. The sampling interval k can be taken as integer nearest to $\frac{T_N}{n}$.

An unbiased estimator of Popln total Y, derived by Hartley and Rao (1962) is given by

$$\hat{Y}_{HR} = \sum_{1}^{n} \frac{y_i}{n_i} = \frac{n_2 \bar{y}}{\bar{y}\, np_i}.$$

The main drawback of this method, as in sys. sampling is that, an unbiased variance estimator on the basis of single sample is

not possible Hartley and Rao (1962) have further considered this selection procedure when the units are arranged at random and assymptotic approach has been used in developing formular for variance and estimated variance. For large values of N and for value of n relatively small compared to N, the approx sampling variance of the estimator can be written as

$$V(\hat{\bar{Y}}_{HR}) \simeq \frac{N}{2} \sum_i \left(\frac{y_i}{p_i} - \bar{Y}\right)^2 p_i \left[1-(n-1)p_i\right]$$

which shown that even the the units are arranged from random, selecting them with PPS Sys sampling is more efficient than PPS SWR for the terms $[1-(n-1)p_i]$ are acting as reduction terms.

Further this variance can be estimated by

$$v(\bar{y}_{HR}) = \frac{1}{n^2(n-1)} \sum_{i}^{n_2} \sum_{j>i}^{n} \left[1 - n(p_i + p_j) + \sum_{j}^{N} p_j^2\right] \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2.$$