UNIT 5

PROBABILITY PROPORTIONAL TO SIZE SAMPLING WITHOUT REPLACEMENT

Structure

5.1 Introduction

Expected Learning Outcomes

5.2 Probability Structures in Probability Proportional to Size Sampling Schemes

Probability Structure in Probability Proportional to Size with Replacement Sampling Scheme

Probability Structure in Probability Proportional to Size without Replacement Sampling Scheme

5.3 Ordered Estimators

Concept of Ordered Estimators

Des Raj's Ordered Estimator of Population Mean for Sample of Size 2

Sampling Variance of Des Raj's Sample Mean Estimator

Estimation of Sampling Variance of Des Raj's Sample Mean Estimator

General Form of Des Raj's Estimator

Das Ordered Estimator

5.4 Unordered Estimators

Existence of an Unordered Estimator

Constructing Unordered Estimator Using Des Raj's Ordered Estimators

5.5 Inclusion Probabilities

Order of an Inclusion Probability

Some Important Results Related to Inclusion Probabilities

5.6 Horvitz-Thompson Estimator

Horvitz–Thompson Estimator for Population Total

Sampling Variance of the Horvitz–Thompson Estimator for Population Total

Another Form of the Sampling Variance of the Horvitz-Thompson Estimator for Population Total

Unbiased Estimators of Variance of Horvitz-Thompson Estimator of Population Total

5.7 Non-negativity Issue of Estimator of the Sampling Variance of Horvitz-Thompson Estimator

Midzuno-Sen Sampling Scheme

Resolving the Issue of Non-Negativity of Estimator of Sampling Variance Using Midzuno-Sen Sampling Scheme

5.8 Summary

5.9 Terminal Questions

5.10 Answers / Solutions

5.1 INTRODUCTION

Till now, in Unit1 we discussed the simplest and most fundamental sampling scheme, namely, Simple Random Sampling with Replacement (SRSWR) and, consequently, in Unit 2 confined to the study of Simple Random Sampling without Replacement (SRSWOR). Both the schemes SRSWR and SRSWOR are members of the class of EPSEM sampling schemes, since, in both of these schemes, the probability of selection of a particular unit of the population at any draw is same, which is 1/N. Further, due to the reason that in some circumstances, it seems reasonable and justifiable to assign unequal selection probabilities to population units; we discussed an advanced sampling scheme in Unit 4, namely, Probability Proportional to Size with Replacement (PPSWR) scheme which we observed to be the generalized version of the SRSWR scheme. It was shown that SRSWR scheme is a special case of PPSWR in the sense that results of the former scheme are easily deducible from those obtained in the latter scheme.

Having acquainted with many of the concepts, theories and results related to Varying Probability Sampling Scheme (VPSS), in general and PPSWR sampling scheme, in particular in the previous unit; we shall confine ourselves in the present unit to the study of Probability Proportional to Size without Replacement (PPSWOR) sampling scheme. It is clear that the difference between PPSWR and PPSWOR schemes is that in the former scheme, unit once selected at any draw was replaced again in the population before the next draw; in the latter, selected units are not replaced in the population again and, therefore, the population size reduced by one after each draw.

In this unit at the outset, Section 5.2 highlights the basic difference between the probability structure of PPSWR and PPSWOR sampling schemes. This helps us to understand why the PPSWOR scheme needs revision of selection probabilities at each draw. Section 5.3 discusses the ordered samples and consequently, concept of ordered estimators. This section also provides a special type of ordered estimators, namely, Des Raj's ordered estimator and derives the sampling variance of the ordered estimator of population mean. In Section 5.4, we discuss the concept of unordered estimators and have proved that corresponding to some ordered estimators, there always exists an unordered estimator having more efficiency. The construction of an unordered estimator on the basis of Des Raj's ordered estimator has also been illustrated in this section. Section 5.5 provides the important concept of inclusion probabilities and describes some important and useful relations between first order and second-order inclusion probabilities. Section 5.6 defines the Horvitz-Thompson (H-T) unordered estimator of population total. The sampling variance of the estimator and its unbiased estimator have also been derived in this section. The expressions of variance of the population total as proposed by Yates-Grundy and Sen have been obtained and the issue of non-negativity of the estimate of the variance has been explained. Section 5.7 provides a condition on inclusion probabilities under which the variance of the estimator becomes positive. In this context, the Midzuno-Sen sampling scheme has also been explained and it has been shown that under this scheme, estimate of the variance of Horvitz-Thompson estimator is always positive.

Expected Learning Outcomes

After studying this unit, you should be able to:

- discuss the Probability Proportional to Size without Replacement sampling scheme and explain about the complexity of such sampling scheme due to its complex probability structure;
- describe the concept of ordered samples and ordered estimators;
- define ordered estimators proposed by several statisticians and discuss Des Raj's ordered estimator and obtain its sampling variance;
- define unordered estimator and show that an unordered estimator, obtained through ordered estimators, always exist with a greater efficiency;
- define Horvitz-Thompson unordered estimator for population total and obtain its sampling variance and also identify an unbiased estimator of the sampling variance;
- explain the non-negativity issue of the estimate of sampling variance of Horvitz-Thompson estimator; and
- describe the method of resolving the Non-negativity issue using Midzuno-Sen sampling scheme.

5.2 PROBABILITY STRUCTURES IN PROBABILITY PROPORTIONAL TO SIZE SAMPLING SCHEMES

We discussed and explained some basic concepts and operational methods of finding selection probabilities of units of the population in PPS sampling scheme and particularly in Probability Proportional to Size with Replacement sampling scheme in previous unit. It was observed that in varying probability sampling schemes, to which PPS scheme is a special case; deciding the probability of selection of a population unit with which it is to be selected in the sample is the most important task. In Probability Proportional to Size with Replacement sampling scheme, we saw that using either the Cumulative Total Method or Lahiri's Method one is assured to select a unit of the population with probability proportional to the size measure of the unit concerned.

5.2.1 Probability Structure in Probability Proportional to Size with Replacement Sampling Scheme

Let us first observe that in Probability Proportional to Size with Replacement scheme whether the selection probabilities which are associated with the population units before starting the selection process, changes as we proceed for drawing units one by one in different draws or remain unchanged during the entire selection process. It is very much important to know for understanding the problems which arise when we deal with the Probability Proportional to Size without Replacement scheme.

Using the usual symbols of the previous unit, we can list the units of the population, their labels and associated selection probabilities as shown below:

Population (U)	U_1	U_2	U_3	…	U_i	…	U_N
Label (i)	1	2	3	…	i	…	N
Probability of Selection (p)	p_1	p_2	p_3	…	p_i	…	p_N

We know that the probabilities of selection $(p_i's)$, in the above table are obtained using a chance mechanism which provide these probabilities proportional to the size of the corresponding unit, that is,

$$p_i \propto X_i \; (i = 1, 2, 3, \ldots, N).$$

$p_i s$; being probabilities, satisfy the conditions:

$$p_i > 0 \; (i = 1, 2, 3, \ldots, N) \quad \text{and} \quad \sum_{i=1}^{N} p_i = 1.$$

Since, these probabilities are obtained before the selection process of the units start, these probabilities are called "Initial Probabilities of Selection". Thus, the selection probabilities as obtained in **Example 1** of the Unit IV, given by,

p_i	0.11	0.07	0.09	0.14	0.11	0.08	0.15	0.18	0.07

are 'Initial Probabilities of Selection', satisfying the two conditions as mentioned above.

Henceforth, we shall call any such set of ordered selection probabilities as "Structure of the Selection Probabilities", which provides us the pairs

$$(1, p_1), (2, p_2), (3, p_3), \ldots, (i, p_i), \ldots, (N, p_N);$$

where in the pair (i, p_i), i stands for the label of the unit and p_i for the corresponding selection probability.

Suppose in Probability Proportional to Size with Replacement scheme, during the selection process, label 3 was selected at the first draw, then obviously, it was selected with selection probability p_3. If you think about the probability structure of selection probabilities after the first draw, it would be $(p_1, p_2, p_4, p_5, \ldots, p_N)$; the total of which would be $1 - p_3 \neq 1$. This shows that the second condition imposed on $p_i s$ is not fulfilled. However, in Probability Proportional to Size with Replacement scheme, since the selected unit is again replaced in the population before the next draw, the probability structure again becomes $(p_1, p_2, p_3, p_4, p_5, \ldots, p_N)$; which is same as the structure of the initial probabilities, and this would happen before and after each draw.

Therefore, we conclude that:

"The structure of the selection probabilities remains unchanged in Probability Proportional to Size with Replacement scheme before and after each draw irrespective of the unit selected at any draw".

In fact, this characteristic of Probability Proportional to Size with Replacement scheme makes it simple in handling during the selection of a sample.

5.2.2 Probability Structure in Probability Proportional to Size without Replacement Sampling Scheme

Now, we shall discuss the nature of the structure of selection probabilities in the Probability Proportional to Size without Replacement scheme, where after making a draw, the selected unit is not replaced again in the population before the next draw, resulting in the decrease of number of units in the population by one after each draw. As above, we write the units along with their labels and initial selection probabilities:

Population (U)	U_1	U_2	U_3	…	U_i	…	U_N
Label (i)	1	2	3	…	i	…	N
Probability of Selection (p)	p_1	p_2	p_3	…	p_i	…	p_N

Let at the first draw, 6th unit of the population be selected with the associated selection probability p_6. Obviously, this unit is not replaced in the population before the next draw. Therefore, now in the population one unit would be absent (6th unit) and, hence, the changed (new) probability structure would consist of the probabilities:

Changed (new) Probability Structure (p)	p_1	p_2	p_3	p_4	p_5	p_7	p_8	…	p_N

so that the total of all the changed probabilities would be $(1-p_6) \neq 1$. Since, the 6th unit is not replaced in the population before the second draw, the sum of new probabilities does not fulfil the important condition of the probability. Obviously, therefore, we have to revise the new probabilities in such a way that their sum over all the units becomes unity. Let us divide each probability as obtained after the first draw by $(1-p_6)$. Then, the revised probabilities after the first draw would be

$$\left(\frac{p_1}{1-p_6}, \frac{p_2}{1-p_6}, \frac{p_3}{1-p_6}, \frac{p_4}{1-p_6}, \frac{p_5}{1-p_6}, \frac{p_7}{1-p_6}, \frac{p_8}{1-p_6}, \ldots, \frac{p_N}{1-p_6}\right);$$

such that $\sum_{(i\neq 6)=1}^{N} \frac{p_i}{1-p_6} = 1.$

Since the sum of that remaining probabilities is $\sum_{i=1}^{N} p_i \neq 1$. However, it is

$$\sum_{i\neq 6=1}^{N} p_i = 1-p_6 \quad \Rightarrow \quad \sum_{i\neq 6=1}^{N} \frac{p_i}{1-p_6} = 1$$

Thus, after the first draw, we see that instead of using initial selection probabilities, we have to use the revised probabilities. You can see that revised probabilities depend upon the unit which was selected at the first draw, since, if the 2nd unit was selected at the first draw, the revised probabilities would have been

$$\left(\frac{p_1}{1-p_2}, \frac{p_3}{1-p_2}, \frac{p_4}{1-p_2}, \frac{p_5}{1-p_2}, \frac{p_6}{1-p_2}, \frac{p_7}{1-p_2}, \frac{p_8}{1-p_2}, \ldots, \frac{p_N}{1-p_2}\right);$$

such that $\sum_{(i\neq 2)=1}^{N} \frac{p_i}{1-p_2} = 1.$

Suppose unit 3 is selected at the first draw and unit 5 is selected at the second draw, then the revised probabilities after the second draw would be

$$\left(\frac{p_1}{1-p_3-p_5}, \frac{p_2}{1-p_3-p_5}, \frac{p_4}{1-p_3-p_5}, \frac{p_6}{1-p_3-p_5}, \frac{p_7}{1-p_3-p_5}, \frac{p_8}{1-p_3-p_5}, \ldots, \frac{p_N}{1-p_3-p_5}\right)$$

such that $\sum_{(i\neq 3\neq 5)=1}^{N} \frac{p_i}{1-p_3-p_5} = 1$ and so on.

These revised probabilities are different than the above revised probabilities. It is clear that the revised probabilities are, in fact, conditional probabilities and depend upon the outcomes of the previous draws. On the basis of the above results, therefore, we reach to the following conclusion:

"It can be concluded that in PPSWOR, the probability structure of selection probabilities after each draw needs to be revised on the basis of units which have been selected in the previous draws. The revised probabilities, therefore, are not the same as the initial probabilities".

A theoretical proof of the above statement is presented in the following theorem:

Theorem 1: In Probability Proportional to Size without Replacement sampling scheme, in which the initial selection probabilities are unequal, the probability of a specified unit of the population at a given draw changes with the draw.

Proof: Let the initial probability of selection of the i^{th} unit of the population, in the sample at the first draw, be $\{p_i; (i = 1, 2, 3, \ldots, N)\}$ where $p_i > 0$ for all i and $\sum_{i=1}^{N} p_i = 1$. Further, let p_{i_k} be the probability of selection of the i^{th} unit at the k^{th} draw, where $k = 1, 2, \ldots, n$. Since, we are considering PPSWOR sampling scheme, a unit drawn at any draw is not replaced in the population.

Then, it is clear that

$$p_{i_1} = p_i \quad \text{for } i = 1, 2, 3, \ldots, N; \qquad \ldots (5.1)$$

indicating that if a unit is selected at the first draw, it will be selected with its initial selection probability. Now, we observe that the probability of selection of the i^{th} unit at the second draw is

$$p_{i_2} = P\begin{bmatrix} i^{th} \text{unit is not selected} \\ \text{at the first draw} \end{bmatrix} \times P\begin{bmatrix} i^{th} \text{unit is selected} \\ \text{at the second draw} \end{bmatrix. \Big/ \begin{matrix} \text{it is not selected} \\ \text{at the first draw} \end{matrix} \end{bmatrix}$$

$$p_{i_2} = \sum_{j(\neq i)=1}^{N} P\begin{Bmatrix} j^{th} \text{ unit is selected at} \\ \text{the first draw } (j\neq i) \end{Bmatrix} \times P\begin{bmatrix} i^{th} \text{unit is selected at} \\ \text{the second draw} \end{bmatrix. \Big/ \begin{matrix} j^{th} \text{ unit is selected at} \\ \text{the first draw } (j\neq i) \end{matrix} \end{bmatrix}$$

since, j^{th} unit may assume any of the N population values except i^{th} value, which itself can assume any of the N population values. So, there is the above summation over all possible combinations of j and i.

Thus,

$$p_{i_2} = \sum_{j(\neq i)=1}^{N} p_j \times \frac{p_i}{(1-p_j)} = \left[\sum_{j=1}^{N} \frac{p_j}{1-p_j} - \frac{p_i}{1-p_i}\right] p_i \neq p_{i_1};$$

$$\text{since,} \quad \sum_{j(\neq i)=1}^{N} \frac{p_j}{(1-p_j)} = \left\{\sum_{j=1}^{N} \frac{p_j}{1-p_j} - \frac{p_i}{1-p_i}\right\}.$$

Therefore, we see that, $p_{i_2} \neq p_{i_1}$, that is, the probability of selection of the i^{th} unit of the population at the second draw is not same as its initial probability of selection of the j^{th} unit at the first draw. Thus, the probability of selection of a particular unit, changes with the draw.

Proceeding in the same way, it can be shown that $p_{i_3} \neq p_{i_2} \neq p_{i_1}$ and so on.

This establishes the theorem.

Remark 5.1: Due to the reason that probability of selection of a particular unit of the population in Probability Proportional to Size without Replacement sampling scheme changes from draw to draw, the Probability Proportional to Size without Replacement scheme is quite complex scheme which is not easy to handle. The order of the selection of units in the sample does affect the results. This means that if units 1, 2 and 3 are selected in the sample with PPSWOR scheme, the samples with order of units (1, 2, 3), (2, 3, 1), (3, 1, 2), (1, 3, 2), etc., would produce different estimators. We shall see afterwards how to tackle this complicated problem of ordering of units in the sample.

Remark 5.2: In SRSWOR, it was observed that the probability of selection of a particular unit of the population in the sample at any draw is same and is given by 1/N. We can see that in PPSWOR scheme, if initial probabilities of selection for all the units is taken to be equal to 1/N, we get the same result as obtained under SRSWOR scheme. Let us prove this statement theoretically as follows:

If probabilities of selection for all units is taken to be equal 1/N, that is, $p_i = \frac{1}{N}$ for all i = 1, 2, …, N, then, $p_{i_1} = p_i = \frac{1}{N}$.

We have given that $p_i = \frac{1}{N}$ for all i = 1, 2, …, N. Now, we have

$$p_{i_2} = \left[\sum_{j=1}^{N} \frac{p_j}{1-p_j} - \frac{p_i}{1-p_i}\right] p_i = \left[\sum_{j=1}^{N} \frac{1/N}{1-1/N} - \frac{1/N}{1-1/N}\right] \frac{1}{N},$$

since $p_i = \frac{1}{N}$ for all i = 1, 2, …, N.

$$\text{But} \sum_{j=1}^{N} \frac{1/N}{1-1/N} - \frac{1/N}{1-1/N} = \sum_{j=1}^{N} \frac{\frac{1}{N}}{\frac{N-1}{N}} - \frac{\frac{1}{N}}{\frac{N-1}{N}}$$

$$= \sum_{j=1}^{N} \frac{1}{(N-1)} - \frac{1}{(N-1)} = \frac{N}{(N-1)} - \frac{1}{(N-1)}$$

Therefore,

$$p_{i_2} = \left[\frac{N}{(N-1)} - \frac{1}{(N-1)}\right] \cdot \frac{1}{N} = \frac{1}{N} = p_{i_1} = p_i.$$

This shows that $p_{i_2} = p_i$ if $p_i = \frac{1}{N}$ for all i.

Proceeding in the same way, we can show that

$$p_{i_1} = p_{i_2} = p_{i_3} = \ldots = p_{i_n} = \frac{1}{N}.$$

Example 1: Below are given initial selection probabilities of the population units:

Population (U):	U_1	U_2	U_3	U_4	U_5	U_6	U_7
Label (i):	1	2	3	4	5	6	7
Probability of Selection (p):	0.15	0.20	0.09	0.32	0.16	0.05	0.03

Let a sample of size 3 be selected from the population using PPSWOR scheme. It is known that the units selected in the three successive draws were U_7, U_3 and U_5. Find the selection probabilities with which these units were selected.

Solution: We have the given data as follows:

Population (U):	U_1	U_2	U_3	U_4	U_5	U_6	$\mathbf{U_7}$
Label (i):	1	2	3	4	5	6	**7**
Probability of Selection (p):	0.15	0.20	0.09	0.32	0.16	0.05	**0.03**

Given that 7th, 3rd and 5th units are respectively selected in the sample in first, second and third draws. Since, 7th unit is selected at the first draw, its selection probability will be 0.03 as per the above table. Now, as soon as the first draw is complete, the probabilities of selection must be revised under the condition that 7th unit is selected at the first draw. Therefore, the revised probability of selection of remaining units would be as follows:

Population (U):	U_1	U_2	U_3	U_4	U_5	U_6
Label (i):	1	2	**3**	4	5	6
Probability of Selection after First Draw	$\frac{0.15}{1-0.03} = 0.1546$	$\frac{0.20}{1-0.03} = 0.2062$	$\frac{0.09}{1-0.03} = 0.0928$	$\frac{0.32}{1-0.03} = 0.3299$	$\frac{0.16}{1-0.03} = 0.1649$	$\frac{0.05}{1-0.03} = 0.0516$

Therefore, at the second draw, the 3rd unit will be selected with selection probability equal to 0.0928 as per above table.

Since at the second draw 3rd unit is selected, before the third draw, we have to compute revised selection probabilities accordingly. Obviously, now revised probabilities are to be computed as follows:

U_1	U_2	U_4	**U_5**	U_6
1	2	4	**5**	6
$\frac{0.15}{1-0.03-0.09} = 0.1704$	$\frac{0.20}{1-0.03-0.09} = 0.2273$	$\frac{0.32}{1-0.03-0.09} = 0.3636$	$\frac{0.16}{1-0.03-0.09} = 0.1818$	$\frac{0.05}{1-0.03-0.09} = 0.0568$

The table shows that 5^{th} unit is selected at the third draw with selection probability equal to 0.1818.

Now, you may like to try to answer the following Self-Assessment Question based on the content of this section:

SAQ 1

In what sense, the selection probability structure of PPSWOR scheme differ than that of PPSWR scheme?

5.3 ORDERED ESTIMATORS

Keeping in view the result of **Theorem 1** and observations made in **Remark 5.1**, it can be concluded that in Probability Proportional to Size without Replacement sampling scheme, selection probabilities of the units of the given population do not remain same in each draw, as these were observed to be same for all the draws in Probability Proportional to Size with Replacement scheme. This change in selection probability over different draws happens because of the order of selection of the units in the sample. This fact can be elaborated in terms of order of arrangement of units in the sampling frame and their associated selection probability as shown below:

$$\begin{pmatrix} Y_1 & Y_2 & Y_3 & Y_4 & \ldots\ldots & Y_8 & Y_9 & Y_{10} \\ p_1 & p_2 & p_3 & p_4 & \ldots\ldots & p_8 & p_9 & p_{10} \end{pmatrix}$$

No doubt, here p_is $(i = 1, 2, 3, \ldots, 10)$ are initial probabilities which are obtained through the methods which assure the probabilities to be proportional to the size measure of the respective units and attached to the units before starting the sample selection process. The initial probability of selection of a unit is workable only if it is selected at the first draw, for example, if unit Y_3 is selected at the first draw, then its selection probability would be p_3. But if Y_4 is selected at the first draw and then Y_3 is selected at the second draw, the selection probability of Y_3 would not be p_3. Instead, its selection probability would be something else, which could be calculated as shown in the Sub-section 5.2.2. This means that in a sample of size 2, consisting units Y_3 and Y_4, the probability of selection of samples (Y_3, Y_4) and (Y_4, Y_3) would be different, whereas both the samples possess the same two units and it is the values of the units which are involved in the estimation process of a population parameter and not the order in which they are selected.

In Probability Proportional to Size without Replacement sampling scheme, therefore, the order in which different units are selected in the sample plays an important role. Owing to this reason, some special types of estimators have been developed for population parameters under Probability Proportional to

Size without Replacement, taking into account the order of the units in which they are selected in the sample.

5.3.1 Concept of Ordered Estimators

Since, the selection probability of a particular unit in Probability Proportional to Size without Replacement scheme changes from draw to draw, the expectation of the study variable (y) also changes over draw for the same unit.

Definition of an Ordered Estimator:

"Ordered Estimators are those estimators which take into account the order of selection of units in the sample, but at the same time, overcome the difficulty of changing expectation with each draw by associating a new variate with each draw such that its expectation is equal to the concerned population parameter".

As per above definition, in order to construct an ordered estimator for any population parameter we need to introduce a new variate with each draw so that the expectation of the variate at that draw becomes equal to the concerned population parameter. If it is possible, it can be seen that the final estimator is an unbiased estimator for the population parameter. This concept of changing the variate-value at each draw can be better understood by considering the estimator of population mean in Simple Random Sampling without Replacement scheme, for illustration purpose:

We know that in Simple Random Sampling without Replacement, the sample mean estimator is an unbiased estimator of population mean, that is,

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

where $\bar{y} = \frac{1}{n}\sum_{r}^{n} y_r = \frac{1}{n}[y_1 + y_2 + y_3 + \ldots + y_r + \ldots + y_n]$;

y_r being the unit of the population drawn at the r^{th} draw (r = 1, 2, 3, …, n).

Therefore, we see that

$$E(\bar{y}) = E\left\{\frac{1}{n}\sum_{r}^{n} y_r\right\}$$

$$= E\left[\frac{1}{n}\{y_1 + y_2 + y_3 + \ldots + y_r + \ldots + y_n\}\right]$$

$$= \frac{1}{n}\left[E(y_1) + E(y_2) + E(y_3) + \ldots + E(y_r) + \ldots + E(y_n)\right]. \qquad \ldots (5.2)$$

Now, as we know, each y_r $(r = 1, 2, 3, \ldots, n)$ is a random variable which may assume any of the N values of the study variable Y in the population, denoted by $Y_1, Y_2, Y_3, \ldots, Y_N$; respectively, with some pre-determined selection probabilities, say, $p_1, p_2, p_3, \ldots, p_N$,; such that conditions: $p_i > 0$ for i = 1, 2, …, N and $\sum_{i=1}^{N} p_i = 1$ are satisfied.

Therefore, using the definition of expectation, we can write

$$E(y_r) = \sum_{i=1}^{N} p_i Y_i \quad \text{for} \quad i = 1, 2, \ldots, N. \qquad \ldots (5.3)$$

In this expression, putting the values of p_i for all i; whatever may be their pre-determined values, we can get the expression of $E(y_r)$ which can be substituted in (5.2) in order to get the value of $E(\bar{y})$. So, we have

$$E(\bar{y}) = \frac{1}{n}\sum_r^n \left[\underbrace{\sum_{i=1}^N p_i Y_i + \sum_{i=1}^N p_i Y_i + \sum_{i=1}^N p_i Y_i + \ldots + \sum_{i=1}^N p_i Y_i}_{n \text{ times}} \right] \qquad \ldots (5.4)$$

Since, in SRSWOR, we have seen that $p_i = \frac{1}{N}$ for all i, therefore, each term in the bracket in (5.4) is equal to $\left(\sum_{i=1}^N \frac{1}{N} Y_i = \bar{Y}\right)$, the population mean.

Therefore, we get

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

Now, think that probability of selection at the first draw, second draw, third draw, …, in (5.2) changes according to the order of the selection of units, then after every draw it is needed to define the value y_r for r = 1, 2, …, n; so that at each draw we have $E(y_r) = \bar{Y}$ in order to make the estimator $(\bar{y})$ unbiased for $\bar{Y}$. In this sense, such an estimator is called 'Ordered Estimator'.

Many of the ordered estimators for the same population parameter are available in the literature of the sampling theory. Some of the popularly known ordered estimators are given by Das (1951), Sukhatme (1953) and Des Raj (1956). All these authors have defined a different type of ordered estimator for the same population parameter by considering a new variate-value at each draw.

The following sub-section deals with the construction of Des Raj's Ordered Estimator and its variance:

5.3.2 Des Raj's Ordered Estimator for Sample of Size 2

Due to the complexity of ordered estimators, it is too difficult to define Des Raj generalized ordered estimator for sample size n. Hence, we shall define the estimator for a sample of size n = 2.

Let y_1 and y_2 be the units drawn, respectively, at the first and second draws, whose respective initial probabilities are p_1 and p_2. Clearly, y_1; being the unit drawn at the first draw, is a random variable which can assume any of the N values of the study variable Y in the population. Then, let us define variables:

$$z_1 = \frac{y_1}{Np_1} \qquad \text{and} \qquad z_2 = \frac{1}{N}\left[y_1 + \frac{y_2(1-p_1)}{p_2} \right].$$

Then, we see that

$$E(z_1) = E\left(\frac{y_1}{Np_1}\right) = \sum_{i=1}^N \left(\frac{y_i}{Np_i}\right) \times p_i$$

$$= \sum_{i=1}^N \frac{Y_i}{N} = \bar{Y} \text{, the population mean.} \qquad \ldots (5.5)$$

Thus, variable z_1 is an unbiased estimator of population mean.

Similarly, we want to show that variable z_2 is also an unbiased estimator of population mean.

In order to find the expectation of z_2, we shall keep in mind that y_1 is already selected at the first draw and, therefore, the expectation of z_1, a function of y_1 was obtained in (5.5) as $\bar{Y}$. Let us now find the $E(z_2)$. We have

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

In this expression, remember that as y_1 is selected at the first draw, the result of the first draw is a constant for the second draw and the expectation of $\frac{y_2(1-p_1)}{p_2}$ will be a conditional expectation that y_1 is given. Clearly then,

$$E\left\{\frac{y_2(1-p_1)}{p_2}\Big/ y_1\right\} = \sum_{j=1}^{N}\frac{y_j(1-p_1)}{p_j}.\frac{p_j}{(1-p_1)} ,$$

where summation is taken over all the values of Y except the value drawn at the first draw. Therefore,

$$E\left\{\frac{y_2(1-p_1)}{p_2}\Big/ y_1\right\} = \sum_{j=1}^{N}\frac{y_j(1-p_1)}{p_j}.\frac{p_j}{(1-p_1)}.$$

$$= \sum_{j=1}^{N} y_j = \sum_{j=1}^{N} Y_j - y_1 = N\bar{Y} - y_1$$

Hence, we get,

$$E(z_2) = \frac{1}{N}\left[y_1 + N\bar{Y} - y_1\right] = \bar{Y}$$

indicating that variable z_2 is also unbiased for $\bar{Y}$.

Therefore, the sample mean of variables z_1 and z_2, given by

$$\bar{z} = \frac{1}{2}(z_1 + z_2) = \frac{1}{2N}\left[\frac{y_1}{p_1} + \left\{y_1 + \frac{y_2(1-p_1)}{p_2}\right\}\right], \qquad \text{… (5.6)}$$

which is Des Raj's ordered estimator for population mean with n = 2.

Remark 5.3: You can see that how the variate-values are changed in the two draws so as to get unbiased estimators of the parameter under concern at both draws. The variate-value $z_1 = \frac{y_1}{Np_1}$ was so defined for the first draw that its expectation comes out to be equal to $\bar{Y}$. Similarly, knowing that unit y_1 was selected at the first draw, variate-value

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

was designed for the second draw in such a way that its expectation is also equal to $\bar{Y}$. To design complicated forms of variate-values at each draw seems not to be a simple task and, therefore, as the sample size becomes more than 2, designing variate-value for the third and more higher draws becomes very much complicated. This is the reason that ordered estimators did not get much popularity in sampling theory.

5.3.3 Sampling Variance of Des Raj's Sample Mean Estimator

Re-writing the Des Raj's estimator, as given in (5.6), we see that it reduces to

$$\bar{z} = \frac{1}{2N}\left[(1+p_1).\frac{y_1}{p_1} + (1-p_1).\frac{y_2}{p_2}\right] \quad \text{… (5.7)}$$

Therefore,

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

$$= \frac{1}{4N^2}\sum_{i(\neq j)=1}^{N}\left\{(1+p_i).\frac{Y_i}{p_i} + (1-p_i).\frac{Y_j}{p_j}\right\}^2 .\frac{p_i p_j}{(1-p_i)} - \bar{Y}^2 \quad \text{… (5.8)}$$

This is the expression of the sampling variance of the estimator $\bar{z}$. However, it can be further reduced to

$$V(\bar{z}) = \frac{1}{4N^2}\left\{\left(2-\sum_{i=1}^{N}p_i^2\right)\sum_{i=1}^{N}\frac{Y_i^2}{p_i} - \sum_{i=1}^{N}Y_i^2 + 2\left(\sum_{i=1}^{N}Y_i\right)\left(\sum_{i=1}^{N}Y_i p_i\right)\right\} - \frac{\bar{Y}^2}{2}$$

5.3.4 Estimation of Sampling Variance of Des Raj's Sample Mean Estimator

Since, estimator $\bar{z}$ is an unbiased estimator for $\bar{Y}$, so, by definition of variance, we have

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

Now, consider

$$E(z_1 z_2) = E[z_1\{E(z_2|y_1\}] = E[z_1\bar{Y}] = \bar{Y}^2$$

$\Rightarrow z_1 z_2$ is an unbiased estimator of $\bar{Y}^2$.

Therefore, we have

$$\text{Est.}V(\bar{z}) = \hat{V}(\bar{z}) = \bar{z}^2 - z_1 z_2 = \left[\frac{1}{2}(z_1 + z_2)\right]^2 - z_1 z_2$$

$$= \frac{1}{4}\left[z_1^2 + z_2^2 + 2z_1 z_2\right] - z_1 z_2 = \frac{1}{4}(z_1 - z_2)^2$$

$$= \frac{1}{4}\left[\frac{y_1}{Np_1} - \frac{1}{N}\left\{y_1 + \frac{(1-p_1)y_2}{p_2}\right\}\right]^2$$

$$= \frac{(1-p_1)^2}{4N^2}\left[\frac{y_1}{p_1} - \frac{y_2}{p_2}\right]^2; \quad \text{… (5.9)}$$

indicating that

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

is an unbiased estimator of the sampling variance of the estimator $\bar{z}$.

5.3.5 General Form of Des Raj's Estimator

On the basis of the structure of Des Raj's estimator, defined for a sample of size 2, as given in (5.6); it is possible to write the general form of this estimator for a sample of size n. Let $y_1, y_2, \ldots, y_n$ be the units selected, respectively, at the first, second, …, n^{th} draw. Let $p_1, p_2, \ldots, p_n$, respectively, be their initial probability of selection, which are known before the selection process starts. Then, define variables $z_1, z_2, \ldots, z_n$ as follows:

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

$$z_i = \frac{1}{N}\left\{y_1 + y_2 + \ldots + y_{i-1} + y_i\frac{1-p_1-p_2-\ldots-p_{i-1}}{p_i}\right\};\ (i = 1, 2, 3, \ldots, n). \qquad \ldots (5.10)$$

We have already seen that z_1 and z_2 are unbiased estimators of the population mean. From (5.10), we find

$$z_3 = \frac{1}{N}\left\{y_1 + y_2 + y_3\frac{1-p_1-p_2}{p_3}\right\}.$$

Hence, as before, we have

$$E\left(y_3\frac{1-p_1-p_2}{p_3}\,|y_1, y_2\right) = \sum_{j(\neq 1,2)=1}^{N} Y_j\frac{1-p_1-p_2}{p_j}\cdot\frac{p_j}{1-p_1-p_2}$$

$$= N\bar{Y} - y_1 - y_2$$

implying that

$$E(z_3) = \frac{1}{N}\left[y_1 + y_2 + N\bar{Y} - y_1 - y_2\right] = \bar{Y}.$$

Thus, variable z_3 is also an unbiased estimator of $\bar{Y}$.

Similarly, we can prove that

$$E(z_4) = E(z_5) = \ldots = E(z_n) = \bar{Y}.$$

Therefore, the estimator

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

is an unbiased estimator of the population mean.

5.3.6 Das Ordered Estimator

We shall only define here Das Ordered Estimator of population mean without treating it mathematically.

Das defined the ordered estimator for population mean as follows:

Let

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

$$z_i = \frac{y_i(1-p_1)(1-p_1-p_2)\ldots(1-p_1-p_2-\ldots-p_{i-1})}{N(N-1)\ldots(N-i+1)p_1 p_2\ldots p_i}, \text{ for } i = 1, 2,\ldots, n.$$

Then, Das ordered estimator for $\bar{Y}$ is given by

$$\bar{z} = \frac{1}{n}\sum_r^n z_r$$

which is an unbiased estimator.

Remark 5.4: Comparing the Des Raj's estimator as defined in Sub-section (5.3.5) and Das estimator, defined in Sub-section (5.3.6); it is clear that, defining the variate-values for each draw in different manners, a number of unbiased ordered estimators for the population parameter can be obtained.

Example 2: Write down all the possible ordered estimators which are associated with the Probability Proportional to Size without Replacement sample (1, 2, 3, 4).

Solution: Given that the sample consists of units (1, 2, 3, 4). The different possible ordered samples would be obtained by arranging these numbers in all possible ways. Arranging them in order, we have ordered samples as

(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3 , 4, 2), (1, 4, 2, 3), (1, 4, 3, 2), (2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4), (2, 3, 4, 1), (2, 4, 1, 3), (2, 4, 3, 1), (3, 1, 2, 4), (3, 1, 4, 2), (3, 2, 1, 4), (3, 2 4, 1), (3, 4, 1, 2), (3, 4, 2, 1), (4, 1, 2, 3), (4, 1, 3, 2), (4, 2, 1, 3), (4, 2, 3, 1), (4, 3, 1, 2) and (4, 3, 2, 1).

Thus, with 4 units in the sample, we get $4! = 24$ ordered samples.

Now, you may try to answer the following Self-Assessment Question:

SAQ 2

Explain what you mean by an ordered sample and ordered estimator. Define the ordered estimator proposed by Des Raj for a sample of size 2 and n.

5.4 CONCEPT OF UNORDERED ESTIMATORS

We observed in the previous section that ordered estimators are quite complex in nature due to changing variate-values which are to be associated with each draw. The construction of these variate-values at each draw is not a simple task and needs a lot of mathematical skill. Although, many of the ordered estimators for population mean are proposed in literature, but these are not very much popular in practice owing to difficulty in their construction. This problem of ordered estimators enforced a number of sampling researchers to think about the construction of estimators which do not depend upon the ordering of the units, selected in the sample and the same time, are more efficient than any ordered estimator. Fortunately, such estimators, popularly known as **"Unordered Estimators"**, were constructed by some sampling researchers which were found to be simple to handle and were found more effective and efficient when compared with ordered estimators.

In **Remark 5.1** under sub-section 5.2.2, we took an example of ordering of the units drawn in the sample which was drawn using Probability Proportional to Size without Replacement sampling scheme. We did not discuss in detail there about the number of ordered samples with a set of selected units in a Probability Proportional to Size without Replacement sample and the ordered samples we got.

Let from a population of size N, n units be selected in a Probability Proportional to Size without Replacement sample. For simplicity, let the population consists of N units with labels {1, 2, 3, …, N} out of which a Probability Proportional to Size without Replacement sample of size 3 be selected. For instance, let the Probability Proportional to Size without Replacement sample selected be (1, 2, 3). You know that in Simple Random Sampling without Replacement, since all the units are selected in the sample with same selection probabilities, it is immaterial to discuss about the order of the selection of units in the sample. Therefore, sample (1, 2, 3) may also be written as (3, 2, 1), (2, 1 ,3) or (1, 3, 2). But when order does matter, we can see how many orders of same labels be possible.

Let the sample be of size 2, that is, n = 2 and the sample be (1, 2). Obviously, these two labels would generate two ordered samples, given by (1, 2) and (2, 1). Now, if n = 3 with labels (1, 2, 3), then it would generate 6 ordered samples given by (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2) and (3, 2, 1). Similarly, in case n = 4 with labels (1, 2, 3, 4); you can see that there would be 24 ordered samples; 6 samples with 1 at the first place and combinations of other 3 labels at the remaining three places, 6 samples in which 2 occupies the first place and remaining places are filled with possible combinations of the labels 1, 3 and 4, and so on. This would generate 24 ordered samples of size 4 in all. If this argument is continued, it is easy to see that total number of ordered samples of size n would be $n!$ (factorial n). This means that given an unordered sample, say, $(y_1, y_2, \ldots, y_n)$, there would be $n! = n(n-1)(n-2)\ldots$ $\ldots 3.2.1$ associated ordered samples.

5.4.1 Existence of an Unordered Estimator

Before moving ahead for discussing unordered estimators, we shall theoretically prove an important result for the existence of an unordered estimators.

Theorem 2: Corresponding to any ordered estimator, there exist an unordered estimator, which does not depend on the order in which the units are drawn, and which is more efficient than the ordered estimator in the sense that it has a smaller variance.

Proof: Since, the PPS sample of size n is selected from the population of size N under without replacement scheme, the total number of samples would be

$\binom{N}{n}$, due to Theorem 3 of sub-section 2.3 of the Unit 2. Let these samples be denoted by $s_1, s_2, \ldots, s_k, \ldots, s_{\binom{N}{n}}$; where, $k = 1, 2, \ldots, \binom{N}{n}$.

Clearly, each of these samples can be ordered in $n!$ (factorial n) ordered samples. Let $j = 1, 2, \ldots, n!$.

Let we wish to estimate a population parameter using these samples. Let the estimator used for the j^{th} ordering of the k^{th} sample be denoted by $T\left(s_k^{(j)}\right)$. Further, let $p\left(s_k^{(j)}\right)$ stands for the probability of getting the j^{th} ordering of the k^{th} sample and $p(s_k)$ stands for the probability of getting k^{th} sample. Clearly, then we have the relation

$$p(s_k) = \sum_{j=1}^{n!} p\left(s_k^{(j)}\right); \qquad \ldots (5.11)$$

since, the selection probability of the k^{th} sample would be the sum of selection probabilities of all the ordered samples generated by the k^{th} sample.

Let us denote by T(u) an unordered estimator of the concerned parameter and define it as

$$T(u) = \sum_{j=1}^{n!} T\left(s_k^{(j)}\right) \cdot \frac{p\left(s_k^{(j)}\right)}{p(s_k)} = \frac{1}{p(s_k)} \sum_{j=1}^{n!} T\left(s_k^{(j)}\right) . p\left(s_k^{(j)}\right). \qquad \ldots (5.12)$$

We can see that $T(u)$ is a weighted average of the estimator $T\left(s_k^{(j)}\right)$; weights being equal to $p\left(s_k^{(j)}\right) \Big/ \sum_{j=1}^{n!} p\left(s_k^{(j)}\right)$.

Let us find the expectation of the estimator T(u). We have,

$$E\left[T(u)\right] = E\left[\sum_{j=1}^{n!} T\left(s_k^{(j)}\right) \cdot \frac{p\left(s_k^{(j)}\right)}{p(s_k)}\right] = \sum_{k=1}^{\binom{N}{n}} \left[\sum_{j=1}^{n!} T\left(s_k^{(j)}\right) \cdot \frac{p\left(s_k^{(j)}\right)}{p(s_k)}\right] p(s_k)$$

$$= \sum_{k=1}^{\binom{N}{n}} \sum_{j=1}^{n!} T\left(s_k^{(j)}\right) p\left(s_k^{(j)}\right) = E\left[T\left(s_k^{(j)}\right)\right].$$

Therefore, we have the result that

$$E\left[T(u)\right] = E\left[T\left(s_k^{(j)}\right)\right] \qquad \ldots (5.13)$$

Now, we find the variance of the estimator $T\left(s_k^{(j)}\right)$. By definition of variance, we see that

$$V\left[T\left(s_k^{(j)}\right)\right] = E\left[T\left(s_k^{(j)}\right) - E\left\{T\left(s_k^{(j)}\right)\right\}\right]^2 = E\left[\left\{T\left(s_k^{(j)}\right)\right\}^2\right] - \left[E\left\{T\left(s_k^{(j)}\right)\right\}\right]^2$$

$$= \sum_{k=1}^{\binom{N}{n}} \sum_{j=1}^{n!} \left\{T\left(s_k^{(j)}\right)\right\}^2 p\left(s_k^{(j)}\right) - \left[\sum_{k=1}^{\binom{N}{n}} \sum_{j=1}^{n!} T\left(s_k^{(j)}\right) p\left(s_k^{(j)}\right)\right]^2. \qquad \ldots (5.14)$$

Obviously, (5.14) provides us the sampling variance of the ordered estimator $T\left(s_k^{(j)}\right)$, based upon the j^{th} ordering of the k^{th} sample.

In order to compare this variance with the variance of the unordered estimator, T(u), we find the expression of $V\left[T(u)\right]$. We have,

$$V\left[T(u)\right]=E\left[T(u)-E\left\{T(u)\right\}\right]^2$$

$$=E\left[\left\{T(u)\right\}^2\right]-\left[E\left\{T(u)\right\}\right]^2$$

$$=\sum_{k=1}^{\binom{N}{n}}\left\{\sum_{j=1}^{n!}T\left(s_k^{(j)}\right).\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}\right\}^2 p\left(s_k\right)-\left[\sum_{k=1}^{\binom{N}{n}}\sum_{j=1}^{n!}T\left(s_k^{(j)}\right)p\left(s_k^{(j)}\right)\right]^2$$

… (5.15)

Hence, from (5.14) and (5.15), we see that

$$V\left[T\left(s_k^{(j)}\right)\right]-V\left[T(u)\right]$$

$$=\sum_{k=1}^{\binom{N}{n}}\sum_{j=1}^{n!}\left\{T\left(s_k^{(j)}\right)\right\}^2 p\left(s_k^{(j)}\right)-\sum_{k=1}^{\binom{N}{n}}\left\{\sum_{j=1}^{n!}T\left(s_k^{(j)}\right).\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}\right\}^2 p\left(s_k\right)$$

$$=\sum_{k=1}^{\binom{N}{n}}p\left(s_k\right)\left[\sum_{j=1}^{n!}\left\{T\left(s_k^{(j)}\right)\right\}^2\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}-\left\{\sum_{j=1}^{n!}T\left(s_k^{(j)}\right).\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}\right\}^2\right]$$

$$=\sum_{k=1}^{\binom{N}{n}}p\left(s_k\right)\sum_{j=1}^{n!}\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}\left[T\left(s_k^{(j)}\right)-\left\{\sum_{j=1}^{n!}T\left(s_k^{(j)}\right).\frac{p\left(s_k^{(j)}\right)}{p\left(s_k\right)}\right\}\right]^2$$

$$=\sum_{k=1}^{\binom{N}{n}}\sum_{j=1}^{n!}p\left(s_k^{(j)}\right)\left[T\left(s_k^{(j)}\right)-T(u)\right]^2>0 \text{ always.}$$

This shows that

$$V\left[T(u)\right]<V\left[T\left(s_k^{(j)}\right)\right].$$

Therefore, we can conclude that the variance of an unordered estimator of any population parameter is always smaller than the ordered estimator which can be generated out of the given unordered estimator.

Remark 5.5: The **Theorem 2** provides a general result for any type of estimator of a population parameter. In fact, based on it, we can construct unordered estimator on the basis of given ordered estimators. As an illustration, we shall show in the next Sub-section how to obtain an unordered estimator for population mean with the help of Des Raj's ordered estimators for n = 2.

5.4.2 Constructing Unordered Estimator Using Des Raj's Ordered Estimators

Consider a sample of size 2, drawn from a population using Probability Proportional to Size without Replacement sampling scheme. Let the units in the sample be (y_i, y_j). Then ordering these units in order, we get the two ordered estimators as: (y_i, y_j) and (y_j, y_i). Let us write the Des Raj's estimators for these ordering. On the basis of the discussions made in the Sub-section 5.3.2, for the first sample, we get the variables z_1 and z_2 as follows:

Sample (y_i, y_j), where i ≠ j.

$$z_1 = \frac{y_i}{Np_i} \text{ and } z_2 = \frac{1}{N}\left[y_i + \frac{y_j(1-p_i)}{p_j}\right],$$

so from (5.6) we get the corresponding estimator

$$\bar{Z}_1 = \frac{1}{2N}\left[(1+p_i).\frac{y_i}{p_i} + (1-p_i).\frac{y_j}{p_j}\right]. \qquad \dots (5.16)$$

Sample (y_j, y_i)

For this ordering, we have

$$z_1 = \frac{y_j}{Np_j} \text{ and } z_2 = \frac{1}{N}\left[y_j + \frac{y_i(1-p_j)}{p_i}\right],$$

so that the estimator is

$$\bar{Z}_2 = \frac{1}{2N}\left[(1+p_j).\frac{y_j}{p_j} + (1-p_j).\frac{y_i}{p_i}\right] \qquad \dots (5.17)$$

We now find the probability of getting the j^{th} ordering of the k^{th} sample, $p\left(s_k^{(j)}\right)$. Here, we see that j = 1, 2; therefore, let j =1 means the ordering (y_i, y_j) and j = 2 means the ordering (y_j, y_i). Clearly, then

$$p\left(s_k^{(1)}\right) = \frac{p_i p_j}{(1-p_i)} \text{ and } p\left(s_k^{(2)}\right) = \frac{p_j p_i}{(1-p_j)}.$$

Further, we get probability of getting the k^{th} sample (see (5.11)), given by

$$p(s_k) = \sum_{j=1}^{n!} p\left(s_k^{(j)}\right) = p\left(s_k^{(1)}\right) + p\left(s_k^{(2)}\right) = p_i p_j\left[\frac{1}{1-p_i} + \frac{1}{1-p_j}\right].$$

Since, in variance expressions, we need the quantity $\frac{p\left(s_k^{(j)}\right)}{p(s_k)}$, it can be obtained for j = 1 and 2, using $p\left(s_k^{(1)}\right)$, $p\left(s_k^{(2)}\right)$ and $p(s_k)$. We get

$$\frac{p\left(s_k^{(1)}\right)}{p(s_k)} = \frac{(1-p_j)}{(2-p_i-p_j)} \text{ and } \frac{p\left(s_k^{(2)}\right)}{p(s_k)} = \frac{(1-p_i)}{(2-p_i-p_j)}.$$

Hence, the unordered estimator corresponding to Des Raj's ordered estimators will be

$$T(u) = \sum_{j=1}^{2} \bar{\bar{Z}}_j \cdot \frac{p\left(s_k^{(j)}\right)}{p(s_k)}$$

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

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

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

Simplifying this expression, it can be reduced to

$$T(u) = \frac{1}{N(2-p_i-p_j)}\left[(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right]. \qquad \dots (5.18)$$

This is Murthy (1957) unordered estimator for population parameter with n = 2. Similarly, one can construct an unordered estimator for population parameter with n = 2 corresponding to Das (1951) ordered estimator as given in the Sub-section 5.3.6.

The sampling variances of the ordered estimators; that is, $V\left(\bar{\bar{Z}}_1\right)$ and $V\left(\bar{\bar{Z}}_2\right)$ can be obtained as described in the Sub-section 5.3.3. However, finding the expression of the sampling variance of the unordered estimator T(u) is not a simple task, but if obtained by consuming enough time and patience, a comparison of V[T(u)] with $V\left(\bar{\bar{Z}}_1\right)$ or $V\left(\bar{\bar{Z}}_2\right)$ will reveal the fact that

$$V\left[T(u)\right] < V\left[T\left(\bar{\bar{Z}}_j\right)\right], \quad \text{where, } j = 1, 2.$$

Remark 5.6: The Sub-section 5.4.2 illustrates the procedure of constructing an unordered estimator corresponding to ordered estimators based on a sample of fixed size. However, the process of construction of unordered estimator involves a lot of mathematical complexity, and therefore, is not very much used in the literature of the sampling theory.

Try to answer of the following Self-Assessment Question on the basis of contents discussed in this section:

SAQ 3

Distinguish between ordered and unordered samples and explain the meaning of an unordered estimator.

5.5 INCLUSION PROBABILITIES

In this section, we shall discuss the concept of another special type of probability associated with a specific unit of the population, namely, **"Inclusion Probability"**, which has different concept than the **"Initial Probability of Selection"**; which we defined and frequently used in Units IV and V.

You know that whether it is Equal Probability Selection Method (EPSEM) or Varying Probability Selection Scheme (VPSS); in both the cases, each unit of the population is assigned some probability of selection in the sample, before starting the selection process, which is generated on the basis of the chance mechanism used for selecting a random sample. Whereas, in SRSWR and SRSWOR sampling schemes, all the units of the population are assigned same probability of selection, equal to 1/N; in PPS sampling schemes, these probabilities are unequal over units. Such a probability, either it is EPSEM or VPSS scheme, which is decided and assigned to each of the population units before starting the selection process and is realized through the chance mechanism used, is popularly termed as 'Initial Probability of Selection of Units'. In PPS schemes, you know that initial probability of selection of a particular unit, say, i^{th} unit, denoted by p_i is proportional to its size measure.

Let us now consider Theorem 4, in Sub-section 1.6.5 in Unit 1 and Theorem 2, in Section 2.3 of Unit 2. Both the theorems tell us about a new type of probability attached to each unit of the population. With population size N and sample size n, this probability has been shown to be n/N in schemes SRSWR and SRSWOR both. According to the statement of the theorems, it is the probability of inclusion in the sample of a specified unit with which it is included in the sample of a given size. This probability of the unit is popularly known as **"Inclusion Probability of the Unit"**. Obviously, it is different from the **"Initial Probability of Selection of the Unit"**.

Since, in Probability Proportional to Size with and without Replacement schemes, the probability of selection of a specific unit varies from draw to draw, the computation of the probability of inclusion of that particular unit in a sample of size n is not so easy in PPS schemes as it is in case of SRS schemes. Moreover, it is too difficult for samples which are ordered in different manners in order to define ordered estimators. However, in samples which are unordered, the computation of inclusion probability of a specific unit might be possible.

In fact, the concept of inclusion probability plays an important role for defining different unordered estimators for population parameters. Therefore, before proceeding to the discussion on unordered estimators of certain population parameter; we shall explain different aspects of inclusion probabilities.

5.5.1 Order of an Inclusion Probability

The inclusion probabilities may be of different orders.

(i) **First-Order Inclusion Probability:** The probability that an element (unit), say, k^{th} unit, will be included in the sample of given size, is known as "First-Order Inclusion Probability". Generally, it is denoted by the notation π_k. If the population size is N, then, we get N first-order inclusion probabilities; $\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_N$.

(ii) **Second-Order Inclusion Probability:** The probability that two different units of the population, say, p^{th} and k^{th}, $(p \neq k)$, will be included in the sample, is known as *"Second-Order Inclusion Probability". The second-order inclusion probabilities are denoted by* $\pi_{12}, \pi_{13}, \ldots, \pi_{pk}, \ldots, \pi_{N-1,N}$. We observe that $\pi_{12} = \pi_{21}, \pi_{13} = \pi_{31}, \ldots, \pi_{pk} = \pi_{kp}, \pi_{N-1,N} = \pi_{N,N-1}$, etc. Thus, for N units, we get in total $\frac{N(N-1)}{2}$ second-order inclusion probabilities.

Inclusion probabilities of higher order than two can be defined in similar way. However, they play a less important role in the literature and, hence, need not to be discussed here.

5.5.2 Some Important Results Related to Inclusion Probabilities

In order to prove some important results for inclusion probabilities, we shall use the following notations with their meanings:

Ω: The population under consideration.

S: The sample selected.

$i, j, k \in \Omega$: Denoting that i^{th}, j^{th}, k^{th} units belong to the population.

$i, j, k \in S$: Denoting that i^{th}, j^{th}, k^{th} units belong to the sample.

$\sum_{i \in \Omega} X_i$: Denoting that the sum of all the X_i – values is taken over the population.

$\sum_{i \in S} x_i$: Denoting that the sum of all the x_i – values is taken over the sample.

Result 1: We have

$$\sum_{i \in \Omega} \pi_i = n. \qquad \ldots (5.19)$$

Proof: The statement of the result is that the sum of all the first-order probabilities taken over all the units of the population is n, the sample size.

Let us define a variable t_i such that it assumes value 1 if the i^{th} unit of the population is included in the sample and assumes value zero if it is not included in the sample. This means that

$$t_i = \begin{cases} 1, \text{ if } i \in S \text{ with probability } \pi_i; \\ 0, \text{ if } i \notin S \text{ with probability } (1-\pi_i). \end{cases}$$

Therefore, t_i is a random variable assuming values 1 and 0 with respective probabilities π_i and $(1-\pi_i)$. Hence, we have expectation of the variable as

$$E(t_i) = 1 \times \pi_i + 0 \times (1-\pi_i) = \pi_i.$$

Since, the sampling scheme is PPSWOR, so unit selected once is not replaced in the population before the next draw. Further, as the sample size is n, there would be n units which would be selected with probability π_i and rest of the (N – n) units would not be selected (included) in the sample. This shows that

$$\sum_{k\in\Omega} t_i = \underbrace{1+1+0+1+0+0+\ldots+1+0}_{\text{n ones and }(N-n)\text{ zeros}} = n$$

Therefore, we have

$$E\left(\sum_{i\in\Omega} t_i\right) = E(n) \Rightarrow \sum_{i\in\Omega} E(t_i) = n \Rightarrow \sum_{i\in\Omega} \pi_i = n.$$

Hence the result.

Result 2: We have

$$\sum_{j(\neq i)\in\Omega} \pi_{ij} = (n-1)\pi_i. \qquad \ldots (5.20)$$

Proof: Let us consider a variable t_{ij} which assumes value 1 if both the i^{th} and j^{th} units of the population are included in the sample and assumes value zero otherwise. Therefore, we have

$t_{ij} = 1,$ if $i, j \in S$ with probability π_{ij};

$0,$ if $i, j \notin S$ with probability $(1-\pi_{ij})$.

Obviously, then

$$E(t_{ij}) = 1\times\pi_{ij} + 0\times(1-\pi_{ij}) = \pi_{ij}.$$

Further, we see that

$$E\left(\sum_{j(\neq i)\in\Omega} t_{ij}\right) = \sum_{j(\neq i)\in\Omega} E(t_{ij}) = \sum_{j(\neq i)\in\Omega} \pi_{ij}.$$

Now, consider $\sum_{j(\neq i)\in\Omega} \pi_{ij}$.

We see that

$$\sum_{j(\neq i)\in\Omega} \pi_{ij} = \sum_{j(\neq i)\in\Omega} \left[P\left\{\text{both } i^{th} \text{ and } j^{th} \text{ are included in the sample}\right\}\right]$$

$$= \sum_{j(\neq i)\in\Omega} P[\, j\in S \mid i\in S]\, P(i\in S), \text{ using multiplicative law of probability}$$

$$= \sum_{j(\neq i)\in\Omega} P[\, j\in S \mid i\in S]\,\pi_i = \pi_i \sum_{j(\neq i)\in\Omega} P[\, j\in S \mid i\in S]$$

It can be seen that in $\sum_{j(\neq i)\in\Omega} P[\, j\in S \mid i\in S]$, the number of population units would be (N – 1). Since, it is given that i^{th} unit is already selected in the sample, therefore, we have to select further (n – 1) units in the sample.

Thus, we have

$$\sum_{j(\neq i)\in\Omega} P[\, j\in S \mid i\in S] = (n-1).$$

Therefore, $\sum_{j(\neq i)\in\Omega} \pi_{ij} = (n-1)\pi_i.$

Hence, the result.

Remark 5.7: We have seen that Theorem 4, Sub-section 1.6.5 in Unit 1 and Theorem 2, in the Section 2.3 of Unit 2 showed that the inclusion probability of a particular unit of the population in SRSWR and SRSWOR schemes is given by n/N.

Therefore, as per above notation, it is the first order inclusion probability and, hence, we write $\pi_i = n/N$.

Clearly, then

$$\sum_{i\in\Omega}\pi_i = \underbrace{n/N + n/N + \ldots n/N}_{N \text{ times}} = n$$

So, the Result 1 is verified in case of SRSWR and SRSWOR sampling schemes.

Remark 5.8: The first-order inclusion probability, π_i, in SRSWOR sampling schemes can be obtained in another way. If sample size is n and the i^{th} unit must be included in all the samples of size n, we can say that one place in the samples is reserved for the i^{th} unit and rest of the (n –1) places are to be filled by (N –1) units of the population, except the i^{th} unit. Using the theory of permutation and combination we know that the number of ways of filling (n –1) places by (N –1) units is $\binom{N-1}{n-1}$.

Therefore, there will be $\binom{N-1}{n-1}$ samples which includes the i^{th} unit. Since, all the samples of size n have the same probability of selection given by $\frac{1}{\binom{N}{n}}$; therefore, the probability that i^{th} unit is included in the sample will be given by $\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$ which is π_i.

Remark 5.9: Let us now consider the second-order inclusion probability in SRSWOR scheme. If two units, say i^{th} and j^{th}, such that $i \neq j$; are to be included in the samples selected, we know that the number of samples of size n, including necessarily i^{th} and j^{th} units would be $\binom{N-2}{n-2}$.

Therefore, the probability that i^{th} and j^{th} units are included in the sample will be given by $\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$.

Which is π_{ij}, that is, $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$.

Example 3: Given the population of size 5 with units (A, B, C, D, E), a sample of size 3 has to be selected using without replacement sampling scheme.

(i) List all the possible samples of size 3.

(ii) Find the first order inclusion probabilities for all the 5 units of the population.

(iii) Find the second-order inclusion probabilities such that unit B and E are included in the sample.

Solution:

(i) Given the population (A, B, C, D, E) we have to select samples of size 3. Clearly, there would be 10 samples given by (A, B, C), (A, B, D), (A, B, E), (A, C, D), (A, C, E), (A, D, E), (B, C, D), (B, C, E), (B, D, E) and (C, D, E).

(ii) The first order inclusion probabilities for all the five units can be obtained as follows:

Since all the 10 samples are drawn with same probability of selection given by 1/10 = 0.1, the first order inclusion probability for the element A is given by 0.1 +0.1 +0.1 +0.1 +0.1+0.1 = 0.6, since only 6 samples include the element A. These samples are (A, B, C), (A, B, D), (A, B, E), (A, C, D), (A, C, E) and (A, D, E).

Similarly, it can be seen that all the remaining elements B, C, D and E are included in only 6 samples.

Therefore, the inclusion probability of elements B, C, D and E are also 0.6. So, we have $\pi_A = \pi_B = \pi_C = \pi_D = \pi_E = 0.6$.

(iii) We see that elements B and E both are included in the samples (A, B, E), (B, C, E) and (B, D, E), the second order probability:

$$\pi_{BE} = 0.1 + 0.1 + 0.1 = 0.3.$$

Now, you may try to answer the following Self-Assessment Question:

SAQ 4

What do you understand by "inclusion probability" of units of the population? Explain the difference between the inclusion probability and initial probability of selection of units. Elaborate your answer in case of SRSWOR sampling scheme.

5.6 HORVITZ-THOMPSON ESTIMATOR

Under Section 5.4, we observed that in Probability Proportional to Size Without Replacement sampling scheme, there exist some unordered estimators of population parameters which are superior than the ordered estimators in the sense that the efficiency of the former type estimators are more than the latter type estimators. Further, we also observed that ordered estimators have very limited applicability, since their construction is very complex and, hence, their treatment is almost unmanageable even with a moderate size sample. In the search of unordered estimators, many of the statisticians, therefore, tried to propose different unordered estimators, such as, Horvitz-Thompson (1952), Murthy (1957), Basu (1958) estimators. In Sub-section 5.4.2, for a sample of size 2, we have already seen how to construct an unordered estimator for population mean on the basis of Des Raj's ordered estimator. It is given in the expression (5.18) and was proposed by Murthy (1957).

Since, Horvitz- Thompson (1952) unordered estimator exhibits a number of unique properties, it is the most popular unordered estimator. We shall describe and discuss this estimator in this section and shall derive some of its important properties.

5.6.1 Horvitz-Thompson Estimator for Population Total

Horvitz –Thompson suggested an unbiased estimator for the population total. It is a linear estimator of sample observations. On the basis of n sample observations, $y_i; i = 1, 2, \ldots, n$, the Horvitz-Thompson Estimator (HTE) of population total is defined as

$$\hat{Y}_{HT} = \sum_{i \in S} d_i y_i; \qquad \ldots (5.21)$$

where d_i, $i = 1, 2, \ldots, n$ are pre-determined real constants or design weights. Let us find the values of the constants d_i, so that the estimator $\hat{Y}_{HT}$ is unbiased estimator for the population total. For this, we prove the following theorem:

Theorem 3: *For the estimator* $\hat{Y}_{HT}$ *to be unbiased for the population total,* $d_i = \frac{1}{\pi_i}, (i = 1, 2, \ldots, N)$.

Proof: Using the variable t_i as used in Result 1 under the Sub-section 5.5.2, we see that

$$E\left[\hat{Y}_{HT}\right] = E\left[\sum_{i \in S} d_i y_i\right] = E\left[\sum_{i \in \Omega} t_i d_i Y_i\right].$$

Note that here d_i and Y_i are constants. Now, we see that

$$E\left[\hat{Y}_{HT}\right] = \sum_{i \in \Omega} d_i Y_i E(t_i) = \sum_{i \in \Omega} d_i Y_i \pi_i.$$

In the above expression, if we choose

$$d_i = \frac{1}{\pi_i},$$

then we have

$$E\left[\hat{Y}_{HT}\right] = \sum_{i \in \Omega} Y_i = Y$$

where Y stands for the population total. This implies that the Horvitz-Thompson, estimator $\hat{Y}_{HT}$ is an unbiased estimator of population total, Y only if,

$$d_i = \frac{1}{\pi_i} \text{ for } (i = 1, 2, \ldots, N).$$

The Horvitz-Thompson estimator then, becomes

$$\hat{Y}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i}. \qquad \dots (5.22)$$

This establishes the Theorem.

Remark 5.10: As a particular case of the estimator $\hat{Y}_{HT}$ for population total, we can consider the SRSWOR sampling scheme. We know that in SRSWOR scheme, $\pi_i = \frac{n}{N}$ for all i (see the Theorem 2 under Section 2.3 in Unit 2).

Therefore, $\hat{Y}_{HT}$ becomes

$$\hat{Y}_{HT} = \sum_{i \in S} \frac{y_i}{\pi_i} = \sum_{i \in S} y_i . \frac{N}{n} = N \sum_{i \in S} \frac{y_i}{n} = N\bar{y} \qquad \dots (5.23)$$

which is the same result as proved in Theorem 6 under Sub-section 2.4.2 of Unit 2. Therefore, the Horvitz-Thompson estimator of population total under SRSWOR scheme is given by $N\bar{y}$.

5.6.2 Sampling Variance of the Horvitz-Thompson Estimator for Population Total

After defining the Horvitz-Thompson Estimator for population total, we can show the following result:

Theorem 4: The sampling variance of the estimator $\hat{Y}_{HT}$ is given by

$$V\left(\hat{Y}_{HT}\right) = \sum_{i \in \Omega} \frac{(1-\pi_i)}{\pi_i} . Y_i^2 + \sum_{i \in \Omega} \sum_{j(\neq i) \in \Omega} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} . Y_i Y_j \qquad \dots (5.24)$$

Proof: Let us define once again the random variable t_i as follows:

t_i = 1, if $i \in S$ with probability π_i;

0, if $i \notin S$ with probability $(1-\pi_i)$

We have

$$V\left(\hat{Y}_{HT}\right) = V\left[\sum_{i \in S} \frac{y_i}{\pi_i}\right] = V\left[\sum_{i \in \Omega} \frac{t_i Y_i}{\pi_i}\right]; \qquad \dots (5.25)$$

since $t_i = 1$ only for those units of the population which are included in the sample and zero for other units.

Since, the sample is taken without replacing the units in the population after selection, so random variables t_i and t_j for $i \neq j$ are not independent to each other.

In such case, we know that variance of the sum of variables $X_1, X_2, \dots, X_k$; that is,

$$V\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} V(X_i) + \sum_{i=1}^{k} \sum_{j(\neq)=1}^{k} Cov(X_i , X_j)$$

Applying this rule of sum of variables, we see that (5.25) is expanded as

$$V\left(\hat{Y}_{HT}\right)=\sum_{i\in\Omega}V\left[\frac{t_iY_i}{\pi_i}\right]+\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}Cov\left[\frac{t_iY_i}{\pi_i}\ ,\ \frac{t_jY_j}{\pi_j}\right] \qquad \ldots (5.26)$$

Let us now find the terms $V\left(t_i\right)$ and $cov\left(t_i,\ t_j\right)$. We have

$$V\left(t_i\right)=E\left[t_i-E\left\{t_i\right\}\right]^2=E\left(t_i^2\right)-\left\{E\left(t_i\right)\right\}^2$$

But, since,

$$E\left(t_i^2\right)=1\times\pi_i+0\times\left(1-\pi_i\right)=\pi_i \quad \text{and} \quad \left\{E\left(t_i\right)\right\}^2=\pi_i^2,$$

we have

$$V\left(t_i\right)=\pi_i\left(1-\pi_i\right).$$

Similarly, we see that

$$Cov\left(t_i\ ,\ t_j\right)=E\left[\left\{t_i-E\left\{t_i\right\}\right\}\left\{t_j-E\left\{t_j\right\}\right\}\right]$$

$$=E\left(t_it_j\right)-\left\{E\left(t_i\right)\right\}\left\{E\left(t_j\right)\right\}$$

$$=E\left(t_it_j\right)-\pi_i\pi_j.$$

But, since $E\left(t_it_j\right)=1\times\pi_{ij}+0\left(1-\pi_{ij}\right)=\pi_{ij};$

we have

$$Cov\left(t_i\ \ t_j\right)=\left(\pi_{ij}-\pi_i\pi_j\right).$$

Therefore,

$$V\left(\hat{Y}_{HT}\right)=\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i^2}V\left(t_i\right)+\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\frac{Y_iY_j}{\pi_i\pi_j}Cov(t_i\ \ t_j)$$

$$=\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i^2}.\pi_i\left(1-\pi_i\right)+\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\frac{Y_iY_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j).$$

After simplification, we get

$$V\left(\hat{Y}_{HT}\right)=\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i}.\left(1-\pi_i\right)+\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\frac{Y_iY_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j).$$

Hence the theorem.

Remark 5.11: Under SRSWOR sampling scheme reduces to

$$V\left(\hat{Y}_{HT}\right)=N^2\frac{(N-n)}{Nn}S^2;$$

which is same as

$$V\left(N\bar{y}\right)=N^2V\left(\bar{y}\right)=N^2\frac{N-n)}{Nn}S^2;$$

the sampling variance of the estimator of population total under SRSWOR scheme. We can show this result as follows:

We know that in Simple Random Sampling without Replacement scheme

$$\pi_i = \frac{n}{N} \text{ and } \pi_{ij} = \frac{n(n-1)}{N(N-1)} \qquad \text{(See Remarks 5.8 and 5.9 of this unit).}$$

Therefore, by substituting these values of π_i and π_{ij} in the expression (5.24), we have

$$V\left(\hat{Y}_{HT}\right)_{SRSWOR} = \sum_{i\in\Omega} \frac{\left(1-n/N\right)}{n/N} Y_i^2 + \sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega} \frac{\frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N}}{\frac{n}{N}\frac{n}{N}} . Y_i Y_j$$

$$= \frac{(N-n)}{n}\sum_{i\in\Omega} Y_i^2 - \frac{(N-n)}{n(N-1)}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega} Y_i Y_j$$

But from the relation

$$Y^2 = \left(\sum_{i\in\Omega} Y_i\right)^2 = \sum_{i\in\Omega} Y_i^2 + \sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega} Y_i Y_j$$

we have

$$\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega} Y_i Y_j = Y^2 - \sum_{i\in\Omega} Y_i^2$$

Therefore, we have

$$V\left(\hat{Y}_{HT}\right)_{SRSWOR} = \frac{(N-n)}{n}\sum_{i\in\Omega} Y_i^2 - \frac{(N-n)}{n(N-1)}\left\{Y^2 - \sum_{i\in\Omega} Y_i^2\right\}$$

$$= \frac{(N-n)}{n}\left[\sum_{i\in\Omega} Y_i^2 - \frac{1}{(N-1)}\left\{Y^2 - \sum_{i\in\Omega} Y_i^2\right\}\right]$$

$$= \frac{(N-n)}{n(N-1)}\left[(N-1)\sum_{i\in\Omega} Y_i^2 - Y^2 + \sum_{i\in\Omega} Y_i^2\right]$$

$$= \frac{(N-n)}{n(N-1)}\left[N\sum_{i\in\Omega} Y_i^2 - Y^2\right]$$

$$= \frac{(N-n)}{n(N-1)}\left[N\sum_{i\in\Omega} Y_i^2 - \left(\sum_{i\in\Omega} Y_i\right)^2\right]$$

$$= \frac{(N-n)}{n(N-1)}\left[N\sum_{i\in\Omega} Y_i^2 - \left(N\bar{Y}\right)^2\right]$$

$$= \frac{(N-n)}{n(N-1)}.N\left[\sum_{i\in\Omega} Y_i^2 - N\bar{Y}^2\right]$$

$$= \frac{(N-n)N}{n}.\frac{1}{N-1}\left[\sum_{i\in\Omega}\left(Y_i - \bar{Y}\right)^2\right]$$

$$= \frac{(N-n)N}{n} S^2 = N^2\frac{(N-n)}{nN} S^2 = V\left(N\bar{y}\right);$$

the variance of the estimator of population total in Simple Random Sampling without Replacement scheme. Thus, the proof of the result is complete.

5.6.3 Another Form of the Sampling Variance of the Horvitz–Thompson Estimator for Population Total

We obtained the expression of sampling variance of the Horvitz-Thompson estimator of population total in Theorem 4 under the Sub-section 5.6.2. It was derived by Horvitz-Thompson themselves. Henceforth, we shall denote this expression by $V\left(\hat{Y}_{HT}\right)_{HT}$. However, Yates and Grundy (1953) and Sen (1953) independently developed another form of $V\left(\hat{Y}_{HT}\right)_{HT}$, which we shall denote by $V\left(\hat{Y}_{HT}\right)_{YG,S}$. We shall derive the expression of $V\left(\hat{Y}_{HT}\right)_{YG,S}$ in the following theorem:

Theorem 5: The expression of the variance of the Horvitz-Thompson estimator, as obtained by Yates and Grundy and Sen independently, is given by

$$V\left(\hat{Y}_{HT}\right)_{YG,S}=\frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right]^2. \qquad \ldots (5.27)$$

Proof: We have

$$V\left(\hat{Y}_{HT}\right)_{YG,S}=\frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right]^2.$$

$$=\frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i^2}{\pi_i^2}+\frac{Y_j^2}{\pi_j^2}-2\frac{Y_iY_j}{\pi_i\pi_j}\right]$$

$$=\frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i^2}{\pi_i^2}+\frac{Y_j^2}{\pi_j^2}\right]-\frac{1}{2}.2\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\frac{Y_iY_j}{\pi_i\pi_j}$$

$$=\frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i^2}{\pi_i^2}+\frac{Y_j^2}{\pi_j^2}\right]+\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}. \qquad \ldots (5.28)$$

It is easy to see that after some adjustments, the first term of (5.28) can be reduced to

$$\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\frac{Y_i^2}{\pi_i^2}=\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i^2}\pi_i\sum_{j(\neq i)\in\Omega}\pi_j-\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i^2}\sum_{j(\neq i)\in\Omega}\pi_{ij}$$

$$=\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i}\left(n-\pi_i\right)-\sum_{i\in\Omega}\frac{Y_i^2}{\pi_i^2}\left(n-1\right)\pi_i;$$

since, $\displaystyle\sum_{j(\neq i)\in\Omega}\pi_j=\sum_{j\in\Omega}\pi_j-\pi_i=\left(n-\pi_i\right)$

and $\displaystyle\sum_{j(\neq i)\in\Omega}\pi_{ij}=\left(n-1\right)\pi_i.$

Therefore, we have (5.28) as

$$V\left(\hat{Y}_{HT}\right)_{YG,S} = \sum_{i\in\Omega}\frac{Y_i^2}{\pi_i}\left(n-\pi_i\right) - \sum_{i\in\Omega}\frac{Y_i^2}{\pi_i}\left(n-1\right) + \sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}\ .$$

$$= \sum_{i\in\Omega}\frac{Y_i^2}{\pi_i} - \sum_{i\in\Omega}Y_i^2 + \sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}$$

$$= \sum_{i\in\Omega}(1-\pi_i)\frac{Y_i^2}{\pi_i} + \sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)Y_iY_j = V\left(\hat{Y}_{HT}\right)_{HT}\ .$$

Thus, $V\left(\hat{Y}_{HT}\right)_{YG,S}$ is an altered form of the $V\left(\hat{Y}_{HT}\right)_{HT}$.

Remark 5.12: It can be seen that the variance expression given in (5.27) becomes zero if $\pi_i \propto Y_i$ for all i, that is, if first order inclusion probability π_i is directly proportional to the value of the study variable. This is because then $\pi_i = kY_i$, where k is the constant of proportionality.

Then we have, $\sum_{i\in\Omega}\pi_i = k\sum_{i\in\Omega}Y_i \Rightarrow n = kY \Rightarrow k = n/Y$

Thus, $\pi_i = \frac{n}{Y}Y_i$, for all $i = 1, 2, \ldots, N$.

Substituting this value of π_i and $\pi_j = \frac{n}{Y}Y_j$ in (5.27), we see that

$$V\left(\hat{Y}_{HT}\right)_{YG,S} = \frac{1}{2}\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left[\frac{Y_i.Y}{nY_i} - \frac{Y_j.Y}{nY_j}\right]^2 = 0.$$

This indicates that the ideal choice of inclusion probabilities would be $\pi_i \propto Y_i$, so that the estimator $\hat{Y}_{HT}$ would have no variance and the estimator would be the best one. But this is a hypothetical situation, as, Y_i values; being the population values, are never known.

5.6.4 Unbiased Estimators of Variance of the Estimators of Population Total

Since, the variance of the estimator $\hat{Y}_{HT}$ depends upon the population values Y_i, Y_j, etc., its computation is not possible if the values of the study variable, Y, are not given for all the population units. It is, therefore, necessary to know that, in such case, how the estimated value of the sampling variance, $V\left(\hat{Y}_{HT}\right)$, can be obtained on the basis of sample values. For this purpose, we shall present in the following theorem the estimated values of $V\left(\hat{Y}_{HT}\right)$ both for the expressions of $V\left(\hat{Y}_{HT}\right)_{HT}$ and $V\left(\hat{Y}_{HT}\right)_{YG,S}$.

Theorem 6: Unbiased estimators of the sampling variances $V\left(\hat{Y}_{HT}\right)_{HT}$ and $V\left(\hat{Y}_{HT}\right)_{YG,S}$ are respectively given by

$$\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{HT}\right\} = \sum_{i\in S}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2 + \sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}. \qquad \text{… (5.29)}$$

and

$$\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{YG,S}\right\} = \frac{1}{2}\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \qquad \text{… (5.30)}$$

Proof: Let us first consider the expression (5.29). Consider the first term in it. Taking the expectation of it, we have

$$E\left[\sum_{i\in S}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2\right] = E\left[\sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2t_i\right],$$

where t_i is the same variable as used in **Result 1** in the Sub-section 5.5.2. But, then

$$E\left[\sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2t_i\right] = \left[\sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2E(t_i)\right]$$

$$= \left[\sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2.\pi_i\right] = \sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i}.Y_i^2$$

The expression $\sum_{i\in\Omega}\frac{(1-\pi_i)}{\pi_i}.Y_i^2$ is the first term of the expression (5.24). This shows that an unbiased estimator of the first term of (5.24) is obtained as

$$\sum_{i\in S}\frac{(1-\pi_i)}{\pi_i^2}.y_i^2.$$

Now, consider the second term in the expression (5.29).

Taking the expectation of it, we have

$$E\left[\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}\right] = E\left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}.t_{ij}\right],$$

where t_{ij} is the same variable as used in ***Result 2*** in the Sub-section 5.5.2. But, then

$$E\left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}.t_{ij}\right] = \left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}.E(t_{ij})\right]$$

$$= \left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}.\pi_{ij}\right]$$

$$= \left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}\right].$$

You can see that the term,

$$\left[\sum_{i\in\Omega}\sum_{j(\neq i)\in\Omega}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}\right]$$

is the second term in the expression (5.24). So, the above result shows that an unbiased estimator of the second term in the expression of $V\left(\hat{Y}_{HT}\right)_{HT}$ is given by

$$\left[\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}\right].$$

Combining these two results, we see that unbiased estimator of $V\left(\hat{Y}_{HT}\right)_{HT}$ is given by

$$\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{HT}\right\}=\sum_{i\in S}\frac{\left(1-\pi_i\right)}{\pi_i^2}y_i^2+\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_iy_j}{\pi_i\pi_j}$$

Now, let us consider about the unbiased estimator of $V\left(\hat{Y}_{HT}\right)_{YG,S}$. Let us take the expectation of the right-hand side term of (5.30). We have

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

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

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

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

$$=\left[\frac{1}{2}\sum_{i\in \Omega}\sum_{j(\neq i)\in \Omega}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2\right]$$

$$=V\left(\hat{Y}_{HT}\right)_{YG,S}$$

This shows that an unbiased estimator of the $V\left(\hat{Y}_{HT}\right)_{YG,S}$ is given by

$$\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{YG,S}\right\}=\frac{1}{2}\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

This completes the proof of the Theorem.

Now, you may try to answer the following Self-Assessment Question:

SAQ 5

Define the Horvitz-Thompson (HT) unordered estimator of population total and show that it is an unbiased estimator for the parameter.

5.7 NON-NEGATIVITY ISSUE OF ESTIMATOR OF THE SAMPLING VARIANCE

We derived an unbiased estimator of the sampling variance of the estimator $\hat{Y}_{HT}$ given by Yates and Grundy and Sen, which is expressed in (5.30) and is presented below:

$$\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{YG,S}\right\} = \frac{1}{2}\sum_{i\in S}\sum_{j(\neq i)\in S}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

This estimator of $V\left(\hat{Y}_{HT}\right)_{YG,S}$, though estimable on the basis of sample values, but has a serious drawback. Obviously, being the estimator of variance of an estimator, it should not possess a negative value, whatever be the sample values y_is. Also, we see that it depends upon the inclusion probabilities π_i, π_j and π_{ij}, which satisfy the conditions $\sum_{i\in\Omega}\pi_i = \sum_{j\in\Omega}\pi_j = n$, and $\sum_{j(\neq i)\in\Omega}\pi_{ij} = (n-1)\pi_i$. The value of the term $\left(\pi_i\pi_j - \pi_{ij}\right)$ in the expression of $\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{YG,S}\right\}$ might be sometimes negative also. The condition for a non-negative value of $\text{Est.}\left\{V\left(\hat{Y}_{HT}\right)_{YG,S}\right\}$, therefore, requires to satisfy the relation $\left(\pi_i\pi_j - \pi_{ij}\right) > 0 \Rightarrow \pi_i\pi_j > \pi_{ij}$. This problem of non-negativity of the estimator of the variance is popularly known as *"Issue of non-negativity of the estimator of sampling variance of the estimator* $\hat{Y}_{HT}$", which must be resolved by any means.

Fortunately, this issue can be resolved by selecting any specific method of selecting a Probability Proportional to Size without Replacement method such that inclusion probabilities π_i, π_j and π_{ij} satisfy the condition mentioned above.

We shall discuss and describe such a PPSWOR selection method which assures the condition of non-negativity as mentioned above.

5.7.1 Midzuno - Sen Sampling Scheme

Midzuno (1952) and Sen (1952) proposed an interesting method of selection of a sample which was found very much useful in many applications and also for resolving some problems faced in sampling theory.

The Midzuno-Sen sampling scheme is a mixture type of scheme in which both the equal probability selection method and unequal probability selection methods are used. Using this method, the unit at the first draw is selected with unequal probability and rest of the units are selected with equal probability and without replacement.

Let us define the random variable t_i in the same way as defined in the Result 1 under Sub-section 5.5.2, that is,

$t_i = 1$, if the i^{th} unit is included in the sample,

0, otherwise.

Then, we have the probability of inclusion of the i^{th} unit in the sample

$$\pi_i = E(t_i) = P\begin{bmatrix} i^{th} \text{ unit is selected} \\ \text{at the first draw} \end{bmatrix} + P\begin{bmatrix} i^{th} \text{ unit is not selected} \\ \text{at the first draw and it is} \\ \text{selected at any of the} \\ \text{subsequent } (n-1) \text{ draws} \end{bmatrix}$$

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

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

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

$$= \left(\frac{N-n}{N-1}\right)p_i + \frac{n-1}{N-1} \;;$$

where p_i stands for the initial probability of selection of the i^{th} unit in the sample.

Similarly, the probability of inclusion of i^{th} and j^{th} unit in the sample

$$E(t_{ij}) = \pi_{ij}$$

$$= P\begin{pmatrix} i^{th} \text{ unit is selected} \\ \text{at the first draw} \\ \text{and } j^{th} \text{ unit is} \\ \text{selected at any} \\ \text{subsequent} \\ (n-1) \text{ draws} \end{pmatrix} + P\begin{pmatrix} j^{th} \text{ unit is selected} \\ \text{at the first draw} \\ \text{and } i^{th} \text{ unit is} \\ \text{selected at any} \\ \text{subsequent} \\ (n-1) \text{ draws} \end{pmatrix} + P\begin{pmatrix} \text{neither } i^{th} \text{ unit} \\ \text{nor } j^{th} \text{ unit is} \\ \text{selected at the} \\ \text{first draw and} \\ \text{both are selected} \\ \text{at any } (n-1) \\ \text{remaining draws} \end{pmatrix}$$

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

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

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

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

Following the same method, as applied above, it can be shown that

$$E(t_{ijk}) = \pi_{ijk} = \left(\frac{n-1}{N-1}\right)\left(\frac{n-2}{N-2}\right)\left[\left(\frac{N-n}{N-3}\right)\{p_i + p_j + p_k\} + \left(\frac{n-3}{N-3}\right)\right]$$

Finally, if in a sample of size n, i^{th}, j^{th}, k^{th}, …, q^{th} units are included in the sample of size n, then we have

$$E\left(t_{ijk\ldots q}\right)=\pi_{ijk\ldots q}=\frac{(n-1)(n-2)\ldots.1}{(N-1)(N-2)\ldots.(N-n+1)}\left[p_i+p_j+p_k+\ldots+p_q\right]$$

$$=\frac{1}{\binom{N-1}{n-1}}\left[p_i+p_j+p_k+\ldots+p_q\right].$$

Therefore, the probability that a sample of size n includes n distinct units of the population then, it is given by

$$\pi_{ijk\ldots g}=\frac{1}{\binom{N-1}{n-1}}\left[p_i+p_j+p_k+\ldots+p_q\right].$$

5.7.2 Resolving the Issue of Non-Negativity of Estimator of the Sampling Variance Using Midzuno-Sen Sampling Scheme

The issue of the possible negative value of the estimate of $V\left(\hat{Y}_{HT}\right)_{YG,S}$, as discussed in the Section 5.7, can be resolved with the application of Midzuno-Sen sampling scheme. We shall show below that using Midzuno-Sen scheme while selecting a PPSWOR sample of size n, the non-negativity condition given by $\left(\pi_i\pi_j-\pi_{ij}\right)>0$ is satisfied; hence, the estimate of the sampling variance $V\left(\hat{Y}_{HT}\right)_{YG,S}$ is always positive with this sampling scheme.

Theorem 7: Under Midzuno-Sen sampling scheme, we have the following condition:

$$\left(\pi_i\pi_j-\pi_{ij}\right)>0$$

for the first and second order inclusion probabilities.

Proof: From the Sub-section 5.7.1, we have

$$\pi_i=\left(\frac{N-n}{N-1}\right)p_i+\frac{n-1}{N-1}\ ;$$

$$\pi_j=\left(\frac{N-n}{N-1}\right)p_j+\frac{n-1}{N-1}$$

and

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

Therefore,

$$\left(\pi_i\pi_j-\pi_{ij}\right)=\left[\left(\frac{N-n}{N-1}\right)p_i+\frac{n-1}{N-1}\right]\left[\left(\frac{N-n}{N-1}\right)p_j+\frac{n-1}{N-1}\right]$$

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

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

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

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

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

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

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

$$=\frac{(N-n)}{(N-1)^2}\left[(N-n)p_i p_j + \left(\frac{n-1}{N-2}\right)(1-p_i-p_j)\right] > 0.$$

This completes the proof of the theorem.

Now, you may try to answer the following Self-Assessment Question with the help of contents of this section:

SAQ 6

What do you mean by the issue of non-negativity of the estimate of sampling variance of the Horvitz-Thompson estimator of population total?

5.8 SUMMARY

In this unit, we have discussed:

- The Probability Proportional to Size without Replacement sampling scheme and the difficulties which are faced in this scheme due to changing selection probabilities in each draw. Also, the comparison of the probability structures of Probability Proportional to Size with Replacement and Probability Proportional to Size without Replacement schemes.
- The concept of ordered estimators in Probability Proportional to Size without Replacement scheme along with the examples. Several ordered estimators proposed by a number of statisticians.
- The most popular ordered estimator for population mean proposed by Des Raj and the sampling variance of this estimator along with the unbiased estimator of the sampling variance. We also discussed in brief the Das ordered estimator.

- The result that for a set of ordered estimators there always exists an unordered estimator which is more efficient than the concerned ordered estimators. We explained with the help of Des Raj ordered estimator, how an unordered estimator can be framed which is easy to handle in comparison to ordered estimators.
- The concept of inclusion probabilities of different orders in contrast to initial probabilities of selection probabilities of the units, along with some basic results related to inclusion probabilities.
- The Horvitz-Thompson unordered estimator for population total and derivation of the expression of its sampling variance.
- The derivation of the expression of sampling variance of Horvitz-Thompson this estimator by Yates and Grundy and Sen.
- The unbiased estimators of the sampling variance of Horvitz-Thompson estimator.
- The problem of the Yates and Grundy and Sen estimator of the sampling variance that it may sometimes assume negative values.
- The Midzuno-Sen sampling scheme and expressions of inclusion probabilities under this scheme. It was also discussed that if sample is selected using this sampling scheme, the estimator of sampling variance proposed by Yates and Grundy and Sen is always positive. Thus, the non-negativity issue of the estimate was resolved.

5.9 TERMINAL QUESTIONS

1. Show that in Probability Proportional to Size without Replacement sampling scheme the probability of selection of a particular unit of the population changes from draw to draw.

2. Obtain the sampling variance of the ordered estimator of population mean proposed by Des Raj.

3. Let the population size and the sample size be respectively N and n. How many random samples using Probability Proportional to Size without Replacement sampling scheme of size n can be selected from the population? Further, how many ordered samples are possible for each Probability Proportional to Size without Replacement samples of size n? Illustrate your answer with the help of the population (a, b, c, d, e) and sample of size 3.

4. Show how would you construct an unordered estimator using ordered estimators of size 2 suggested by Des Raj.

5. Define the first order and second order inclusion probabilities. If the sample size is n, what would be sum of all the first-order inclusion probabilities of a unit when added over the entire population? Show your answer mathematically.

6. Obtain the sampling variance of the Horvitz-Thompson estimator of population total. State for what values of first order and second-order inclusion probabilities this variance reduces to the variance of the estimator of population total as obtained under Simple Random Sampling without Replacement sampling scheme.

7. Show that the issue of non-negativity of the estimate of sampling variance of the Horvitz-Thompson estimator of population total is resolved when inclusion probabilities are as follows:

$$\pi_i = \left(\frac{N-n}{N-1}\right)p_i + \frac{n-1}{N-1};$$

$$\pi_j = \left(\frac{N-n}{N-1}\right)p_j + \frac{n-1}{N-1}$$

and

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

5.9 ANSWERS / SOLUTIONS

Self-Assessment Questions (SAQs)

1. **Hint:** See Sub-sections 5.2.1 and 5.2.2 for your answer.

2. **Hint:** Consult the Section 5.3 as well as Sub-sections 5.3.1, 5.3.2 and 5.3.5 for completing your answer.

3. **Hint:** See the Section 5.4 for your answer to the question.

4. **Hint:** Consult the Section 5.5 and **Remarks 5.7, 5.8** and **5.9** under Sub-section 5.5.2.

5. **Hint:** Consult the Sub-section 5.6.1 and Theorem 3 under Sub-section 5.6.1.

6. **Hint:** See the Sub-section 5.7.1.

Terminal Questions (TQs)

1. **Hint:** See Theorem 1 under the Sub-section 5.2.2 for answering the question.

2. **Hint:** Consult the Sub-section 5.3.3 for searching answer to the question.

3. From the proof of the **Theorem 2** under Sub-section 5.4.1, we observe that with population size N and sample size n there would be $\binom{N}{n}$ Probability Proportional to Size without Replacement samples and each of these samples can be ordered in $n!$ ways. Since, here the population size is 5 and sample size is 3, we get $\binom{5}{3}$ Probability Proportional to Size without Replacement samples given as:

(a, b, c), (a, b, d), (a, b, e), (a, c, d), (a, c, e), (a, d, e), (b, c, d), (b, c, e), (b, d, e) and (c, d, e).

Further, each of these 10 samples can be ordered in $3! = 6$ ways.

For instance, sample (a, b, c) generates the following 6 ordered samples:

(a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b) and (c, b, a).

4. **Hint:** Refer to Sub-section 5.4.2 for searching your answer.

5. **Hint:** See the Sub-section 5.5.1 and Result 1 under the Sub-section 5.5.2.

6. **Hint:** Refer to Sub-section 5.6.2 and **Theorem 4** and **Remark 5.11** under Sub-section 5.6.2.

7. **Hint:** See the Sub-section 5.7.2.