UNIT 4

PROBABILITY PROPORTIONAL TO SIZE SAMPLING WITH REPLACEMENT

Structure

4.1 INTRODUCTION

In the first three units in this block, we focused our concern on giving emphasis to some basic concepts of sample surveys, sampling theories and then few sampling schemes which were virtually Equal Probability Selection Methods (EPSEM) in the sense that while selecting some of the population units, using some chance mechanism, in a number of draws for constituting a random sample of given size, all the units were having equal probability of selection in each draw. The assurance of assigning equal probability of selection to the population units was seen to be provided by the chance mechanism used. In this context, we defined and discussed the most basic and fundamental Equal Probability Selection Method (EPSEM), namely, Simple Random Sampling (SRS) when the characteristic under study was either quantitative or qualitative by nature. We also explained the basic difference between 'SRS with replacement' (SRSWR) and 'SRS without replacement' (SRSWOR) schemes. In fact, by assigning equal probability of selection to the units of the population in the sample is equivalent to consider all the units of the population with equal weightage (preference) of being selected in the sample without giving more or less weightage (selection probability) to some of the units over other units.

No doubt, the above sampling schemes provide us with such estimation procedures which are considered to be the 'first hand' estimated values, in the sense that; being the simplest type of sampling schemes, these assume that inherently all the units of the population are more or less of same nature in all respects. Under such an assumption, EPSEM works very well since it treats all the units with equal weightage while including them into the sample. But practically, this situation is quite rare in most of the sample surveys. Therefore, in this context, results obtained using EPSEM schemes, in fact, provide only rough estimates ignoring some kind of variations which are prevalent over the units, and which must be taken into consideration while designing a survey.

Generally, while designing a sample survey, it sometimes appears that all the units must not be selected into the sample with equal probability and that it seems appropriate to assign unequal probabilities of selection to population units before starting the selection procedure, rather than treating all of them equally probable. Actually, this feeling is very common in most of the sample surveys. We wish to explain the reason behind such feeling of assigning unequal probabilities of selection to the units of the population in most of the surveys through the following example:

Although, it is not directly related with sampling problems, but in order to understand the weightage to be given to units according to their importance in the study, we can mention a household survey in which the per month consumption of different food items has to be estimated. For this, say, we categorize the food items broadly into the categories as: cereals, pulses, vegetables, milk products, snacks, and spices. It is clear that; being a subject matter of estimation process, consumptions of items are unknown to us; therefore, we can decide how much importance or weightage are to be given to these items only with the help of some other characteristic which might

indicate the weightage of these items they contribute to the consumption pattern. Obviously, such a characteristic might be the variable 'expenditure incurred on these items per month' which is expected to exhibit a high degree of correlation with the consumption patterns. Why, because it is a very common fact in the households that consumptions of cereals and pulses are at the highest level as compared to milk products, spices and snacks; spices being at the lowest level, and therefore, expenditure on cereals and pulses are maximum as compared to expenditure on milk products, snacks and spices. This fact suggests and seems advisable too, not to assign as much selection probability of selection to milk products and spices as assigned to cereals and pulses. Cereals and pulses are more consumed, so these items must be assigned more selection probabilities for including them in the sample in comparison to selection probabilities assigned to milk products and spices. This fact enforces the investigator to assign *unequal selection probabilities* to the units of the population, which should be proportional to their expected contribution in the consumption pattern.

The aim of this unit is to highlight importance and different concepts, definitions and salient features of the unequal probability of selection methods. At the beginning, in Section 4.2, we shall discuss the points which sometimes make it necessary to use varying probability selection methods of sampling instead of equal probability selection methods. This section also highlights the importance of role of auxiliary variables in selecting the unequal probabilities of selection. Section 4.3 is devoted to the introduction of a special type of varying probability of selection scheme, namely, Probability Proportional to Size (PPS) sampling scheme. It also defines the concept of Size Measure of a unit in Probability Proportional to Size scheme and provides a lucid definition of Probability Proportional to Size scheme. The meaning of Probability Proportional to Size with replacement (PPSWR) scheme is also explained in this section. Section 4.4 explains the methods of selection of a sample using Probability Proportional to Size with Replacement scheme, namely, Cumulative Total Method and Lahiri's Methods along with their merits and demerits. Section 4.5 is devoted to the problem of estimation of population parameters, specially, population total and population mean along with their properties, through a Probability Proportional to Size with Replacement sample. Section 4.6 provides the method of finding sampling variance of these estimators. Section 4.7 shows how one can estimate the sampling variance of the sample mean estimator on the basis of the sample selected.

Expected Learning Outcomes

After studying this unit, you should be able to:

- highlight the reasons for assigning varying probabilities of selection to population units so as to constitute a random sample of fixed size;
- explain the meaning of 'Varying Probability Sampling Scheme (VPSS)' and the role of an auxiliary variable in selecting the unequal probabilities of selection which are to be assigned to the units of the population before selection process starts;
- discuss the concept of 'size measures' of population units and their use for selecting a sample under Varying Probability Sampling Scheme;

- describe the methods of selection of unequal probabilities of selection using size measures of the units;
- discuss the concept of 'Probability Proportional to Size with Replacement' (PPSWR) sampling scheme;
- discuss the matter of estimation of some population parameters on the basis of Probability Proportional to Size with Replacement scheme and highlighting the salient properties of the concerned estimators;
- describe the method of finding the expressions of variance of these estimators and to show that Simple Random Sampling with Replacement scheme is a special case of Probability Proportional to Size with Replacement scheme; and
- obtain the expression of the estimator of sampling variance of the estimate of population mean.

4.2 VARYING PROBABILITY SAMPLING SCHEME

In previous section, we have highlighted the reason for assigning the varying probabilities of selection to population units so as to constitute a random sample of fixed size. Now, in this section, we shall discuss the points which sometimes make it necessary to use varying probability selection methods of sampling instead of equal probability selection method through an example. We shall also highlight the importance of role of auxiliary variables in selecting the unequal probabilities of selection.

4.2.1 Reason for Assigning Unequal Probability of Selection

We mentioned in the previous section that Varying Probability Sampling Scheme are seen to be more appropriate in providing more efficient results as compared to equal probability sampling methods, due to the assignment of unequal probability of selection of the units in the sample; the selection probability assigned according to their weightage they contribute in the estimated value. To illustrate this, we consider the following example:

Let there be 10 students in a class who appeared in a mid-term class test. The marks secured by them out of 50 in the test are given in the following table:

Roll Numbers of students	Marks obtained in class test (out of 50)
1	46
2	14
3	21
4	44
5	48
6	8
7	16
8	34

9	19
10	23
Total	**273**

Let we wish to find an estimate of the average mark of the class in the mid-term class test. Let it be estimated with the help of a random sample of size 4.

Let us first estimate the average mark using Simple Random Sampling without Replacement scheme, that is, using Equal Probability Selection Method (EPSEM). Obviously, then each and every mark presented in the table would be selected with selection probability equal to 1/10. Since, the population size is too small, we use lottery method (chit method) for selecting four marks under Simple Random Sampling without Replacement scheme. Therefore, we prepare 10 identical chits. Let the chits selected randomly with labels 2, 4, 6 and 9.

Obviously, then the sample mean would be

$$\bar{x} = \frac{(14 + 44 + 8 + 19)}{4} = 21.25$$

an estimate of the population average which is

$$\bar{X} = \frac{273}{10} = 27.3.$$

Now, let us prepare in all 273 identical chits and mark 46 chits with label 1, 14 chits with label 2, 21 chits with label 3 and so on, that is, mark as many chits as the mark of a particular student with label same as the roll number of that student. We, thus, have 273 chits in all. Obviously, now each student has unequal probability of selection in the sample; clearly, first student assigned probability 46/273, second student assigned probability 14/273, third student assigned probability 21/273, …., sixth student assigned probability 8/273, and so on. On the basis of random number table, we selected 4 numbers from 1 to 273 (starting in the second row with last three digits of random numbers) and got selected labels as 123, 248, 265 and 52. This indicates that roll numbers 4, 9, 10 and 2 were selected in the sample.

The sample mean estimate then obtained as

$$\bar{x} = \frac{(44 + 19 + 23 + 14)}{4} = \frac{100}{4} = 25.00.$$

The estimate is closer to the actual value 27.3.

Remark 4.1: The above example is just an attempt to show that using unequal (varying) probability of selection, many times we may get more accurate results as obtained from the sample values. You can see that, as far as the contribution of marks towards the average mark of the population is concerned, roll numbers 1, 4 and 5 contribute a lot as compared to roll numbers 2, 6, 7 and 9. Accordingly, we developed a method of selection which gives more weightage (probability of selection) to more intelligent students as compared to other students. On the other hand, in EPSEM schemes, we give equal weightage to all the students without considering their intelligence level,

which is a salient characteristic of students as far as their marks are concerned.

Remark 4.2: There are more examples where the varying probability selection method seems to be more justifiable. In case if one wishes to estimate the average expenditure per month of families in a locality, assignment of unequal probability of selection to families, considering their sizes seems scientific, since, more the number of members in the family more would be the expenditure. If annual average income of families is under consideration, unequal probabilities of selection can be assigned to families on the basis of number of earning members in it, as, more earning members in the family means more income of the family.

4.2.2 Symbolic Presentation of Varying Probability Sampling Scheme

The Varying Probability Sampling Scheme can also be represented using symbols of Section 1.5 of Unit 1. We have

Label (i): $\{1, 2, 3, ..., i-1, i, i+1, ..., N\}$; where i being integers.

Population (U): $\{U_1, U_2, U_3, ..., U_{i-1}, U_i, U_{i+1}, ..., U_N\}$; where U_i being the i^{th} unit of the population.

The units of the population may be any object on/from which necessary information are gathered when they are selected in the sample.

Characteristic Under Study (Y): $\{Y_1, Y_2, Y_3, ..., Y_{i-1}, Y_i, Y_{i+1}, ..., Y_N\}$; where, Y_i being the value of the study variable on the i^{th} unit of the population.

In sample surveys there is always at least one characteristic under study, the value of which is recorded/measured for each and every selected unit. The characteristic might be either of quantitative nature or of qualitative nature.

Probability of Selection Assigned to Units (p):
$\{p_1, p_2, p_3, ..., p_{i-1}, p_i, p_{i+1}, ..., p_N\}$; where, p_i being the probability of selection, assigned to the i^{th} unit of the population.

We know that before starting the selection procedure of units, we develop/search some compatible chance mechanism in order to associate with each unit of the population some non-zero probability of selection in the sample. These probabilities might be equal for all the units, as in case of Equal Probability Selection Method or might be unequal over unit to unit, as in case of Varying Probability Sampling Scheme. Symbolically, we see that $p_i = \frac{1}{N}$ for $i = 1, 2, 3, \ldots, N$; in case of Equal Probability Selection Method; like Simple Random Sampling with Replacement, Simple Random Sampling without Replacement; whereas in Varying Probability Sampling Scheme, p_is differ over $i = 1, 2, 3 \ldots, N$ such that these satisfy the following two conditions:

(i) $0 < p_i < 1$; for all i and (ii) $\sum_{i=1}^{N} p_i = 1$.

Remark 4.3: It is easy to see that Equal Probability Selection Method is a special case of Varying Probability Sampling Scheme. Whenever, in Varying Probability Sampling Scheme, p_is are equal for all i, it converts into Equal Probability Selection Method. Thus, the results obtained under Varying Probability Sampling Scheme are generalization of the Equal Probability Selection Method, like Simple Random Sampling schemes.

Remark 4.4: In Simple Random Sampling schemes (Simple Random Sampling with Replacement, Simple Random and Sampling without Replacement), the selection of units from the population for including them into the sample, we used some chance mechanisms, such as Lottery method (Chit method) and Random Number Tables method; both of which assured that each unit of the population has an equal probability of selection given by $\frac{1}{N}$ at each draw. Thus, the mechanisms were compatible with the underlying assumption of equal probability of selection. Since, in Varying Probability Sampling Scheme, the underlying assumption is that probabilities of selection are unequal for all the units, we cannot use the above two methods of selection in the same way as we applied in Equal Probability Selection Method; rather we should develop some compatible chance mechanisms for Varying Probability Sampling Scheme. We shall mention afterwards in a separate section some selection methods which assure varying probabilities of selection.

Based on what we discussed and presented till now in this section; you can try to answer the following Self-Assessment Question:

SAQ 1

What do you mean by Varying (Unequal) Probability Sampling Scheme?

4.2.3 Use of an Auxiliary Characteristic for Finding Probabilities of Selection

Unlike the Equal Probability Selection Method, in Varying Probability Sampling Scheme, a problem arises that how one can decide about the probability of selection, p_i to be assigned to the i^{th} unit of the population such that these satisfy the conditions:

(i) $0 < p_i < 1$ for all i

and (ii) $\sum_{i=1}^{N} p_i = 1.$

No doubt, one can abruptly decide the selection probabilities without giving any scientific reason, but then he/she is not sure about the accuracy of the results derived from the sample values. Moreover, if for a given population, the set of probabilities are selected by a number of samplers according to their own choice, there would be a number of resulting values for the same population parameter. As for example, let a population consists of five units $\{U_1, U_2, U_3, U_4, U_5\}$; the abruptly selection probabilities might be any of the following sets:

(i) $\{p_1 = 0.21, p_2 = 0.11, p_3 = 0.36, p_4 = 0.09, p_5 = 0.23\}$;

(ii) $\{p_1 = 0.41, p_2 = 0.17, p_3 = 0.04, p_4 = 0.30, p_5 = 0.08\}$;

(iii) $\{p_1 = 0.17, p_2 = 0.11, p_3 = 0.15, p_4 = 0.34, p_5 = 0.23\}$;

(iv) $\{p_1 = 0.25, p_2 = 0.15, p_3 = 0.33, p_4 = 0.11, p_5 = 0.16\}$;

Since, the conditions imposed are satisfied by all the sets, theoretically, any one of them can be considered without any doubt. However, since probabilities are abruptly chosen by the samplers applying their decision, the derived results obtained will be questionable by common people and experts too.

To avoid abrupt choices of the selection probabilities and to give a scientific base to the selection process, a specific type of Varying Probability Sampling Scheme is developed in literature. It is popularly known as "**Probability Proportional to Size" (PPS)** sampling scheme.

In fact, a PPS sampling scheme defines a specific rule of deciding the selection probabilities to be associated with each population unit. This rule is devised on the basis of another variable, other than the study variable. In sampling theory, such a variable for which value on each unit is known from some sources or only a function of values is known, and the variable exhibit a good degree of correlation with the study variable; is popularly called an '**Auxiliary Variable' or 'Ancillary Variable'**. The role of such variables in sampling theory is manifold. These variables might be used at the early stages of the survey for the purpose of improving the procedure of selection of units for constituting a better sample or these might be used in a later stage of the survey for improving the estimation procedure. In whatever form these variables are used, it is seen that using an auxiliary characteristic along with the study characteristic one could get improved results derived from the sample values. We shall show here how use of such auxiliary variables may provide us selection probabilities in Probability Proportional to Size scheme.

In the next section we shall discuss about the concept, definition and other theories related to Probability Proportional to Size sampling scheme.

Based on what we discussed and presented in this sub-section; you can try to answer the following Self-Assessment Question:

SAQ 2

Explain the role of an auxiliary variable in selecting the unequal probabilities of selection in sampling schemes.

4.3 PROBABILITY PROPORTIONAL TO SIZE SAMPLING SCHEME

In order to understand the concept of Probability Proportional to Size sampling scheme, first of all, it becomes necessary to be familiar with the term ***"Size Measure of the Unit"***, the use of an auxiliary variable for deciding the 'Size Measure of the Unit' as well as the procedure of finding the selection probabilities. The following sub-sections explain these concepts:

4.3.1 Size Measure of the Unit

We know that unlike the Equal Probability Selection Method schemes, in Varying Probability Sampling Scheme selection probabilities p_is are different from unit to unit. Therefore, in Varying Probability Sampling Scheme the main problem is first to obtain the values of the pre-fixed probability p_i, for i = 1, 2, …, N, associated with the unit U_i of the population. We have seen that while selecting some of the units of the population in a number of draws so as to constitute a random sample; we needed a compatible chance mechanism which assured that a particular unit was selected in the sample with the pre-decided probability of selection which was assigned to that unit well before the selection procedure.

Since, Y, being the characteristic under study, its values Y_is for i = 1, 2, …, N over population units are not known to the sampler, therefore, in order to find the probabilities p_is, the study variable Y is not useful. In such situations, the only way is to use another characteristic (variable), different than the study characteristic Y, which exhibit a good degree of correlation with the study variable, that is, to use an Auxiliary Variable. In sample surveys, in most of the situations such auxiliary variables are readily available. Let us denote an auxiliary variable by X. Let us assume that the value of the variable X is known for each unit in the population. Whenever, in sample surveys an auxiliary variable, exhibiting close relation with the study variable, is available, it is called a ***"*Measure of Size of the Unit"** (or, **"Size Measure of the Unit"**) belonging to the population under consideration.

Remark 4.5: In sample surveys there is always a number of characteristics which are associated with each unit of the population. Out of these characteristics, generally, any one of the characteristics is used for study purpose, which is known as 'Characteristic under Study' (study variable). All other remaining characteristics, therefore, might be considered to be the auxiliary variables which can be used as size measure of the units of the population. Obviously, the best possible auxiliary variable would be that one which exhibit a good degree of correlation with the study variable. For example, in a survey, designed for estimating the average income per month of persons belonging to certain families (obviously, 'income' is the study variable in this case), family expenditure per month, total number of family members or number of earning persons in the family might help as a size measures; but the best possible size measure would be the family expenditure per month, if known from the past records of the families; because in comparison to other size measures, it is expected to have the closest relationship with income. Similarly, in agricultural surveys for estimating the yield of crops, the area under the crop, amount of manure applied, soil fertility, amount of rainfall, etc., might be considered to be the size measures; but the area under the crop is expected to be the best possible size measure due to obvious reason. In surveys of manufacturing industries for estimating total output of finished materials; number of workers, amount of raw material used, total working hours per unit of time might be used as size measures and perhaps the number of workers would be the best possible size measure amongst others.

Remark 4.6: It is, therefore, clear from the above discussions that in order to use Probability Proportional to Size sampling scheme in any sample survey, one should clearly state what size measure of the population units he/she is going to use. It is also necessary for the sampler to explore the possibility of having knowledge of values of the size measure on the units considered in the survey, either from current records or past records kept in the concerned organization, from past experience of similar type of survey or from any other reliable source.

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

SAQ 3

In the context of PPS sampling scheme, mention the concept of Size Measure of a unit of the population under study.

4.3.2 Defining Probability Proportional to Size Sampling Scheme

Probability Proportional to Size sampling scheme is a special type of Varying Probability Sampling Scheme. In Probability Proportional to Size scheme, there is a set rule of finding the selection probabilities, whereas Varying Probability Sampling Scheme in general, does not define such a rule which could be accepted as a scientifically designed rule of selection. The Probability Proportional to Size sampling scheme is defined as follows:

The sampling scheme in which units of the given population are selected one by one in a number of draws, in order to include them into a random sample of given size, in such a way that the probability of selection for each unit is proportional to some size measure of the unit, is known as "**Probability Proportional to Size (PPS) Sampling Scheme**".

Using symbols we defined earlier, the definition can be elaborated in a simple manner. Below we have sequentially arranged the given population with its units, values the units assumed on the study variable, probability of selection assigned to each unit and the value of the size measure of the units:

Population (U): $\{U_1, U_2, U_3, ..., U_i, ..., U_N\}$

Study Variable (Y): $\{Y_1, Y_2, Y_3, ..., Y_i, ..., Y_N\}$

Probability of Selection (p): $\{p_1, p_2, p_3, ..., p_i, ..., p_N\}$

Size Measure (X): $\{X_1, X_2, X_3, \ldots, X_i, \ldots, X_N\}$

Symbolically, Probability Proportional to Size scheme states the rule of selection of the units in each draw as

$$p_i \propto X_i \text{ for } i = 1, 2, \ldots, N. \qquad \ldots (4.1)$$

From (4.1), we have $p_i = k.X_i$; where k stands for constant of proportionality.

Summing both the sides over N, we have

$$\sum_{i=1}^{N} p_i = k\sum_{i=1}^{N} X_i$$

Let us denote the sum of all X_is by X, then we have

$$\sum_{i=1}^{N} p_i = k\sum_{i=1}^{N} X_i$$

$$\Rightarrow 1 = k.X \Rightarrow k = \frac{1}{X} \qquad \text{Since } \sum_{i=1}^{N} p_i = 1$$

Substituting the value of k in the expression $p_i = k.\ X_i$;we have the expression

$$p_i = \frac{X_i}{X} = \frac{X_i}{\sum_{i=1}^{N} X_i} \quad \text{for all i.} \qquad \ldots (4.2)$$

4.3.3 Types of Probability Proportional to Size Sampling Scheme

We are familiar with the fact that a random sample may be selected in two ways; (i) 'With Replacement' of units once selected in any draw again in the population before making the next draw and (ii) 'Without Replacement' of selected units again in the population. Since, Probability Proportional to Size sampling is also a random sampling scheme with some rules imposed on defining the selection probabilities only, a Probability Proportional to Size sampling scheme may be operated either with any one of them. Accordingly, we come across with the following:

(i) Probability Proportional to Size with Replacement sampling scheme

(ii) Probability Proportional to Size without Replacement sampling scheme.

No doubt, then in Probability Proportional to Size with Replacement sampling scheme, a particular unit of the population is selected in a draw following the selection probability rule as given in (4.2). After noting down its details, the selected unit is supposed to be replaced in the population before making the next draw so that it has some chance of being selected in the sample again at any forthcoming draw. In this sense, Probability Proportional to Size with Replacement scheme has no difference with Simple Random Sampling with Replacement in operational procedure of selection except that probability of selection of a particular unit at a particular draw is different in both the schemes.

In the present unit, we shall confine our study only on the Probability Proportional to Size with Replacement scheme.

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

SAQ 4

Define the Probability Proportional to Size sampling scheme. Show that in Probability Proportional to Size sampling, the selection probability, p_i is given by $\left(X_i / X\right)$ where the symbols have their usual meanings.

4.4 SELECTION PROCEDURES OF A PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SAMPLE

Let us now discuss the methods which can be used for selecting a random sample of a given size from the population using probability proportional to size. There are two methods of selection which assure that the units in the sample are selected with the prescribed selection probabilities, assigned to each unit of the population. The methods are

(a) Cumulative Total Method

(b) Lahiri's Method.

4.4.1 Cumulative Total Method

Let there be N units in the population and the auxiliary variable X has values $X_1, X_2, X_3, \ldots, X_N$, respectively, on these population units, serially arranged as the first, second, third, …, N^{th} unit, which are known by some means. Let the total of , X_i values be given by $X = \sum_{i=1}^{N} X_i$. The cumulative total method is essentially based on assigning each of the unit a set of consecutive natural numbers such that the number of numbers in the set is equal to the size of the unit. Let us arrange the unit labels (i), the values of the study variable (Y_i), the corresponding sizes of units (X_i), Cumulative Totals (T_i) and selection probabilities (p_i) serially as shown in the following table:

Serial Number of Units (Labels i)	Values of the Study Variable (Y_i)	Size Measure of Units (X_i)	Cumulative Totals (T_i)	Probability of Selection $\left(p_i = \frac{X_i}{X}\right)$
1	Y_1	X_1	$T_1 = X_1$	X_1/X
2	Y_2	X_2	$T_2 = X_1 + X_2$	X_2/X
3	Y_3	X_3	$T_3 = X_1 + X_2 + X_3$	X_3/X
4	Y_4	X_4	$T_4 = X_1 + X_2 + X_3 + X_4$	X_4/X
.	.	.	.	.
.	.	.	.	.
N	Y_N	X_N	$T_N = X_1 + X_2 + X_3 + ... + X_N$	X_N/X
Total	**---**	$X = \sum_{i=1}^{N} X_i$.	**---**	**1.00**

Steps for Finding Probability of Selection (p_i):

(1) In the table, write the labels (i) of the units of the population in the first column of the table. Since, the population size is N, the N labels would be first N integers 1, 2, 3, …, N.

(2) Write the values of the study variable, Y, for the units of the population in the next column of the table. However, this column is optional and may not be included in the table.

(3) In the third column of the table, write the size measures (X_i) of the units. Find the total of X_is at the last row of the column. It is denoted by X.

(In other words, we can say that with each unit we associate a number of numbers equal to its size). Thus, we associate X_1 numbers (1 to X_1) to the first label, X_2 numbers (X_1+1 to $X_1 + X_2$) to the second label, ($X_1 + X_2$ +1 to $X_1 + X_2 + X_3$) to the third label and so on.

(4) Find the Cumulative Totals, denoted by T_i (i = 1, 2, 3 …, N) in the fourth column of the table for each label i; where

$$T_i = X_1 + X_2 + X_3 + \ldots + X_i = T_{i-1} + X_i.$$

(5) Using any chance mechanism or conveniently using random number table, choose a number R at random from 1 to T_N.

(6) Select the label i equivalently, the unit U_i for its inclusion in the sample if $T_{i-1} < R \leq T_i;\ \ T_0 = 0.$

(7) After noting down the details of this unit, assume that it is replaced again in the population so that the population size N remains same, and that the same unit may be selected again in some other draw.

(8) Repeat the process n times, so that n units are included in the sample.

Example 1: In a primary school, the aim is to estimate the average age of children studying with the help of a sample of students. There are in all 9 sections in the school with respective strength of students as 47, 30, 40, 60, 45, 36, 65, 75 and 26. It was decided to select a sample of five sections for selecting students from the selected sections. Considering the strength of students of the sections as a size measure, select the Probability Proportional to Size with Replacement sample.

Solution: The following table shows the labels allotted to sections (i), size measures associated with the sections (X_i), cumulative totals (T_i) and selection probabilities (p_i) for each label:

Serial Number of Units (Labels i)	Values of the Study Variable (Y_i)	Size Measure of Units (X_i)	Cumulative Totals (T_i)	Probability of Selection $\left(p_i = \frac{X_i}{X}\right)$
1	47	47	$T_1 = 47$	$47/424 = 0.11$
2	30	30	$T_2 = 47 + 30 = 77$	$30/424 = 0.07$
3	40	40	$T_3 = 47 + 30 + 40 = 117$	$40/424 = 0.09$
4	60	60	$T_4 = 47 + 30 + 40 + 60 = 177$	$60/424 = 0.14$
5	45	45	$T_5 = 222$	$45/424 = 0.11$
6	36	36	$T_6 = 258.$	$36/424 = 0.08$
7	65	65	$T_7 = 323$	$65/424 = 0.15$
8	75	75	$T_8 = 398$	$75/424 = 0.18$
9	26	26	$T_9 = 424$	$26/424 = 0.07$
Total	**---**	$\sum_{i=1}^{9} X_i = 424$		**1.00**

The cumulative totals T_is corresponding to different labels given in the fourth column show the cumulative total of number of numbers associated with each unit. For example, numbers 1 to 47 are associated with the first label; numbers 48 to 77, that is, 30 numbers are associated with second label and so on. So, the cumulative total for label 1 is 47 and for label 2 is 47 + 30 = 77, etc.

Probabilities of selection (p_i) are calculated with the formula $p_i = \frac{X_i}{X}$

Thus, $p_1 = \frac{47}{424} = 0.11;\ p_2 = \frac{30}{424} = 0.07$ and so on.

Using the random number table given in **Appendix – A** of the Unit 1, we start the selection of random numbers from the third column and move column-wise. Since T_9 is 424, we use the first three digits of the random numbers. We observe that the first random number selected is 267 which means that label 7 is selected with probability 0.15 since, the condition 258 < R < 323 (that is $T_6 < R < T_7$) is satisfied for 7th label. The second number is 129 indicating that label 4 is selected in the sample with probability 0.14. Proceeding in the same way, we observe that the next three numbers are 372, 164 and 394 indicating that labels 8, 4 and 8 are selected in the sample.

Thus, sections with labels {7, 4, 8, 4, 8} are selected in the sample under Probability Proportional to Size with Replacement sampling scheme with respective selection probabilities {0.15, 0.14, 0.18, 0.14, 0.18}.

Remark 4.7: Since, the selection of Probability Proportional to Size sample is made using with replacement scheme, obviously, the total number of Probability Proportional to Size samples will be N^n. For this example, therefore, we will have total $9^5 = 59049$ samples.

We shall now show one important property of Cumulative Total method that under this method, i^{th} unit of the population is selected in the sample with selection probability equal to (X_i/X); where symbols have their usual meanings. We have the following theorem:

Theorem 1: The selection probability of a unit to be selected in the sample in Probability Proportional to Size sampling scheme is proportional to its size.

Proof: Let us consider the i^{th} unit of the population whose size be given by X_i (i = 1, 2, …, N) and selection probability be given by p_i. We have to show that

$$p_i \propto X_i \qquad \text{for } i = 1, 2, \ldots, N$$

We have seen that the i^{th} unit of the population is selected in the sample at any draw if the randomly selected number R satisfies the condition:

$$T_{i-1} < R \leq T_i$$

that is, $$T_i - T_{i-1} = X_i$$

numbers out of total cases X are favourable cases to the happening of the i^{th} unit in the draw. Therefore, the probability of selection of the i^{th} unit at any draw is given by

$$P(U_i) = \frac{X_i}{X}$$

Treating $\frac{1}{X}$ as constant of proportionality, the above expression can be reduced to

$$P(U_i) \propto X_i$$

This shows that the selection probability of the i^{th} unit is proportional to its size.

This establishes the theorem.

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

SAQ 5

What is 'Cumulative Total Method' of selecting a Probability Proportional to Size with Replacement sample?

4.4.2 Lahiri's Method

Another method of selecting a Probability Proportional to Size sample was introduced by Lahiri in 1951 in a research paper. This method is a modification over Cumulative Total Method. We have seen that in the previous method, we need to list all the size measures and their cumulative totals, which is a time consuming and costly affair if the population size is large enough. Lahiri's method avoids the cumulation process.

Lahiri's method needs to associate with each unit of the population the same number of numbers out of which only the numbers of numbers equal to its size are considered to be 'effective numbers' and rest numbers are considered to be 'ineffective numbers'. Let us denote by the same number of numbers to be associated with each unit by M, where M may be any number greater than or equal to the maximum size. In other words, we select the number $M \geq \max.(X_1, X_2, \ldots, X_N)$. This means that, in fact, we associated in all NM numbers out of which only X_i numbers are considered to be *effective* numbers for the unit U_i for i = 1, 2, …, N, and rest $(M - X_i)$ numbers are considered to be *ineffective* numbers for this unit.

The process of Lahiri's method consists of the following steps:

(1) Select a pair of random numbers (u, v) such that $1 \leq u \leq N$ and $1 \leq v \leq M$.

(2) If $u = k$, say; then k^{th} unit of the population is provisionally selected for the sample.

(3) Now, if we see that $v \leq X_k$ then k^{th} unit of the population is *finally* selected for the sample. Otherwise, in case $v > X_k$ then the k^{th} unit of the population is *not selected at all* in the sample and, therefore, the draw was treated to be in vain.

(4) We again select a pair of random numbers and repeat the steps (1), (2) and (3) in order to select a unit of the population in the sample.

(5) Repeat the above steps till we get n units in the sample.

Example 2: Use the data given in the **Example 1** for illustrate the procedure of selection of a sample of 5 units using Lahiri's method with Probability Proportional to Size with Replacement sampling scheme.

Solution: We have N = 9 and n = 5. We arrange the information needed for the process in the following table:

Labels (i)	Size Measures (X_i)	Effective Numbers	Ineffective Numbers
1	47	1 - 47	48 - 75
2	30	1 - 30	31 - 75
3	40	1 - 40	41 - 75
4	60	1 - 60	61 - 75
5	45	1 - 45	46 - 75
6	36	1 - 36	37 - 75
7	65	1 - 65	66 - 75
8	75	1 - 75	---
9	26	1 - 26	27 - 75
Total	**424**		

Let us decide the value of M, the equal number of numbers to be associated with each unit. We choose it equal to maximum size, which is 75. So, M = 75. We can choose M any other number greater than maximum size, say 100; but as we shall see afterwards that it is advisable to choose M = max X_i for some reason. We then classified the M numbers associated with each unit into two categories as effective numbers and ineffective numbers. Effective numbers are the numbers which are equal to the size of each unit and rest of the numbers out of M are ineffective numbers. Thus, third column depicts effective numbers as 1 - 47, 1 - 30, 1- 40 and so on, where 47, 30 and 40 are the sizes of the first three labels. Fourth column shows the ineffective numbers for each label.

Now, let us start the process of selecting the units for the sample one by one. We shall select two random numbers u and v such that $1 \leq u \leq N$ and $1 \leq v \leq M$. We can use the Random Number Table for this purpose. Obviously, u will be one-digit numbers and v will be two-digit numbers. Let us start with the first column of the table and select the unit-digit number of random numbers for u and start with the third column of the table with last two digits for v.

We shall now list all the draws below and the result of these draws:

Selected Value of u	Selected Value of v	Result of the Draw
6	9	**Label 6** is provisionally selected and since v is less than 36, its size; so, it is **finally selected** in the sample **at the first draw**.
3	75	**Label 3** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
3	35	**Label 3** is provisionally selected and since v is an effective number for the label, it is **finally selected** in the sample **at the third draw**.
7	71	**Label 7** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
5	55	**Label 5** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.

7	20	**Label 7** is provisionally selected and since v is an effective number for this label, it is **finally selected** at the **sixth draw**.
7	41	**Label 7** is provisionally selected and since v is an effective number for this label, it is **finally selected** at the **seventh draw**.
2	41	**Label 2** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
3	64	**Label 3** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
5	47	**Label 5** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
2	53	**Label 2** is provisionally selected but since v is an ineffective number for this label, it is **not selected** in this draw. The draw is then **ineffective draw**.
6	30	**Label 6** is provisionally selected and since v is an effective number for this label, it is **finally selected** at the **twelfth draw**.

Since, five units are finally selected in the sample, we stop as soon as we attain the sample size. The labels which are selected in the sample are {6, 3, 7, 7, 6}. This constitutes a Probability Proportional to Size with Replacement sample.

Remark 4.8: However, in Lahiri's method, we observed that some of the draws remained ineffective and, hence, could not provide a unit in the sample. For selecting five units in the sample, we required to go up to twelve draws.

Remark 4.9: Although Lahiri's method seems to be cumbersome, but it is more important that whether the process really provides a sample using the rule of Probability Proportional to Size, which is $p_i \propto X_i$. that is, the probability of selection of a particular unit is proportional to its size.

We shall theoretically prove that Lahiri's method also follows the rule $p_i \propto X_i$. For this, we shall state and prove the following theorem regarding the condition of Probability Proportional to Size sampling scheme, that is, units are selected with probability proportional to their size:

Theorem 2: In Lahiri's method of selection of units in the sample, the condition of Probability Proportional to Size sampling scheme is satisfied, that is, units are selected with probability proportional to their size.

Proof: Since, in all NM numbers are associated with the units of the population, out of which X_k numbers are favourable for the provisional selection of the k^{th} unit at any draw, the probability of selection of k^{th} unit is given by

$$\frac{X_k}{NM} = q_k, (\text{say}). \qquad \ldots (4.3)$$

Therefore, the probability that no unit is selected in a draw $= 1 - \sum_{k=1}^{N} q_k$.

But

$$1 - \sum_{k=1}^{N} q_k = 1 - \sum_{k=1}^{N} \frac{X_k}{NM}$$

$$= 1 - \frac{1}{NM}\sum_{k=1}^{N} X_k$$

$$= \left(1 - \frac{\bar{X}}{M}\right) = P(k), \text{ (say)}; \quad \text{where } \bar{X} = \frac{\sum_{k=1}^{N} X_k}{N}.$$

We know that the k^{th} unit, U_k, may or may not be selected at a particular draw and the number of ineffective draws may be theoretically infinite, the probability that it is selected at the first effective draw, denoted by $P(U_k)$ will be given by

$$P(U_k) = P_1(U_k) + P_2(U_k) + P_3(U_k) + \ldots \text{ up to } \infty;$$

where $P_r(U_k)$ stands for the probability that U_k is selected at the r^{th} draw. We see that

$$P(U_k) = \frac{X_k}{NM} + \left(1 - \frac{\bar{X}}{M}\right)\left(\frac{X_k}{NM}\right) + \left(1 - \frac{\bar{X}}{M}\right)^2 \frac{X_k}{NM} + \left(1 - \frac{\bar{X}}{M}\right)^3 \frac{X_k}{NM} + \ldots.$$

$$= \frac{X_k}{NM}\sum_{r=0}^{\infty}\left(1 - \frac{\bar{X}}{M}\right)^r$$

Since, $0 < \left(1 - \frac{\bar{X}}{M}\right) < 1$, we have

$$P(U_k) = \left(\frac{X_k}{NM}\right) \times \frac{1}{\left\{1 - \left(1 - \frac{\bar{X}}{M}\right)\right\}}$$

$$= \left(\frac{X_k}{NM}\right) \times \frac{M}{\bar{X}} = \frac{X_k}{N\bar{X}} = \frac{X_k}{\sum_{i=1}^{N} X_i} = \frac{X_k}{X}$$

This shows that the probability of selection of the k^{th} unit is proportional to its size X_k, where k = 1, 2, 3, …, N.

This establishes the theorem.

Remark 4.10: We observed that in Lahiri's method, a number of the draws becomes ineffective draws in the sense that no unit is selected in these draws. Obviously, no one can afford a large number of ineffective draws, since, it increases the total time consumed in the entire process. How it can be minimized? In the above theorem, we have obtained the probability of an ineffective draw which is

$$P(k) = \left(1 - \frac{\bar{X}}{M}\right).$$

In the expression, the only quantity which is in our command is M, the number of numbers we associate with each of unit. From the expression of P(k), it is clear that the term $\frac{\bar{X}}{M}$ decreases with increasing M and, hence, P(k) goes on increasing as $M \to \infty$. So larger is the choice of M, the larger would be the

probability of having ineffective draws. This indicates that since, M $\geq$ max. (X_1, X_2, …, X_N), the minimum possible value of M should be equal to max. X_i. However, the condition states that M may be taken greater than max. X_i also, if one desires so, but then it will increase unnecessarily the number of ineffective draws. For example, in the above example, we took M =75; the max. X_i. So, the probability of ineffective draws would be

$$P(k) = \left(1 - \frac{\bar{X}}{M}\right) = \left(1 - \frac{424/9}{75}\right) = (1 - 0.6281) = 0.3719;$$

whereas, taking M = 150, we have,

$$P(k) = \left(1 - \frac{\bar{X}}{M}\right) = \left(1 - \frac{424/9}{150}\right) = 1 - 0.3141 = 0.6859$$

which is approximately 1.84 times more than when M = 75.

Merits and Demerits of Lahiri's Method compared to Cumulative Total Method:

(i) Merits

1. Cumulative Total method needs a lot of time in the process of cumulation of sizes which is not necessary in the case of Lahiri's method. So, this process is less time consuming.
2. There is no need to know the sizes of all the units; only the maximum size and the sizes of the provisionally selected unit are required.

(iii) Demerits

1. Lahiri's method seems to be comparatively complex. We need to select two random numbers u and v for the selection process. Moreover, there is a number of draws which are ineffective because of which we have to make more number of draws than the sample size, which is wastage of time.

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

SAQ 6

What are 'effective' and 'ineffective' numbers in Lahiri's method of sample selection?

4.5 ESTIMATION OF SOME PARAMETERS IN PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT SCHEME

After discussing the selection methods of selecting a Probability Proportional to Size with Replacement sample from the given population, we shall now move towards the problem of estimation of some population parameters on the basis of a random sample selected under Probability Proportional to Size with Replacement sampling scheme.

4.5.1 Estimation of Population Total

Let the population total be denoted by Y_T. Then, we have,

$$Y_T = Y_1 + Y_2 + Y_3 + \ldots + Y_N = \sum_{i=1}^{N} Y_i \qquad \ldots (4.4)$$

We wish to estimate the population total using a sample selected with Probability Proportional to Size with Replacement sampling scheme. We shall consider two cases: (i) estimation with a sample of size one and (ii) estimation with a sample of size n.

(a) Estimation with a Sample of Size One

Let a unit is selected in the sample with Probability Proportional to Size with Replacement sampling scheme. Let it be y_i. Obviously, the unit will be selected with selection probability p_i. Since, y_i is a sampled unit with probability p_i it would be a random variable assuming values $Y_1, Y_2, Y_3, \ldots, Y_N$ with corresponding selection probabilities $p_1, p_2, p_3, \ldots, p_N$; such that $\sum_{i=1}^{N} p_i = 1$. Let us consider the function of y_i as:

$$\hat{Y}_T = \frac{y_i}{p_i} \qquad \ldots (4.5)$$

Then, we have

$$E\left(\hat{Y}_T\right) = E\left(\frac{y_i}{p_i}\right) = \sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i}\right) = \sum_{i=1}^{N} Y_i = Y_T \qquad \ldots (4.6)$$

since, by definition of expectation, $E\left(y_i\right) = \sum_{i=1}^{N} p_i Y_i$

The result (4.6) shows that $\hat{Y}_T$ is an unbiased estimator of the population total. The sampling variance of the estimator $\hat{Y}_T$ will be derived in the Sub-section 4.6.1 afterwards.

(b) Estimation with a Sample of Size n

Let a sample of size n be drawn from the given population under the Probability Proportional to Size with Replacement sampling scheme. Let us denote by $\left(y_r, p_r\right)$, respectively the observational value of the study variable Y and the probability of its selection of the unit selected at the r^{th} draw $\left(r = 1, 2, \ldots, n\right)$. Then, the variables $\frac{y_r}{p_r}$ are independent random variables each assuming N values $\frac{Y_i}{p_i}$ with probabilities $p_i \left(i = 1, 2, \ldots, N\right)$.

The following theorem defines the estimator of population total and provides its unbiasedness property:

Theorem 3: In Probability Proportional to Size with Replacement sampling scheme, an unbiased estimator of population total Y_T is given by

$$\hat{Y}_{PPSWR} = \frac{1}{n} \sum_{r=1}^{n} \left(\frac{y_r}{p_r}\right) \qquad \ldots (4.7)$$

Proof: We have,

$$E\left(\hat{Y}_{PPSWR}\right) = E\left[\frac{1}{n}\sum_{r=1}^{n}\left(\frac{y_r}{p_r}\right)\right] = \frac{1}{n}\sum_{r=1}^{n}E\left(\frac{y_r}{p_r}\right)$$

$$= \left(\frac{1}{n}\right)\times n\sum_{i=1}^{N}\left(\frac{Y_i}{p_i}\times p_i\right) = \sum_{i=1}^{N}Y_i = Y_T$$

This shows that the estimator of population total, Y_T, is given by $\hat{Y}_{PPSWR}$ as given in (4.7). Further, the estimator is unbiased for the parameter 'Population Total'.

Remark 4.11: The estimator $\hat{Y}_{PPSWR}$ was proposed by Hansen and Hurwitz in his research publication in 1953.

4.5.2 Estimation of Population Mean

We know that the population mean, $\bar{Y}$ of the study variable is given by

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N}Y_i = \frac{1}{N}\times Y_T \qquad \ldots (4.8)$$

In order to find an estimator of the population mean, let us define the quantity,

$$Z_i = \frac{Y_i}{Np_i}$$

in the population for $i = 1, 2, \ldots, N$.

If a sample of size n is selected from the population, then we can define the sample mean of the variable Z_i as

$$\bar{Z}_n = \frac{1}{n}\sum_{r=1}^{n}\frac{y_r}{Np_r}. \qquad \ldots (4.9)$$

Now, let us prove the unbiasedness property of sample mean $\bar{Z}_n$ in Probability Proportional to Size with Replacement Sampling Scheme.

Theorem 4: In Probability Proportional to Size with Replacement sampling scheme, an unbiased estimator of population mean is given by

$$\bar{Z}_n = \frac{1}{n}\sum_{r=1}^{n}\frac{y_r}{Np_r}$$

Proof: We have

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

$$= \frac{1}{n}\sum_{r=1}^{n}\sum_{i=1}^{N}p_i\left(\frac{Y_i}{Np_i}\right)$$

$$= \frac{1}{n}\sum_{r=1}^{n}\frac{1}{N}\sum_{i=1}^{N}Y_i$$

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

This shows that the estimator of population mean is $\bar{Z}_n$ which is unbiased for the population mean $\bar{Y}$. This completes the proof of the theorem.

Remark 4.12: We can see that if all p_i^s are taken to be (1/N), that is, if we consider the Simple Random Sampling with Replacement scheme, then we have

$$Z_i = \frac{Y_i}{N\left(\frac{1}{N}\right)} = Y_i$$

and $\bar{Z}_n = \frac{1}{n}\sum_{r=1}^{n}\frac{y_i}{Np_i} = \frac{1}{n}\sum_{r=1}^{n} y_i = \bar{y}$

which is the sample mean estimator as defined in Theorem 5 under the Sub-section 1.7.1 in the Unit 1.

This shows that results of Simple Random Sampling with Replacement scheme can be considered to be a special case of Probability Proportional to Size with Replacement scheme. Probability Proportional to Size with Replacement results convert into the results of Simple Random Sampling with Replacement scheme when, instead of taking unequal probabilities p_i^s, we consider equal selection probabilities for all the units of the population.

Try to answer of the following Self-Assessment Question:

SAQ 7

Suggest a suitable estimator for the population total using PPSWR sample of size n and show that it is unbiased estimator for the concerned parameter.

4.6 SAMPLING VARIANCE OF THE ESTIMATORS

In this section we shall derive the expressions of the sampling variance of the estimators obtained in the above section.

4.6.1 Variance of the Estimator of Population Total

In the sub-section 4.5.1(a), we defined the estimator $\hat{Y}_T$ an estimator of population total when sample size was one. It was found that it is an unbiased estimator of the population total Y_T.

Therefore, the variance of the estimator $\hat{Y}_T$ will be given by

$$V\left(\hat{Y}_T\right) = E\left(\frac{y_i}{p_i} - Y_T\right)^2 = \sum_{i=1}^{N} p_i\left(\frac{Y_i}{p_i} - Y_T\right)^2$$

$$= \sum_{i=1}^{N} p_i \frac{Y_i^2}{p_i^2} + \sum_{i=1}^{N} p_i Y_T^2 - 2Y_T \sum_{i=1}^{N} p_i \frac{Y_i}{p_i}$$

$$= \sum_{i=1}^{N} \frac{Y_i^2}{p_i} - Y_T^2$$

Thus, we have

$$V\left(\hat{Y}_T\right) = \sum_{i=1}^{N} \frac{Y_i^2}{p_i} - Y_T^2 \qquad \ldots (4.10)$$

Remark 4.13: We shall now see what happens if instead of taking probabilities of selection proportional to the size measure, an auxiliary variable X; taken proportional to the variable under consideration, Y, even if values of Y are made known. Let us assume that

$p_i \propto Y_i$ for all i.

Then, we have

$$\sum_{i=1}^{N} p_i = k\sum_{i=1}^{N} Y_i \quad \Rightarrow k = \frac{1}{Y_T} \quad \Rightarrow \quad p_i = \frac{Y_i}{Y_T}$$

Substituting this value of p_i in (4.10), we get

$$V\left(\hat{Y}_T\right) = Y_T\sum_{i=1}^{N} Y_i - \left(\sum_{i=1}^{N} Y_i\right)^2 = Y_T^2 - Y_T^2 = 0$$

This shows that if we select probabilities of selection proportional to the study variable, Y, the sampling variance, $V\left(\hat{Y}_T\right)$, of the estimator of population total would be always zero. This is a situation which every researcher would prefer always, because then, the estimated value would be exactly same as the actual value. However, we never get the values Y_i's, as, being the population values, they are always unknown to the sampler. Therefore, the sampler searches an auxiliary variable, X, in place of variable, Y; which is highly correlated with the study variable so that it may best predict the nature of the study variable.

4.6.2 Variance of the Estimator of Population Mean

The unbiased estimator of the population total, Y_T, with sample size was n, was obtained as $\bar{Z}_n$ which is given in (4.9).

Therefore, the sampling variance of the estimator will be

$$V\left(\bar{Z}_n\right) = E\{\bar{Z}_n - E\left(\bar{Z}_n\right)\}^2 = E\left(\bar{Z}_n{}^2\right) - \bar{Z}^2, \qquad \text{where } \bar{Z} = E\left(\bar{Z}_n\right)$$

But,

$$E\left(\bar{Z}_n{}^2\right) = E\left(\frac{1}{n}\sum_{r=1}^{n} Z_r\right)^2 = \frac{1}{n^2}E\left[\sum_{r=1}^{n} Z_r^2 + \sum_{r\neq s}^{n} Z_r Z_s\right]$$

$$= \frac{1}{n^2}E\left(\sum_{r=1}^{n} Z_r^2\right) + \frac{1}{n^2}E\left(\sum_{r\neq s}^{n} Z_r Z_s\right)$$

Now we know that, by definition of expectation

$$E\left(Z_r^2\right) = \sum_{i=1}^{N} p_i Z_r^2 \qquad \text{and} \qquad E\left(Z_r Z_s\right) = E\left(Z_r\right)E\left(Z_s\right)$$

since units are drawn independently. Further, we know that

$$E\left(Z_r\right) = \frac{1}{N}\sum_{i=1}^{N} Z_i = \bar{Z} \qquad \text{and} \qquad E\left(Z_s\right) = \frac{1}{N}\sum_{i=1}^{N} Z_i = \bar{Z}.$$

Therefore, we get

$$V\left(\bar{Z}_n\right) = \frac{1}{n^2} E\left[\sum_{r=1}^{n} Z_r^2 + \sum_{r \neq s}^{n} Z_r Z_s\right] - \bar{Z}^2.$$

$$= \frac{1}{n} E\left(Z_r^2\right) + \frac{n(n-1)}{n^2} E\left(Z_r\right) E\left(Z_s\right) - \bar{Z}^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} p_i Z_i^2 + \frac{n-1}{n} \bar{Z}^2 - \bar{Z}^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} p_i Z_i^2 - \frac{1}{n} \bar{Z}^2 = \frac{1}{n}\left(\sum_{i=1}^{N} p_i Z_i^2 - \bar{Z}^2\right)$$

$$= \frac{1}{n} \sum_{i=1}^{N} p_i \left(Z_i - \bar{Z}\right)^2 = \frac{\sigma_z^2}{n};$$

where $\sigma_Z^2 = \sum_{i=1}^{N} p_i \left(Z_i - \bar{Z}\right)^2$, is, by definition, the variance of the variable Z.

Thus, the variance of the estimator $\bar{Z}_n$ is given by

$$V\left(\bar{Z}_n\right) = \frac{\sigma_Z^2}{n} = \frac{1}{n} \sum_{i=1}^{N} p_i \left\{\frac{Y_i}{N p_i} - \left(\frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{N p_i}\right)\right\}^2 \quad \ldots (4.11)$$

Remark 4.14: We can see that expression (4.11), giving the variance of the estimator of population mean in Probability sampling scheme, reduces to the variance of the estimator of population mean obtained in Simple Random Sampling with Replacement scheme, as obtained in Theorem 8 in Sub-section 1.7.3 of Unit 1, when $p_i = \frac{1}{N}$ for i = 1, 2, …, N.

We use $p_i = \frac{1}{N}$ in (4.11). We, then, have

$$V\left(\bar{Z}_n\right) = \frac{1}{n} \sum_{i=1}^{N} \frac{1}{N} \left\{Y_i - \left(\frac{1}{N} \sum_{i=1}^{N} Y_i\right)\right\}^2$$

$$= \frac{\sigma_Y^2}{n} = \frac{N-1}{N} \cdot \frac{S_Y^2}{n} = V\left(\bar{y}\right).$$

This result is an indication that Simple Random Sampling with Replacement scheme is a special case of Probability Proportional to Size with Replacement scheme and results obtained under Simple Random Sampling with Replacement scheme in Unit 1 are easily deducible from the results obtained under Probability Proportional to Size with Replacement scheme.

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

SAQ 8

Find the expression of the sampling variance of sample mean estimator as obtained under Probability Proportional to Size with Replacement sampling scheme.

4.7 ESTIMATING THE SAMPLING VARIANCE OF ESTIMATOR OF POPULATION MEAN

We shall now prove that there exists an unbiased estimator of the sampling variance of the estimator of population mean in Probability Proportional to Size with Replacement scheme. Let us consider the following theorem:

Theorem 5: In Probability Proportional to Size with Replacement scheme, an unbiased estimator of the sampling variance, $V(\bar{Z}_n)$, is given by

$$\hat{V}(\bar{Z}_n) = \frac{s_z^2}{n}; \qquad \ldots (4.12)$$

where $s_z^2 = \frac{1}{n-1}\sum_{r=1}^{n}(Z_r - \bar{Z}_n)^2$.

Proof: We have given, $s_z^2 = \frac{1}{n-1}\sum_{r=1}^{n}(Z_r - \bar{Z}_n)^2$

Therefore, we have

$E(s_z^2) = \frac{1}{n-1}\sum_{r=1}^{n}E(Z_r^2) - \frac{n}{n-1}E(\bar{Z}_n^2)$ by expanding the term $\sum_{r=1}^{n}(Z_r - \bar{Z}_n)^2$.

But, by definition of variance, we have

$$V(\bar{Z}_n) = E\{\bar{Z}_n - E(\bar{Z}_n)\}^2 = E(\bar{Z}_n^2) - \bar{Z}^2 .$$

$$\Rightarrow E(\bar{Z}_n^2) = V(\bar{Z}_n) + \bar{Z}^2 = \frac{\sigma_z^2}{n} + \bar{Z}^2$$

Therefore,

$$E(s_z^2) = \frac{1}{n-1}\sum_{r=1}^{n}\sum_{i=1}^{N}p_i Z_i^2 - \frac{n}{n-1}\left\{\frac{\sigma_z^2}{n} + \bar{Z}^2\right\}$$

$$= \frac{n}{n-1}\sum_{i=1}^{N}p_i Z_i^2 - \frac{\sigma_z^2}{n-1} - \frac{n}{n-1}\bar{Z}^2$$

$$= \frac{n}{n-1}\left[\sum_{i=1}^{N}p_i Z_i^2 - \bar{Z}^2\right] - \frac{\sigma_z^2}{n-1}$$

$$= \frac{n}{n-1}\left[\sum_{i=1}^{N}p_i (Z_i - \bar{Z})^2\right] - \frac{\sigma_z^2}{n-1}$$

$$= \left[\frac{n}{n-1}\sigma_z^2 - \frac{\sigma_z^2}{n-1}\right] = \sigma_z^2$$

Implying that $E(s_z^2) = \sigma_z^2$. Since $V(\bar{Z}_n) = \frac{\sigma_z^2}{n}$ therefore, the estimated value of $V(\bar{Z}_n)$ will be given by

$$\text{Est. } V(\bar{Z}_n) = \hat{V}(\bar{Z}_n) = \frac{s_z^2}{n}$$

This unbiased estimator of $V(\bar{Z}_n)$, since, s_z^2 s unbiased for σ_z^2.

4.8 SUMMARY

In this unit, we have discussed:

- The reasons as to why sometimes unequal probabilities of selection of units, selected for the sample, are preferred over equal probability of selection method.
- The concept of Varying Probability Sampling Scheme (VPSS).
- A special case of Varying Probability Sampling Scheme, namely, Probability Proportional to Size (PPS) sampling scheme.
- The methods of selection of a sample using Probability Proportional to Size with Replacement (PPSWR) sample. It was also shown theoretically that these methods actually provide the sample with selection probabilities proportional to size.
- The problem of estimation of population total as well as population mean with Probability Proportional to Size with Replacement sample and the properties of these estimators.
- The expression of sampling variances of the estimators of Population Total and Population mean were derived and consequently, it was shown that the results of Simple Random Sampling with Replacement scheme can be derived from the results of Probability Proportional to Size with Replacement scheme.
- The Simple Random Sampling with Replacement scheme is a special case of Probability Proportional to Size with Replacement scheme.
- The unbiased estimator of the sampling variance of the sample mean estimator.

4.9 TERMINAL QUESTIONS

1. State the reasons which enforce a sampler not to prefer always Equal Probability Selection Methods.

2. There are 14 workers in a section of a factory with their salaries varying from Rs. 1000 to 4000 per month. Due to large variations in their salaries, it was decided to select a sample of workers using varying probability selection method. Using Lottery method, select a sample of size 5, if the number of chits associated with each worker is as given in the Table below:

Label Assigned to Workers	1	2	3	4	5	6	7
Number of Chits Associated	8	14	10	20	5	15	10
Label Assigned to Workers	8	9	10	11	12	13	14
Number of Chits Associated	18	9	6	9	13	7	6

3. What is the use of size measures in the sampling?

4. Explain, with the help of the following example, the steps of the Cumulative Total method.

Unit Label	1	2	3	4	5	6
Size	6	10	16	9	7	3

5. Select a sample of size 2 using Lahiri's method for the data given in question 4 above.

6. What are the steps to be performed in Lahiri's method? Show that in this method the units are selected with probability proportional to size.

7. The population mean of the given finite population of size N is to be estimated by selecting a sample of size n following PPSWR sampling scheme. Suggest an unbiased estimator for this purpose and show that it is equivalent to the sample mean estimator used in SRSWR scheme if selection probabilities are taken equal for each unit of the population.

8. "The sampling variance of the sample mean estimator as obtained under PPSWR scheme reduces to that obtained under SRSWR scheme". Is this a correct statement? Prove it.

4.10 ANSWERS / SOLUTIONS

Self-Assessment Questions (SAQs)

1. **Hint:** For the answer to the question, you are referred to Sub-section 4.2.1.

2. **Hint:** For the answer to the question, you are referred to Sub-section 4.2.3.

3. **Hint:** For answering the question, see the Sub-sections 4.3.1.

4. **Hint:** See the Sub-section 4.3.2 for your answer.

5. For answering the question, you are referred to the Sub-section 4.4.1. For the steps of the method, see the Sub-section 4.4.1.

6. For answering the question, consult the Sub-section 4.4.2.

7. **Hint:** For answering the question, you are advised to consult Sub-section 4.5.1(B) and the Theorem 3.

8. **Hint:** For answering the question, consult the Sub-section 4.6.2.

Terminal Questions (TQs)

1. **Hint:** For answering the question, you are referred to the Section 4.1.

2. Prepare the following table showing the data given in the problem along with the sequence of integers allotted to a particular chit and the probability of selection:

Label	No. of Chits Allotted	Cumulative Total	Labels on Chits	Probability of Selection
1	8	8	1 - 8	0.053
2	14	22	9 – 22	0.093
3	10	32	23 - 32	0.067
4	20	52	33 – 52	0.133
5	5	57	53 – 57	0.033
6	15	72	58 – 72	0.100
7	10	82	73 – 82	0.067
8	18	100	83 – 100	0.120
9	9	109	101 – 109	0.060
10	6	115	110 – 115	0.040
11	9	124	116 – 124	0.060
12	13	137	125 – 137	0.087
13	7	144	138 – 144	0.047
14	6	150	145 – 150	0.040
Total	150	-	-	1.000

The third column of the table shows the sequence of integers written on the chits with particular label. Column four presents the Cumulative Totals of number of chits belongs to each label Column five presents the probability of selection of each label, for example:

$p_1 = {}^{8}\!/_{150} = 0.053;\ p_2 = {}^{14}\!/_{150} = 0.093$; and so on.

In order to select 5 labels randomly in the sample, we use the Random Number Table given in the **Appendix - A** in Unit 1 for selecting a number R such that $1 \le R \le 14$. Let us choose randomly sixth row in the table and move row-wise. We select two samples, one, using with replacement and the other, without replacement.

Since, N = 14, we choose the last two digits of each selected random number. Let us select the with replacement sample. From the table we see that random numbers selected are: {10, 13, 4, 5, 6}.

Therefore, in the with replacement sample, we have units as: {U_{10}, U_{13}, U_4, U_5, U_6}.

In order to select without replacement sample, we start with the sixth column and move column-wise. The labels selected are: {1, 14, 3, 10, 5}, indicating that units {U_1, U_{14}, U_3, U_{10}, U_5} are included in the sample.

3. **Hint:** For answering the question, see the Sub-section 4.3.2.

4. The following table is prepared for showing the labels of units, their size measures, cumulative totals for each label, numbers associated with each label and probability of selection of each unit of the population:

Serial Number of Units (Labels i)	Size Measure of Units (X_i)	Cumulative Totals (T_i)	Numbers Associated with Label	Probability of Selection $\left(p_i = \frac{X_i}{X}\right)$
1	6	6	1 – 6	0.118
2	10	16	7 – 16	0.196
3	16	32	17 – 32	0.314
4	9	41	33 – 41	0.176
5	7	48	42 – 48	0.137
6	3	51	49 - 51	0.059
Total	51	---	---	1.00

Steps 1 to 4 of the method are performed in the above table. For completing step 5, we use the Random Number Table given in the **Appendix - A** of Unit 1 for selecting a number R such that $1 \le R \le T_6$ where T_6 =51. We have selected a random number as 7624 (starting with fifth column). Considering the last two digits of the number, which is 24, lying between 1 to 51, is, therefore, selected randomly in the sample at the first draw. Thus, the label 3 is selected. Similarly repeating this process, other units can be selected.

5. We chose the number M, which is maximum X_i. In the problem, therefore, M would be 16. Now, we prepare the columns for effective and ineffective numbers for each unit. We have the following table for this:

Serial Number of Units (Labels i)	Size Measure of Units (X_i)	Effective Numbers	Ineffective Numbers
1	6	1 – 6	7 – 16
2	10	1 – 10	11 – 16
3	16	1 – 16	---
4	9	1 – 9	10 – 16
5	7	1 – 7	8 – 16
6	3	1 – 3	4 – 16
Total	51	---	---

Now, we select two random numbers, u and v, such that $1 \le u \le N$ and $1 \le v \le M$. Here, N = 6 and M = 16, so, we have $1 \le u \le 6$ and $1 \le v \le 16$. Since, u is a one-digit number, we select the last digit of the random numbers and, since, M is a two-digit number, we select last two digits of random numbers. Let us choose the second column and fourth column in Random Number Table, respectively, for selecting u and v. In the random number table, given in the **Appendix - A** given in the Unit 1, we first select u. The first number selected is 3, so label 3 is selected provisionally in the sample. Now we select v. The first value of v obtained in the range 1 -16 is 04, which lies in the effective numbers, therefore, label 3 is finally selected in the sample. For selecting the second unit in the sample, again we select u. We see that it is 4, so label 4 is provisionally selected in the sample. Further, we select v. The next

value of v is obtained as 06, which is an effective number for the label 4, therefore, label 4 is finally selected in the sample. Thus, the selected sample of size 2 will consist of labels {3, 4}.

6. **Hint:** For the answer of the first part of the question, consult the Sub-section 4.4.2. For the second part of the question, see the Theorem 2.

7. **Hint:** For answering the question, see the complete Sub-section 4.5.2.

8. **Hint:** For the answer to the question, you are referred to the Remark 4.14 under the Sub-section 4.6.2.