Unit 6 Sample Survey Methods

6.1 Introduction

The method or process of selecting a sample from a population under study is called **sampling**. There are various methods of sampling techniques used for the selection of sample from the population. The method of selecting a sample from the population is of fundamental importance in sampling theory and depends very much on the objective and scope of the inquiry or investigation, nature of the universe or population, the size of the sample, available resources etc. Sampling methods can be classified into the following broad categories.

i) Random sampling (Probability Sampling)

ii) Non random sampling (Non- Probability Sampling)

iii) Mixed sampling

6.2 Random Sampling (Probability Sampling)

Random sampling or probability sampling is the scientific method of selecting samples from the population according to some laws of chance in which each and every unit in the population has some definite pre-assigned probability of being selected in the sample. Probability samples are selected in such a way so as to be representative of the population. There are various types of sampling in which:

i. Each sample unit has an equal chance of being selected.

ii. Sampling units have different probability of being selected.

iii. Probability of selection of a unit is proportional to the sample size.

6.2.1 Types of Random Sampling

a) Simple Random Sampling: The random sampling in which the units are selected in such a way that each and every unit in the population has equal and independent chance of being selected from the population. It is the simplest and most common method of sampling in which the sample is drawn unit by unit, with equal probability of selection for each unit at each draw. This is also known as the equal probability sampling. It is the most elementary random sampling.

If a sample is to be taken at random, the following two methods can be used:

(i) Lottery method (ii) Use of table of random numbers.

i) Lottery method: It is the simplest method of simple random sampling. Suppose one is interested to select 'n' units out of 'N'. Distinct numbers from 1 to N are assigned for each unit and these numbers are written on 'N' homogenous slips. These slips are put on a box or a bag and mixed thoroughly. Then 'n' slips are drawn one by one.

ii) Use of table of random numbers: Another easiest way of selecting samples is through the use of random number tables. These random numbers are generally generated by computer or by a table of random numbers. Tippet's random number tables, Fisher and Yates tables, Kendall and Smith's tables are some of commonly used random number tables. This method can also be considered as better than the lottery method since lottery method is time consuming.

There are two types of simple random sampling (*SRS*)

a. **SRSWOR:**

If the unit selected in any draw is not replaced in the population before making the next draw, then it is known as simple random sampling without replacement (SRSWOR).

The probability of selecting (drawing) of a sample of size '*n*' from a population of size *N* in SRSWOR is $\frac{1}{N_{C_n}}$.

Note: If the sample is drawn without replacement, then the standard error of sample mean is

$$\text{S.E. }(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Similarly, if the sample is drawn without replacement, the standard error of the sample proportion of is

$$S.E.(P) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$$

b. **SRSWR:** If a unit is selected and noted and then returned back or replaced back in the population before making the next draw, then the sampling procedure is called simple random sampling with replacement (SRSWR). The probability of selecting (drawing) of a sample of size *n* from a population of size *N* in SRSWR is $\frac{1}{N^n}$.

Note: If the sample is drawn with replacement, then the standard error of sample mean is

$$\text{S.E. }(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Similarly, if the sample is drawn with replacement, the standard error of the sample proportion is

$$\text{S.E. }(p) = \sqrt{\frac{PQ}{n}}$$

Merits

i) Each item has equal chance of being selected. So, it depends upon the chance but not on the personal judgment.

ii) This method is quite economic and comparatively saves time and money.

iii) It is more representative of the population as compared to the judgment or purposive sampling.

Demerits

i) Expensive and time consuming especially when the population is large.

ii) If the population is heterogeneous in nature, it may not give accurate results.

iii) It requires completely up-to-date list of population units from which samples to be drawn.

b) **Stratified Sampling:** Stratified random sampling is used when the characteristics of elements of population are heterogeneous. In a stratified random sampling approach, a population is first divided into subpopulations or strata, based upon one or more classification criteria in such a way that the characteristics of units **are homogeneous within strata and heterogeneous between strata**. Then sample is drawn from each stratum using simple random sampling method. These samples are combined (pooled) to form the required sample of the population. Thus, the procedure in which first stratification and then simple random sampling is known as stratified random sampling. The number of units to be sampled from each stratum depends on its size relative to the population.

For example, if we want to study about the examination results of 2000 students of Mechi Multiple Campus Jhapa. At first divide the no. of students into different faculties such as management, humanities and science. If the number of students in management, humanities, and science are 800, 700 & 500 respectively. Then we select 20% of each for the samples i.e. selecting samples consisting of 160, 140 & 100 students from each faculty respectively. The samples taken from each stratum are then pooled (combined) to form the overall sample (required sample). Similarly, in survey of average crop yield per hectare etc.

A stratified sample is a mini-reproduction of the population. Before sampling, the population is divided into characteristics of importance for the research. For example, by gender, social class, education level, religion, etc. Then the population is randomly sampled within each category or stratum.

A stratified sampling may be either (i) Proportionate or (ii) Disproportionate

In a proportionate stratified sampling, the number of items drawn from each stratum is proportional to the size of the strata. On the other hand, if an equal number of items are drawn from each stratum regardless of how the stratum represented in the population (universe), it is called disproportionate sampling.

Merits:

i) The units selected represents whole universe.

ii) The estimation of population parameters is more efficient.

iii) For large and heterogeneous population, stratified sampling is the best design.

Demerits

i) This method requires more time and cost.

ii) If each stratum of the population is not homogeneous the result obtained may not be reliable.

iii) The samples from each stratum should be selected only by the experts or experienced persons.

c) Systematic Sampling: A random sampling in which only the first unit is selected at random and the remaining units are automatically selected according to pre-determined pattern (i.e. at fixed equal intervals from one another) is known as systematic sampling. Systematic sampling is a commonly used technique if the complete and up-to-date sampling frame is available i.e. complete and up-to-date list of sampling units is available.

Suppose N units of the population are numbered from 1 to N in some order. Let $N = n\ k$ where n sample size and K is an integer known as sampling interval. Thus, $k = \frac{N}{n}$. Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. This is random sampling with a system. From the sampling frame, a starting point is chosen at random and choice thereafter are at regular intervals. For example, suppose you want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116. This sampling is mostly used in forest surveys, fisheries surveys etc.

Merits

i) This method is simple and convenient to use.

ii) In selecting the sample by this method, it takes less time and labour.

iii) Most of the results obtained from this method are satisfactory.

iv) If the complete list of the population is available and if the items are arranged systematically, this method is more efficient.

Demerits

i) The sample selected may not be the representative of the population in some cases.

ii) The items of the population must be arranged in some order otherwise the result obtained will be misleading

d) Cluster Sampling: A Cluster sampling is a method or technique of random sampling in which the population is divided into different groups, called clusters, in such a way that the characteristics of **units within the cluster are heterogeneous and between the clusters are homogenous** so that the number of sampling units in each cluster should be approximately same. Then a cluster is selected as a sample by using simple random sampling. This method of random sampling is called cluster sampling. In cluster sampling, individual clusters are representative of the population as a whole. For example, Let us suppose that we want to study about the economic condition of people of Kathmandu metropolitan city. First of all, metropolitan city is divided into different wards in such a way that the economic condition of people within ward is heterogeneous and between wards are homogeneous. Then, a ward is selected as sample by using simple random sampling method and we can study about the economic condition of people in Kathmandu metropolitan city.

Merits

i) It is less costly than simple random sampling and stratified sampling

ii) It is useful even when the sampling frame of elements may not be available.

iii) Elements (units) selection by well-designed cluster sampling procedures is easier, faster cheaper and more convenient than simple random sampling and stratified sampling.

Demerits

i) The efficiency decreases with increase in cluster size.

ii) The efficiency cost per unit may be more in cluster sampling.

iii) Enumeration of the sampling units within the selected clusters is difficult when the population is large.

e) Multistage Sampling: Multi-stage sampling is a further development of the principle of cluster sampling. Multi-stage sampling is also a random sampling in which sampling procedure is carried out i.e. done in various stages. At first stage, the population is divided into large sample units i.e. large groups called first stage units (fsu) or primary stage units and cluster is selected as the sample at random from them. At the second stage the selected clusters at the first stage are further divided or sub-sampled into smaller sample units. The method selecting only some of the units of selected cluster is called two –stage sampling. And if it is generalized into third stage or more stages until get to ultimate units of sample size is known as multi-stage sampling. Multistage sampling is a complex form of cluster sampling. Although cluster sampling and stratified sampling bear some superficial similarities, they are substantially different. In stratified sampling, a random sample is drawn from all the strata, where in cluster sampling only the selected clusters are studied, either in single- or multi-stage.

For example, in crop surveys for estimating yield to a crop in a district, VDC can be considered as primary sampling unit (psu), crop fields as second stage units and a plot of fixed size as the ultimate unit of sampling.

Merits:

i) It is more convenient when area of investigation is very large.

ii) It is commonly used in large scale survey.

iii) This method is also more flexible than other sampling methods.

iv) As sample size is reduced in each stage, this sampling technique saves time and cost.

Demerits:

i) If the samples are not carefully taken from the different stages, it may give the faulty result.

ii) In this method, there is high chance of occurring sampling error when the selected sampling units are decreased.

6.2.2 Non Random Sampling (Non- Probability Sampling)

Non-random sampling is the method of selecting samples, in which the choice of selection of sampling units depends entirely on the discretion or judgment or convenience, beliefs, biases of sampler or investigator. This method is mainly used for opinion surveys but cannot be recommended for general use as it is subject to the drawbacks of prejudice and bias of the investigator. This is method of selecting of selecting samples in which the choice of selection of sampling units depends entirely on the discretion or judgment of sampler or investigator. This method is mainly used for opinion surveys but can not be recommended for general use as it is subject to the drawbacks of prejudice and bias of the investigator. However, if the researcher is experienced and expert , it is possible that judgment sampling may yield useful results. However, this method suffers from a serious defect that it is not possible to compute the degree of precision of estimate from the sample values.

Types of Non-random sampling or Non-probability sampling are as follows:

a) Judgment sampling: A sampling method, in which the choice of sample items depends entirely upon the judgement of the investigator is called judgement sampling. In this method of sampling, the choice of sampling items depends exclusively on the judgement of the investigator.

In other words, the investigator uses self judgment in the choice and includes only those items of the universe which are convenience to the investigator. It is the method for quick decision.

For instance, if we want to study of corruption in Nepalese society, we can select a sample of twenty of the senior professors of T.U. to give their opinion on the subject. We consider that the judgement of these professor is much superior to a convenience. Then we can get the desired information.

Merits:

i) It is the simple method of sampling for quick decision.

ii) It gives the better result when sample size is small.

Demerits:

i) It gives unreliable conclusion if the investigator is personally biased.

ii) Though simple, the method is not scientific and it is not in general use.

iii) Sampling error can not be estimated because it is not based on random sampling.

b) Convenience sampling (Accidental sampling): A sampling method, in which the researcher selects the sample neither by probability nor by judgment but by convenience, is called convenience sampling. It is also called the accidental sampling. Selection of sampling units is totally based on the convenience of the researcher. The results obtained by this method can hardly be representative of the population. They are generally biased and unsatisfactory. However, convenience sampling is often used for making pilot studies. For instance, if any one wants to conduct 'man-on-the-street' interviews, he/she stands up in corner of the street and interviews the desired number of passers-by. Then, required information can be obtained.

Merits:

i) It is useful for making pilot studies.

ii) It is the simple method of sampling for quick decision.

iii) When both time and money are limited, convenience sampling is widely used i.e. it is less expensive and less time consuming.

Demerits:

i) The results obtained by this method can hardly be representative of the population.

ii) Sampling error cannot be estimated because it is also not based on random sampling.

c) Quota Sampling: A non-random sampling method in researcher are given quotas to be filled from the different strata and within pre-assigned quotas, the process of drawing selecting the required samples from these strata by judgment sampling is called quota sampling. Quota sampling is a type of judgment sampling. Sample quotas may be fixed according to some specified characteristics such as income group, sex, occupation, political or religious affiliation etc.

For instance, for the comment about the fiscal year budget in radio listening survey, the interviewer may select a sample of 50 persons choosing from different areas such as 20 officials, 10 Professors, 10 businessmen, 5 farmers and 5 students. Here, interviewer is free to select the people to be interviewed for the comment.

Merits:

i) It saves time and money rather than other sampling methods.

ii) It is stratified–cum-purposive so investigator enjoys the benefits of both.

Demerits:

i) It may be biased because of the personal believes and prejudices of investigator.

ii) Sampling error cannot be estimated because it is also not based on random sampling

d) Snowball sampling or Network sampling: Snowball sampling is a special type of non-probability sampling and is used when the desired sample characteristic is very rare. Therefore, this sampling design is widely used in applications where respondents are difficult to identify and are best located through referral networks. Hence, this sampling is also known as **chain referral sampling or network sampling**. In this sampling, an initial group is discovered and then subsequent respondents, possessing similar characteristics are identified based on referrals provided by the initial respondents.

This sampling is particularly used to study drug cultures (use), teenage gang activities, prostitution study, political activities, illegal activities etc.

Merits:

i) This method is cheap, simple and cost –efficient.

ii) It is useful for rare populations for which no sampling frames are readily available.

iii) The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods.

Demerits:

i) It is difficult to apply when the population is large.

ii) It does not ensure the inclusion of all elements in the list

Workout of Example

Example 6.2 A population variable consists the values: 1, 2, 3, 4, 5.

i) Draw all possible samples of size two which can be drawn from the population without replacement.

ii) Show that the mean of the sampling distribution of the sample mean.

1, 2, 3, 4, 5,

1, 2, 3, 4, 5,

Solution:

i) Population size (N) = 5

Sample size (n) = 2

Possible number of samples of size 2 which can be drawn from the population without replacement = ${}^{N}C_n = {}^{5}C_2 = 10$

Possible samples are (1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5), (4, 5).

ii) Population mean (μ) = $\dfrac{1+2+3+4+5}{5} = 3$

Calculation of sample distribution of sample means

Sample No.	Sample value (X)	Sample mean ($\bar{X}$)
1	(1, 2)	1.5
2	(1, 3)	2
3	(1, 4)	2.5
4	(1, 5)	3
5	(2, 3)	2.5
6	(2, 4)	3
7	(2, 5)	3.5
8	(3, 4)	3.5
9	(3, 5)	4
10	(4, 5)	4.5
		$\Sigma(\bar{X}) = 30$

Mean of sample means ($\bar{\bar{X}}$) = $\dfrac{\Sigma\bar{X}}{10} = \dfrac{30}{10} = 3 = \mu$

Hence, mean of sampling distribution of means is equal to the population mean i.e.,

$$\bar{\bar{X}} = \mu \quad \text{i.e.,} \quad E(\bar{X}) = \mu$$

iii) Calculation of standard error of the sampling distribution of sample mean

Sample mean ($\bar{X}$)	$\bar{X} - \bar{\bar{X}} = \bar{X} - 3$	$(\bar{X} - \bar{\bar{X}})^2$
1.5	–1.5	2.25
2	–1	1
2.5	–0.5	0.25
3	0	0
2.5	–0.5	0.25
3	0	0
3.5	0.5	0.25
3.5	0.5	0.25
4	1	1
4.5	1.5	2.25
		$\Sigma(\bar{X} - \bar{\bar{X}})^2 = 7.5$

Now, standard error of mean is

$$\text{S.E.}\ (\bar{X}) = \sqrt{\text{Var}\ (\bar{X})} = \sqrt{\frac{\Sigma(\bar{X} - \bar{\bar{X}})^2}{{}^{N}C_{n}}} = \sqrt{\frac{7.5}{10}} = 0.866$$

Example 6.3 Let the five observation be 3, 4, 5, 6,7 of a universe. (i) select all sample of size 2. (ii) compute their means. (iii) compare the mean of other sample means with the mean of the universe. (iv) calculate the standard error of the sampling distribution of the sample mean.

Solution: Here,

Population size, $N = 5$

Sample size, $n = 2$

Population observations $= 3, 4, 5, 6, 7$

i) Possible number of sample of size 2 which can be drawn from the population without replacement is ${}^{N}C_{n} = {}^{5}C_{2} = 10$

Possible samples are: (3, 4), (3, 5), (3, 6), (3, 7), (4, 5), (4, 6), (4, 7), (5, 6), (5, 7), (6, 7)

ii) Calculation of sample means

Sample number	Sample value (X)	Sample means ($\bar{X}$)
1	(3, 4)	3.5
2	(3, 5)	4
3	(3, 6)	4.5
4	(3, 7)	5
5	(4, 5)	4.5
6	(4, 6)	5
7	(4, 7)	5.5
8	(5, 6)	5.5

9	(5, 7)	6
10	(6, 7)	6.5
		$\Sigma\bar{X} = 50$

iii) Calculation of mean of universe and mean o the sample means.

$$\text{Mean of universe} = \text{Population mean} = \mu = \frac{3+4+5+6+7}{5} = 5$$

$$\text{Mean of the sample mean} = \bar{X} = \frac{\Sigma\bar{X}}{{}^{N}C_{n}} = \frac{50}{10} = 5$$

Hence, of the sample means is equal to population mean i.e, $\bar{\bar{X}} = \mu$ i.e., $E(\bar{X}) = \mu$

iv) Calculation of standard error of the sampling distribution of sample mean.

Sample number $\bar{X}$	$\bar{X} - \bar{\bar{X}} = \bar{X} - 5$	$(\bar{X} - \bar{\bar{X}})^2$
3.5	–1.5	2.25
4	–1	1
4.5	–0.5	0.25
5	0	0
4.5	–0.5	0.25
5	0	0
5.5	0.5	0.25
5.5	0.5	0.25
6.	1	1
6.5	1.5	2.25
		$\Sigma(\bar{X} - \bar{\bar{X}})^2 = 7.5$

Standard error of sampling distribution of sample mean is

$$\text{S.E.}(\bar{X}) = V(\bar{X}) = \sqrt{\frac{\Sigma(\bar{X} - \bar{\bar{X}})^2}{{}^{N}C_{n}}} = \sqrt{\frac{975}{10}} = \sqrt{0.75} = 0.866 = 0.87$$

Example 6.4 A random sample of size 36 from a finite population consisting 101 units. If the population standard deviation is 12.6, find the standard error of sample mean when the samle is drawn (i) with replacement (ii) without replacement.

Solution: Sample size, $n = 36$; population size, $N = 101$

Population s.d., $\sigma = 12.6$

i) If the sample is drawn with replacement, then the standard error of sample mean is

$$\text{S.E. }(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{12.6}{\sqrt{36}} = 2.1$$

ii) If the sample is drawn without replacement, then the standard error of sample mean is

$$\text{S.E. }(\bar{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{12.6}{\sqrt{36}}\sqrt{\frac{101-36}{101-1}} = 1.69$$

Example 6.5 A simple random sampling of size 9 is drawn without replacement from a finite population consisting of 25 units. If the number of defective units in the population be 5, find the standard error of the sample proportion of defective.

Solution: Here,

Sample size, $n = 9$

Population size, $N = 25$

Population proportion of defective units, $P = \frac{5}{25} = \frac{1}{5}$

$Q = 1 - p = \frac{4}{5}$

If the sample is drawn without replacement, the standard error of the sample proportion of defectives is

$$S.E.(P) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{\frac{1}{5}\times\frac{4}{5}}{9}}\sqrt{\frac{25-9}{25-1}} = \sqrt{\frac{4}{225}}\sqrt{\frac{16}{24}} = 0.1089$$

Example 6.6 A random sample of 500 oranges was taken from a large consignment and it was observed that 65 were found to be bad. Find the standard error of bad oranges.

Solution: Here, Sample size, $n = 500$

Bad sample $= 65$

Sample proportion, $p = \frac{65}{500} = 0.13$

$q = 1 - p = 0.87$

Now, the standard error of sample proportion of bad oranges (for large population) is

$$\text{S.E. }(p) = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{pq}{n}} \quad (\hat{p} = P \text{ for large samples})$$

$$= \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$$

Example 6.7 A population consists of five numbers 1, 3, 5, 7 and 9

(i) Enumerate all possible samples of the size two which can be drawn from the population without replacement.

(ii) Calculate the mean and variance of the population.

(iii) Show that the mean of the sampling distribution of the sample means is equal to the population mean.

(iv) Calculate the variance of the sampling distribution of sample mean.

(v) Standard error of mean

Solution: Here, $N = 5$, $n = 2$

i. Number of possible sample of size 2 can be drawn from the population without replacement ${}^{N}C_{n} = {}^{5}C_{2} = 10$

Thus the possible samples are: (1, 3), (1, 5), (1, 7), (1, 9), (3, 5), (3, 7), (3, 9), (5, &), (5, 9), (7, 9)

ii. Calculation of population mean and population variance:

Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
1	– 4	16
3	– 2	4
5	0	0
7	2	4
9	4	16
$\sum Y = 25$		$\sum(Y - \bar{Y})^2 = 40$

Population mean = $\bar{Y} = \frac{\Sigma Y}{N} = \frac{25}{5} = 5$

Population variance = $\frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{40}{5} = 8$

iii. Calculation of sample means and variance of the sampling distribution of means:

Sample no.	Sample values (y)	Sample means ($\bar{y}$)	$\bar{y} - \bar{\bar{y}} = \bar{y} - 5$	$(\bar{y} - \bar{y})^2$
1	(1, 3)	2	– 3	9
2	(1, 5)	3	– 2	4
3	(1, 7)	4	– 1	1
4	(1, 9)	5	0	0
5	(3, 5)	4	– 1	1
6	(3, 7)	5	0	0
7	(3, 9)	6	1	1
8	(5, 7)	6	1	1
9	(5, 9)	7	2	4
10	(7, 9)	8	3	9
		50		30

Mean of the sample means ($\bar{\bar{y}}$) = $\frac{\sum \bar{y}}{10} = \frac{50}{10} = 5$

Since the mean of the sample means ($\bar{\bar{y}}$) = 5 equal to the population mean $\bar{Y}$ = 5, we conclude that the mean of the sampling distribution of the sample means is equal to the population mean.

iv. Variance of the sample means is

$$V(\bar{y}) = \frac{1}{n}\sum(\bar{y} - \bar{\bar{Y}})^2 = \frac{30}{10} = 3$$

v. The standard error of sample mean is given by

$$\text{S.E.}(\ \bar{y}) = \sqrt{V(\bar{y})} = \sqrt{3} = 1.732$$

Example 6.8 Enumerate all possible samples of size 2 taken from the population whose elements 6, 4, 3, 7 and 5 by simple random sampling with replacement and find mean and variance of sampling distribution of sample mean.

Solution: Population size $N = 5$

Sample size $x = 2$

Possible number of sample of size z which can be drawn from the population with replacement is $N^n = 5^2 = 25$ ways.

Possible samples are

(6, 6)	(6, 4)	(6, 3)	(6, 7)	(6, 5)
(4, 6)	(4, 4)	(4, 3)	(4, 7)	(4, 5)
(3, 6)	(3, 4)	(3, 3)	(3, 7)	(3, 5)
(7, 6)	(7, 4)	(7, 3)	(7, 7)	(7, 5)
(5, 6)	(5, 4)	(5, 3)	(5, 7)	(5, 5)

Mean and variance are calculated by same process.

6.2.3 Mixed Sampling

If the samples are selected partly according to some laws of chance and partly according to a fixed sampling rule (i.e. no assignment of probabilities) they are termed as fixed samples and the technique of selecting such samples is known as mixed sampling.

6.3 Ratio and Regression Method of Estimation under Simple and Stratified Random Sampling

6.3.1 Introduction of Auxiliary Information Ratio and Regression Estimation

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. The knowledge of auxiliary information may be exploited at the estimation stage. The estimator can be developed in such a way that is makes use of this additional information. Ratio estimator, regression estimator are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sapling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost and more time consuming.

6.3.2 Method of Estimation Under Simple Random Sampling

The theory of SRS is based on only estimates related to sample arithmetic means of observed value in the sample that is average value. There are other method of estimation which makes use of the information as an auxiliary variable which is highly correlated with the variable under study such method are ore precise (accurate) and give more reliable estimate of the population values than those based on simple average. The methods are :

(i) Ratio method of estimation (ii) Regression method of estimation

6.3.3 Ratio Estimator

In the ratio method, an auxiliary variate x_i correlated with y_i, is obtained for each unit in the sample. The population X of x_i must be known. In practice, x_i is often the value of y_i at some previous time when a complete census was taken. The aim in this method is to obtain increased precision by taking advantage of the correlation between y_i and x_i. If there exists high correlation between x_i and y_i then ratio $\frac{y_i}{x_i}$ varies little from unit to unit.

Let us assume simple random sample of size n is drawn from the population. Let $\bar{y}$ and $\bar{x}$ be the sample means of y_i and x_i respectively ($i = 1, 2, 3, \cdots, n$). y and x be the sample total of y_i and x_i respectively. Also let $\bar{X}$ be the population mean of X_i and X be the population total of X_i then the ratio estimate of population total Y of Y_i is $\bar{Y}_R = \frac{y}{x} X = \frac{\bar{y}_i}{\bar{x}_i} X$ where, y and x are the sample totals of y_i and x_i respectively.

If the quantity is estimated to be $\bar{Y}$, the population mean value of y_i, the ratio estimate is $\bar{Y}_R = \frac{y}{x} \bar{X}$.

Similarly, in case of stratified random sampling $\hat{Y}_{Rs} = \sum_i \frac{\bar{y}_i}{\bar{x}_i} X_i$ where y_i and x_i are the sample total in the i^th stratum and X_i is the Stratum total.

Notation and Terminology

x_i = Auxiliary variable

y_i = Variable under study

x = Sample total of x_i

y = Sample total of y_i

$\bar{x}$ = Sample mean of x_i

$\bar{y}$ = Sample mean of y_i

X_i = Auxiliary population variable

X = Population total of X_i

$\bar{X}$ = Population mean of X_i

$\hat{Y}_R$ = Ratio estimate of population total $= \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X$

$\hat{Y}_R$ = Ratio estimate of population mean $= \frac{y}{x} \bar{X} = \frac{\bar{y}}{\bar{x}} \bar{X}$

$B(\hat{R})$ = Bias of ratio estimate

6.4 Ratio Estimate Under Stratified Sampling

Let y_{ni} be the variable under study and x_{ni} be the auxiliary variable. Let the population be stratified into L strata as follows:

Strata	**Value of Y_N in the population**			
1	Y_{11}	Y_{12}	...	Y_{1N_1}
1	Y_{21}	Y_{22}	...	Y_{2N_2}
L	Y_{L1}	Y_{L2}	...	Y_{LN_L}

Where $N_1 + N_2 + N_3 + \ldots + N_L = N$ (size of population)

Let the sample be drawn from given population from all strata.

Strata	**Value of y_{ni} in sample**			
1	y_{11}	y_{12}	…	y_{1n_1}
2	y_{21}	y_{22}	…	y_{2n_2}
L	y_{L1}	y_{L2}	…	y_{Ln_L}

Where $N_1 + n_2 + n_3 + \dots + n_L = n$

Let the corresponding value of auxiliary variable x_{ni} in the sample are

Strata	**Value of x_{ni} in sample**			
1	x_{11}	x_{12}	…	x_{1n_1}
2	x_{21}	x_{22}	…	x_{2n_2}
L	x_{L1}	x_{L2}	…	x_{Ln_L}

$$n_1 + n_2 + n_3 + \cdots + n_L = n$$

... Situation there are two ways of obtaining a ratio estimate of the population total these

(i) Separate ratio estimate (ii) Combined ratio estimate

Separate Ratio Estimate

$$X_y = \sum_{i=1}^{N_i} X_{N_i} = \text{stratum total}$$

$$x_n = \sum_{i=1}^{N_i} X_{N_i} = \text{sample total of } y_i \text{ of } i^{\text{th}} \text{ stratum}$$

$$y_n = \sum_{i=1}^{N_i} X_{N_i} = \text{sample total of } y_i \text{ of } i^{\text{th}} \text{ stratum}$$

Then separate ratio estimate of population is given by $\hat{Y}_{RS} = \sum_{n=1}^{L} \frac{y_n}{x_n} X_n = \sum_{n=1}^{L} \frac{y_n}{x_n} X_n$

6.5 Combined Ratio Estimate

Let $\hat{Y}_{st} = \sum_{i=1}^{L} N_i \overline{Y}_i$ = stratified estimate of the population total Y

$\hat{X}_{st} = \sum_{i=1}^{L} N_i \overline{X}_i$ = stratified estimate of the population total X

X = population total, then combined ratio estimate of population is given by

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\overline{X}_{st}} X = \frac{\hat{Y}_{st}}{\overline{X}_{st}} X, \text{ where } \overline{Y}_{st} = \hat{Y}_{st}/N \text{ and } \overline{X}_{st} / N$$

6.6 Regression Estimator Under Stratified Random Sampling

Like a ratio estimate, the linear regression estimate is designed to increase precision by the use of an auxiliary variate x_i that is correlated with y_i. When the relation between y_i and x_i ··· examined, it may be found that although the relation is approximately linear, the line does not go through the origin. This suggests an estimate based on the linear regression of y_i on x_i rather than on the ratio of two variables.

We suppose that y_i and x_i are each obtained for every unit in the sample and that the population mean $\overline{X}$ of x_i is known. The linear regression estimate of $\overline{Y}$, the population mean of y_i, is $\overline{y}_{ir} = \overline{y} + b(\overline{X} - \overline{x})$

Where $\bar{y}$ and $\bar{x}$ are sample mean of y_i and x_i respectively. $\bar{X}$ be the population mean of x_i's. The subscript $1r$ denotes linear regression and b is an estimate of the change in y when x is increased by unity.

If $b = 0$ then, $\bar{y}_{1r} = \dfrac{\sum_{i=1}^{L} y_i}{n}$

If $b = \dfrac{\bar{y}}{\bar{x}}$

then $\bar{y}_{1r} = \bar{y} + \dfrac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \dfrac{\bar{y}}{\bar{x}} = \bar{Y}$

The regression estimate of the population is $\bar{Y}$.

6.7 Probability Proportion to Size Sampling (PPS)

In the case of simple random sampling the selection probabilities are equal on all the units of the population. If the units vary considerably in sizes, the simple random sampling is not appropriate method because in such situation, this method will not give the importance of large units in the population. In order to consider the importance of large units, an auxiliary information can be utilized in selecting sample to get more efficient estimator of the population parameters. One simple method is to assign unequal probabilities of selection of different units in the population depending on their sizes.

Some auxiliary characteristics x closely related to main character y of interest are available for all the units of the population. For example, the village with larger geographical area are likely to have larger population and larger area under food crops. It may be desired to provide a sampling scheme in which villagers are selected with probability proportional to their population or to their geographical areas.

When the units vary in their sizes and variates under study is highly correlated to their sizes of the units, the probability proportional to the size (PPS) sampling can be used to have an efficient estimator.

Technique of selecting a sample.

Two different methods are used to get desirable sample of specific size.

Cumulative total method (Cumulative method)

To draw a sample of size n from a population of size N with probability proportional to their sizes, proceed as follows:

Let the size to i^{th} unit be x_i ($i = 1, 2, \cdots, n$) the total of this size be $X = \sum_{i=1}^{n} X_i$.

The number 1 to x_1 is associated with the first unit and $x_1 + 1$ to $x_1 + x_2$ with the second unit and so on. A random number R is selected from a random number table. If $x_1 + x_2 + \cdots + x_{i-1} \leq x_1 + x_2 + \cdots + x_i$, then the item is selected associated with this random number as a sample unit. Note that $R < X$.

For example:

Let us consider a village with 10, holdings consisting of following field area under paddy of 50, 30, 45, 25, 40, 26, 44, 35, 28, 27 ropanies respectively. Select a sample of four holdings with replacement methodizing PPs as follows.

First of all construct successive cumulative totals.

TABLE

Selection of *pps* sample using cumulative total method

Holdings	Size (X_i)	Cumulative size

1	50 (x_1)	50
2	30 (x_2)	80
3	45 (x_3)	125
4	25 (x_4)	150
5	40 (x_5)	190
6	26 (x_6)	216
7	44 (x_7)	260
8	35 (x_8)	295
9	28 (x_9)	323
10	27 (x_{10})	350

Suppose, random number R = 272 is selected. Here R is less than 350, 272 lies between 261 and 350, which corresponds to 8th item is selected. This process is repeated till 4 units are selected.

Lahiri's Method

Main drawback of the cumulative method is that construction of range for the items is time consuming as well as costly, if population consists of large number of units.

As an alternative to this method. Lahiri (1951) developed a simple method which does not require to construct cumulative total.

In this procedure, a random number is selected in between 1 to N, the size of item associated with this selected random number is provisionally noted. Again, a random number which lies in between 1 to M, where M is the maximum value of the items, is selected. If the second random number us less than the values of provisionally noted items then that item is considered as a sample unit.

If $1 \leq i \leq N$, note that the item and its values provisionally say X_i. For a second selection, $1 \leq i \leq N$, note that the item and is values provisionally say X_i. For a second selection, $1 \leq j \leq M$, M = maximum of X_i. Now if $j < X_i$, then whose process is repeated. This process is repeated till desirable size of sample is obtained.

Table

Selection *pps* sample using Lahiri's method

Random no. *i*	Random no. *j*	Observation	Units selected
7	38	$R < x_7$	7
1	45	$R < x_1$	1
7	49	$R > x_7$	
9	38	$R > x_9$	
10	25	$R < x_{10}$	10
5	49	$R > x_5$	
6	32	$R > x_6$	
7	49	$R > x_7$	
4	38	$R > x_4$	

5	10	$R > x_5$	5

The selected samples are 7, 1, 10 and 5.

Estimation of Population Total and its Variance

Let us consider a population of size N and y_i be the value of i^{th} unit in the population ($i = 1, 2, \cdots, N$) of the study variate. Suppose the probability proportional to size $p_i = \frac{x_1}{X}$ be the probability associated with i^{th} unit u_i, where X is the total for auxilary auxiliary variate so that $\sum_{i=1}^{N} p_i = 1$.

Let 'n' independent selection be made with replacement method and the value of y_i for each selected unit be observed. Let (y_i, p_i) be the value and its probability for the i^{th} observation then $\frac{y_i}{p_i}$ will be independently and identically distributed. Let us consider a problem of estimating a population mean $\bar{Y}$ and population total $\bar{Y}$ based on the sample of size 'n' using *PPS* with replacement.

To estimate $\bar{Y}$, let us define $z_i = \frac{y_i}{Np_i}$ as an estimator of population $\bar{Y}$. Then,

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{\sum_{i=1}^{n} \frac{y_i}{Np_i}}{n}$$

Since, z_i can take any value of N unit in the population with probability proportional to the size p_i.

$$E(z_i) = \sum_{i=1}^{n} p_i\, z_i = \sum_{i=1}^{n} p_i \frac{y_i}{Np_i} = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{Y}$$

Now the estimate of $\bar{Y}$,

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

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

In the case of *PPS* sampling the estimate $\bar{z} = \frac{\sum_{i=1}^{N} \frac{y_i}{Np_i}}{n}$ is an unbiased estimate of the population mean $\bar{Y}$.

Again, the variance of z_i can be defined as

$$V(z_i) = E(z_i - \bar{Y})^2 = \sum_{i=1}^{N} p_i\, (z_i - \bar{Y})^2 = \sum_{i=1}^{N} p_i \left(\frac{y_i}{Np_i} - \bar{Y}\right)^2.$$

Since, the units are selected with replacement,

$$V(\bar{z}) = V\left(\frac{\sum_{i=1}^{N} z_i}{n}\right) = \frac{\sum_{i=1}^{N} V(z_i)}{n^2} = \frac{\sigma^2_z}{n} = \frac{\sum_{i=1}^{N} p_i \left(\frac{y_i}{Np_i} - \bar{Y}\right)^2}{n}$$

6.8 χ^2, *t*, *F* Distribution

6.8.1 Chi Squared Distribution

The square of a standard normal variate is known as chi-squared variate with 1 d.f. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma}$ is $N(0, 1)$ and $Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$ is a chi-squared variate with 1 d.f.

In general, if X_i; $i = 1, 2, ..., n$, are n independent normal variates with mean μ and variance σ^2, $i = 1, 2, \cdots, n$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \text{ is a chi-square variate with } n \; d.f.$$

In other words,

$$Z^2 = \frac{X - \mu}{\sigma} \sim N(0, 1) \qquad \text{then,} \qquad Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2_{(1)}$$

and when $X_i \sim \mathrm{N}(\mu_i, \sigma^2{}_\mathrm{i})$ then $\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2_{(1)}$

Its probability density function is given by

$$f(x) = \begin{cases} \dfrac{e^{\frac{n}{2}} x^{\frac{n}{2} - 1}}{2^{\frac{n}{2}} \left|\overline{\frac{n}{2}}\right.} & \text{for all } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

This is a probability density function of gamma distribution with parameters $n/2$ and $1/2$. So, gamma distribution is special case of Gamma distribution with $\alpha = n/2$ and $\beta = 1/2$.

Properties

1. Mean and variance of chi-square distribution with n degree of freedom is n and $2n$ respectively.
2. Mode is $n - 2$ for $n > 2$.
3. Karl Pearson coefficient of Skewness is $\sqrt{\frac{2}{n}}$
4. $\beta_1 = \frac{8}{n}$ and $\gamma_1 = 2\sqrt{\frac{2}{n}}$
5. $\beta_2 = 3 + \frac{12}{n}$ and $\gamma_2 = \frac{12}{n}$
6. It does not have parameters.
7. Sum of two chi square variate is chi square variate i.e., if $X_1 \sim \chi^2_n$ and $X_1 \sim \chi^2_n$. Then, $X_1 + X_2 \sim \chi^2_{n+m}$
8. Moment generating function of χ^2_n is $(1 - 2n)^{-n/2}$.
9. If $n = 1$, the chi square distribution gives the positive half of normal distribution curve.
10. If $n = 2$, it gives exponential distribution with mean 2.
11. As $n \to \infty$, the χ^2_n tends to normal distribution with mean n and variance $2n$.

12. If $n > 30$, $\sqrt{2\chi^2}$ follows approximately normal distribution with mean $\sqrt{(2n-1)}$ and variance 1.

Uses of χ_n^2 Distribution

χ_n^2 distribution is used in testing of hypothesis as follows.

1. To test the significance of sample variance.
2. To test goodness of fit.
3. To test of independence of attributes.
4. To test the independence of estimates of population variance, correlation coefficient etc.
5. To find the distribution of sample variance.

6.8.2 Student's *t* Distribution (i.e., *t* Distribution)

A random sample x_i, $i = 1, 2, \cdots, n$ is drawn from the normal population with mean μ and variance σ^2. The student's t statistic is defined as

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

It follows student's t distribution with $(n - 1)$ degrees of freedom, where $\overline{x}$ is the sample men and s^2 is sample variance (unbiased estimator of population variance σ^2).

The random variable t is said to follow student's t distribution with $n - 1$ degrees of freedom if its probability density function is given by

$$f(t) = \frac{1}{\sqrt{\nu}\ \beta\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{n}{2} - 1}} ; -\infty \leq \infty$$

where $\nu = n - 1$ is the degrees of freedom.

6.8.3 Fisher's *t* Distribution

It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If $X \sim N(0, 1)$ and $Y \sim \chi_n^2$ such that X and Y are independent then Fisher's t is given by

$$t = \frac{X}{\sqrt{\frac{Y}{n}}}$$

It follows student's t-distribution with n degree of freedom.

Its probability density function is given by

$$f(t) = \frac{1}{\sqrt{\nu}\ \beta\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{n}{2} - 1}} ; -\infty \leq \infty$$

Where, $\nu = n$ is the degrees of freedom.

Properties

Some of the properties of the t distribution are as follows:

1. Odd ordered raw moments about origin is zero. i.e., $\mu_{2r+1} = 0$; $r = 0, 1, 2, \cdots$
2. Odd ordered central moments are zero. i.e., $\mu_{2r+1} = 0$; $r = 0, 1, 2, \cdots$
3. Mean is zero.
4. Even ordered moment is given by

$$\mu_{2r} = \mu'_{2r} = n^r \frac{(2r-1)(2r-3)\cdots 3.1}{(n-2)(n-4)\cdots(n-2r)}$$

5. Variance is given by

$$\mu_2 = \frac{n}{(n-2)}; n > 2$$

6. $\beta_1 = 0$ and $\beta_2 = \dfrac{3(n-2)}{(n-4)}$
7. Mode is zero.
8. As $n \to \infty$, t distribution tends to normal distribution.
9. As $n \to \infty$, t^2 distribution tends to F distribution with $(1, n)$ degrees of freedom.

Use of *t* Distribution

1. *t*-distribution is used to test the significance of sample mean when population variance is unknown.
2. *t*-distribution is used to test the significance of difference of sample mean when population variance is unknown.
3. Test the significance of correlation coefficients and regression coefficients.

Exercise 6.1

1. Describe types of sampling survey methods.

2. What do you mean by simple random sampling? Describe the method of selecting sampling with replacement and without replacement in the sampling.

3. What do you mean by random sampling? Describe difference by of random sampling.

4. Write short notes on (a) simple random sampling (b) Stratified sampling (c) Systematic sampling (d) Cluster Sampling (e) Multistage sampling.

5. What is *pps* sampling? Differentiate between *srs* and *pps* sampling.

6. When is *pps* sampling is more effective than simple random sampling?

7. Describe the method of selecting *pps* sampling with replacement?

8. Find the relation for unbiased estimate of population mean and its variance.

9. Find the relation for unbiased estimate of population total and its variance.

Exercise 6.2

Multiple Choice Questions circle (O) the correct answer.

1. A sample consists of

(a) all units of the population (b) 50% units of the population

(c) 5% of the population (d) any fraction of the population

2. Sampling is in evitable in the situations.

(a) blood test of a person (b) when the population is infinite

(c) testing of life of dry battery cells (d) all the above

3. In case of systematic sampling

(a) sample mean is biased estimator population mean

(b) sample mean is unbiased estimator population mean

(c) sample mean cannot estimate population mean

(d) sample mean may equal to population mean

4. Mean of χ^2-distribution with n degrees of freedom is

(a) 1 (b) 0 (c) $2n$ (d) n

5. Variance of χ^2-distribution with n degrees of freedom is

(a) 1 (b) 0 (c) $2n$ (d) n

6. Probability of selection varies at each subsequent draw in:

(a) sampling without replacement (b) sampling with re placement

(c) both (a) and (b) (d) neither (a) or (b)

7. An unordered sample of size n can occur in

(a) n ways (b) $n!$ ways (c) one way (d) n^2 ways

8. Probability of any one sample of size n being drawn out of N units is:

(a) $1/N$ (b) n/N (c) $1/n!$ (d) $1/\binom{N}{n}$

9. Probability off including a specified unit in a sample of size n selected out of N units is:

(a) $1/n$ (b) $1/N$ (c) n/N (d) $\frac{N}{n}$

10. A selection procedure of a sample having no involvement of probability is known as:

(a) purposive sampling (b) judgment sampling(c) subjective sampling (d) all the above

11. An estimate based on a fixed set of values of a sample always possess:

(a) a single value (b) any value (c) a value equal to one (d) all the above

12. Students-t is categorized as:

(a) an estimate (b) an estimator (c) a statistic (d) none of above

13. If each and every unit of a population has equal chance of being included in the sample, it is known as:

14. The most important factor in determining the size of a sample is:

(a) the availability of resources (b) purpose of the survey

(c) heterogeneity of population (d) none of the above

15. If n units are selected in a sample from N population units, the sampling fraction is given as:

(a) $\frac{N}{n}$ (b) $\frac{1}{N}$ (c) $\frac{1}{n}$ (d) $\frac{n}{N}$

Answer Key

1. (d)	2. (d)	3. (d)	4. (d)	5. (c)	6. (a)	7. (b)	8. (d)	9. (b)	10. (d)
11. (a)	12. (c)	13. (d)	14. (c)	15. (d)					