UNIT 3

SIMPLE RANDOM SAMPLING FOR ATTRIBUTES

Structure

3.1 INTRODUCTION

The first two units of this block have been exclusively devoted to the extensive and detailed study of Simple Random Sampling (SRS) scheme with an aim to acquaint you with the most basic and fundamental sampling scheme in sampling theory, which, in fact, prepares a firm base for other sampling schemes which are considered to be the advanced and refined versions of SRS. While Unit 1 dealt with the scheme known as “Simple Random Sampling with Replacement” (SRSWR), the Unit 2 dealt with “Simple Random Sampling without Replacement” (SRSWOR) scheme.

You might have noticed in these two units that the variable under study, denoted by Y was quantitative in nature, that is, we could measure the values of the variable in some scale and could apply various mathematical techniques which are used for analysing the numerical figures. Examples of such quantitative characteristics are age, height, weight, income, blood pressure of persons; number of children per family; number of rooms per household; profit and loss per month of retail shops in a mall, etc. In all these situations, characteristics under study are quantitative in nature and, hence, are popularly called *“Variables”.*

Sometimes, we come across with such characteristics, which are not quantitative in nature since they cannot be measured in some units of measurement and cannot be represented in terms of numerical figures. Examples of such characteristics are sex of persons, colour of flowers, honesty of persons, beauty of females, nationality of persons, opinion of experts on some matter, etc. In all these characteristics, value of measurement has no meaning, therefore, mathematical operations which are applicable to variables cannot be applied to such characteristics. These types of characteristics are very commonly known as **“Qualitative Characteristics”** or **“Attributes”**. What mathematical operations we can apply with such observed elements is only to classify the elements into a number of classes to which they belong. As for example, sex of persons can be classified into two classes: ‘male’ and ‘female’; colour of flowers may be classified as red, yellow, pink, blue, etc.; nationality of persons can be classified as Indian, Russian, Australian, German, American; etc.

As far as the sample surveys are concerned; in many surveys, according to the purpose and aim of the survey, the investigator might be interested to gather information on one or more characteristic(s), which might be of qualitative in nature instead of quantitative characteristics. In other words, units of the population might be interviewed/measured for gathering information on a single qualitative characteristic or a number of qualitative characteristics. In such situations, sample surveys which deal with samples selected from the given population are termed as **“Survey Sampling for Attributes”**, in which sampling is made to collect some units in the sample for gathering information on qualitative characteristics.

If in any study, attributes are the study characteristics, the question arises how the collected data can be combined together to obtain a representative sample value or statistic or estimator for drawing inference on some

population parameters and what types of analytical techniques may be used in case of qualitative characteristics for getting an idea about the sampling error of the estimator used.

You might have observed from the first two units, devoted to the study of Simple Random Sampling scheme, that the *'selection procedure'* in any sampling scheme is not affected by the analytical techniques to be adopted later in 'estimation procedure'. This means that as far as the selection of a sample from the given population is concerned, all types of sampling schemes can be used without bothering about the type of data to be collected on the population units, whether it is quantitative or qualitative. The type of data affects the estimation techniques later on, due to reason that quantitative and qualitative data involve different types of analysis techniques. In this light, we can discuss the application of the basic sampling scheme, that is, SRS, even when the attribute is the characteristic under study.

In this context, we shall discuss and describe the theory of Simple Random Sampling in the case when some qualitative type characteristics are observed during the process of sample selection and later on, at the time of exploring different types of estimators for population parameters.

In this unit, we shall discuss about the applications of Simple Random Sampling scheme in order to select some units in the sample and then for defining the estimators based on the values of qualitative characteristic under study along with their properties. In this sense, this unit may be looked upon as an extension of previous results obtained in the previous two units. In Section 3.2, we shall show how the results of Simple Random Sampling are applicable on attributes also. We shall show that many of the results derived on attributes are directly derivable from the previous results, obtained for variables. Since, in case of attributes, we classify the units into different categories, we are more interested into the proportions of units in each class. Accordingly, in Section 3.3, we shall discuss the procedure of selecting a random sample from the population when the characteristic under study is an attribute. Section 3.4 describes the type of parameters which are generally considered in case of attributes.

Section 3.5 devoted to obtaining an estimator of the population proportion of the attribute along with its important properties, assuming that the attribute can be classified only into two classes. Section 3.6 discusses the estimation of several population proportions assuming that the attribute under consideration can be classified into more than two classes. Section 3.7 has been devoted to the problems of determining the optimum sample sizes under Simple Random Sampling with Replacement, Simple Random Sampling without Replacement schemes with quantitative characteristic and Simple Random Sampling without Replacement schemes scheme when an attribute is the characteristic of interest.

Expected Learning Outcomes

After studying this unit, you should be able to:

- explain sampling for attributes and particularly Simple Random Sampling schemes with a qualitative characteristic (attribute);

- discuss the necessity of sampling with attributes arises in real world life;
- explore the examples of attributes and its levels in which an attribute can be classified;
- describe method of selecting a random sample using Simple Random Sampling schemes, if we are interested to gather information on a qualitative characteristic;
- determine the population parameters which mainly depend upon the qualitative type of characteristics;
- use Simple Random Sampling schemes to find their estimators based on a random sample in case there are (i) only two levels and (ii) more than two levels of qualitative characteristic;
- discuss the important properties of the estimator of population proportion when Simple Random Sampling schemes (both SRSWR and SRSWOR) are used for estimation purpose;
- obtain the estimate of sampling variance of the sample proportion estimator;
- explain the concept of optimum samples size and its importance in sampling scheme; and
- determine the value of the optimum sample size when the characteristic under study is quantitative type and qualitative type both.

3.2 SAMPLING FOR ATTRIBUTES

In this section, we shall first discuss about the classification of an attribute into different categories with the help of some real examples, followed by the necessity of using appropriate sampling schemes when the characteristic under study is a qualitative characteristic.

3.2.1 Examples of Attributes and Their Classification

Let we be given a population of human beings (persons), which are elementary units, and, therefore, be contacted for seeking necessary information. Naturally, a particular person possesses a number of characteristics, some of which are quantitative by nature, such as, age, height, weight, blood pressure level, daily working hours, etc., and some qualitative characteristics also, like, sex, nationality, marital status, eye colour, dietary habits, etc. In a sample survey, therefore, purposefully, the study characteristic might be a qualitative characteristic on which the required information be collected from the units selected in the sample. Contrast to the sampling for variables, in such surveys, the information could not be obtained in terms of numerical numbers, rather, we could simply know in which class of the attribute the person belongs.

Case 1: Estimation of Population Proportion

In the context of a population of human beings, let the information be sought whether the person belonging to the given population has some kind of

addiction or not and in case the answer is 'Yes'; we might be interested to know what type of addiction, whether addicted to alcohol, smoking or drugs. Thus, at the first place, the information has two levels: 'Yes' or 'No' and at the second place there are three levels of addiction: alcohol, smoking and drugs. No doubt, at both the places, information is qualitative by nature; having two levels of classification at the first place and three levels of classification at the second place. We, therefore, can state that what one can do in case of qualitative characteristics is simply to classify the data into a number of categories (sub-populations) and then we might be interested into the estimation of proportion of each of the sub-populations in the population.

Case 2: Estimation of Sex Ratio in the Population

Let the problem be to know what is the existing sex ratio in a large population of 5000 persons residing in a society?

Obviously, on the basis of a sample of appropriate size, it would be possible to estimate the existing sex ratio in the society. The selected persons could be asked about their sex, which is purely a qualitative characteristic of the persons. The answers could be classified broadly into two sub-populations: 'Male' and 'Female'. Counting the numbers of each sub-population, the ratio between the number of males and the number of females can easily be obtained as an estimate of the sex ratio in the population. Remember that the characteristic under this study is 'sex' which is qualitative in nature.

Case 3: Estimation of Population Proportion of Nationals

Let the capital city of a country shows the following composition of different nationals residing in the capital:

Nationality	Indian	Britisher	Russian	South Korean	American	Japanese	Total
Number	N_1	N_2	N_3	N_4	N_5	N_6	**N**

Since the characteristic "Nationality" in the population of size N is an attribute possessing 6 levels, the population is categorized (classified) into 6 sub-populations. Let the problem be to estimate the population proportion of Russian and American nationals in the capital, respectively, given by $\frac{N_3}{N} = P_3$ and $\frac{N_5}{N} = P_5$ say.

Obviously, for this purpose, a sample survey has to be conducted in the capital city in order to select a random sample of appropriate size. The sample proportions of these nationals would be then immediate estimators of population proportions.

Let the composition of the selected sample be as follows:

Nationality	Indian	Britisher	Russian	South Korean	American	Japanese	Total
Number	n_1	n_2	n_3	n_4	n_5	n_6	**n**

Clearly then, an estimate of population proportion P_3 and P_5, say, denoted respectively by $\hat{P}_3$ and $\hat{P}_5$ will be $\frac{n_3}{n}$ and $\frac{n_5}{n}$, obtained from the sample.

Since, the sample is selected randomly from the population, nobody can

predict how many cases of the i^{th} level will appear in the sample, except the knowledge that the cases coming from the i^{th} sub-population will be between 0 to n_i, both values inclusive. In this sense, all n_i's are random variables.

Let $\frac{n_3}{n} = p_3$ and $\frac{n_5}{n} = p_5$.

Note that whereas population proportions P_i (i = 1, 2, …, 6) are unknown, being population parameters; sample proportions p_i, for (i =1, 2, …, 6) would be known.

Sometimes, quantitative characteristics are also classified into different levels of measurement so as treat the data as a qualitative characteristic and analyse the data purposefully into a different manner. We present here an example for illustration.

Example 1: The ages of 650 persons residing in a small society be classified into 5 age groups. The age groups and corresponding number of persons belonging to that age group are as follows:

Age Group (in Years)	Number of Persons	Proportion of the Group
1 -9	78	0.120
10 – 25	145	0.224
26 – 49	220	0.338
50 – 74	117	0.180
75 and above	90	0.138
Total	**650**	**1.000**

Select a random sample of size 25 persons from the society and find estimates of the population proportions of children under age 10 and senior citizens with age 75 years and more, using the Simple Random Sampling without Replacement schemes scheme.

Solution: Even though the characteristic 'Age of Persons' is a variable (quantitative characteristic), the analysis of the data in the problem has to be made in a different way. No doubt, in the designed survey, the characteristic under study is of quantitative nature and, hence, the necessary information to be gathered from the units, by conducting a sample survey, will be the age of persons. After selecting the sample of desired size from the population, units of the sample are to be contacted for asking their ages. The ages then are to be classified in the five age groups along with the number of persons in each age group. The sample proportion of a particular age group will be the estimate of the population proportion of that age group.

This is an example which shows that even though the information is gathered on a quantitative characteristic, but the analysis techniques are those which applied in case of qualitative characteristic.

You know that in order to select a random sample, first of all, the units belonging to the population are to be labelled from 1 to 650, in any order, and, thus, the sampling frame is to be prepared. However, in this example, persons are classified in different age groups, so to identify persons

separately within age groups is not possible. However, we can assume that society office has such age-wise records of persons and, hence, this record can be used for assigning labels to persons. In order to assign labels to persons, therefore, we may assign labels 1 to 78 to children of age group 1 - 9 years; labels 79 to 223 to the persons belonging to 10 - 25 age group; labels 224 to 443 to persons of age group 26 - 49 years and so on.

Let us now select a random sample of size 25 using Simple Random Sampling without Replacement scheme with the help of random number tables. Starting from the fourth row and moving row-wise, we select random numbers. Considering only three left-most digits of each random number, we have the following sample consisting of labels of units:

{580, 569, 616, 500, 553, 238, 285, 629, 92, 129, 585, 216, 33, 472, 37, 201, 97, 648, 372, 619, 355, 108, 614, 545, 373}.

Now after drawing the sample, we now arrange the number of persons age-wise selected in the sample in tabular form as follows:

Age Group (in Years)	Number of Persons	Proportion of the Group
1–9	02	0.08
10–25	06	0.24
26–49	05	0.20
50–74	04	0.16
75 and above	08	0.32
	25	1.00

We observe that in the sample, 2 persons are selected from age group 1 - 9; 6 persons from age group 10 - 25; 5 persons from age group 26 - 49; 4 persons from age group 50 - 74 and 8 persons from age group 75 and more. So, the sample proportions of age groups 1 - 9 years and 75 and above are found respectively to be $\frac{2}{25}=0.08$ and $\frac{8}{25}=0.32$.

3.2.2 Necessity of Using Sampling Schemes for Attributes

In the previous two units it was made clear that conducting survey sampling is a must whenever the problem of estimation of population parameters on the basis of only a small part of the population, that is, sample, arises. Since some parameters such as proportions, ratios, rates are mostly associated with a qualitative characteristic and its classification into a number of sub-populations; the estimation of such parameters on the basis of a selected sample from the population necessitates the application of some appropriate sampling scheme so as to select a random sample of pre-fixed size and then to apply suitable analysis techniques for obtaining an efficient estimate from the sample used.

Keeping in view the nature of the population, characteristic under study, estimation techniques to be used and the efficiency of the estimates to be derived from the sample, the investigator is free to use any of the sampling schemes, available in the Theory of Sampling. However, since till now, we are familiar with the Simple Random Sampling scheme only, we shall see

how Simple Random Sampling scheme can be used in case the sample survey considers the characteristic under study which is of qualitative nature.

3.3 SIMPLE RANDOM SAMPLING FOR ATTRIBUTES

Let a sample survey involves SRS scheme for (i) selecting a random sample of pre-fixed size, (ii) obtaining required information on the sampled units selected and (iii) estimating population parameters based on the information gathered on units selected in the sample. From the discussions and descriptions made in the Sub-section 1.6.3 in Unit 1 and in the Section 2.2 in Unit 2; you might have learnt the process of selecting a random sample of fixed size, respectively under SRSWR and SRSWOR schemes. In this section we shall be discussing selection of simple random sample for attributes.

3.3.1 Selecting Simple Random Sample with Attribute as Study Characteristic

You might have also noted that as far as the process of sampling the units in the sample in both the schemes are concerned, it is only the labels of units which are selected during the selection procedure and the nature of the characteristic under study does not play any role in the process of selection. This implies that whether it is sampling for variable (or, quantitative characteristic) or sampling for attribute (or, qualitative characteristic); the selection procedure remains same and the procedure does not have any impact of the nature of the study characteristic.

In the following figure, a population of size N consisting of M units with attribute A and (N – M) units not possessing. The attribute has been sampled to get a sample of size n SRS scheme. The sample can be partitioned into two classes: x units possessing the attribute A and (n – x) units not possessing the attribute. Obviously, here x would be a random variable.

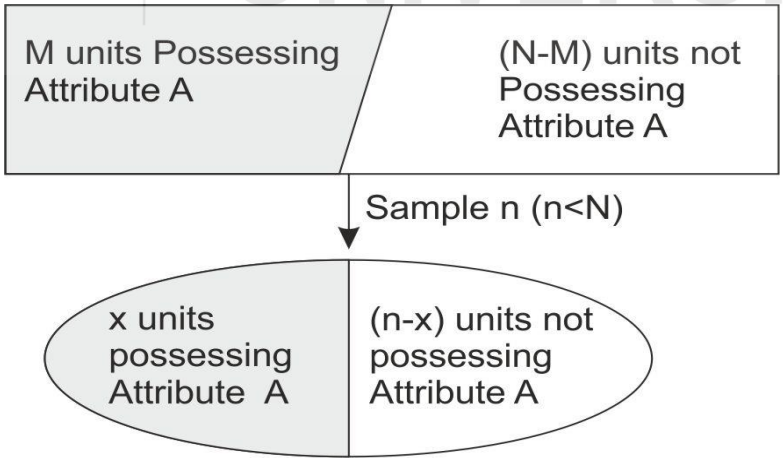

Fig. 3.1 Simple Random Sampling Scheme for Attributes

The conclusion drawn in the above paragraph indicates that the selection procedure of a random sample from the population, following SRSWR scheme in the case of attributes, would be exactly same as described in sub-section 1.6.3 of the Unit1. Similarly, selection procedure of a random sample,

following SRSWOR scheme for attributes, would be exactly same as described in Section 2.2 of Unit 2.

Example 2: Let us consider the case 3 discussed in the Sub-section 3.2.1 providing the nationality of nationals residing in the capital city of a country. Let the number of different nationals is as given in the following table:

Nationality	Indian	Britisher	Russian	South Korean	American	Japanese	Total
Number	45	55	18	15	200	27	**360**

With the aim to estimate the proportions of Russian and American residents in the population, a random sample of size 10% of the population size is to be selected from the population using Simple Random Sampling without Replacement scheme. Draw the sample using Random Number Table given in the **Appendix - A** of Unit 1.

Solution: The size of the sample is provided as 36. Let the nationals be assigned labels from 1 to 360 in the same order in which they appear in the above table. Then, obviously labels 1 to 45 will be assigned to 45 Indians in whatever order they are arranged within the group; labels 46 to 100 to be assigned to Britishers arranged in any manner within the group; labels 101 to 118 to be assigned to Russians and so on. Since, the largest label is a three digital number, we choose the last three digits of the random numbers in the table. Let us start with the second column and move forward column-wise. Then since, it is SRSWOR scheme, we skip the labels once selected at any draw. Going through the random numbers one by one, we see that the following labels are selected for the sample:

{154, 271, 341, 112, 177, 280, 257, 143, 73, 93, 298, 264, 248, 101, 176, 179, 150, 44, 277, 35, 188, 245, 169, 45, 162, 191, 304, 267, 276, 243, 222, 212, 258, 282, 303, 81}.

Now, we classified the sampled persons as per the Nationality of the persons out of 36:

Nationality	Number	Proportion of the Group
Indian	03	0.08
Britisher	03	0.08
Russian	02	0.06
South Korean	00	0.00
American	27	0.75
Japanese	01	0.03
	36	1.00

Remark 3.1: While going through the random numbers starting from the second column and reaching to the first number in the fifth column in Random Number Table in order to complete the selection of 36 labels in the sample, we observed that the label 150 appeared at two different places, one, in the third column and the second, in the fourth column. Since, this label was selected at the seventeenth draw, we skipped the same label when it appeared in the fourth column since we required to use Simple Random Sampling without Replacement scheme.

Remark 3.2: The estimated value of the population proportions of Russians and Americans can be easily obtained from the labels appeared in the sample. Observing the labels of nationals as assigned to them, we see that 3 Indians, 3 Britishers, 2 Russians, zero South Koreans, 27 Americans and 1 Japanese are selected in the sample. Therefore, the estimated proportion of Russians and Americans are respectively $\frac{2}{36}=0.06$ and $\frac{27}{36}=0.75$, whereas the respective population proportion of these nationals are found to be 0.05 and 0.56.

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

SAQ 1

What do you understand by sampling for attribute? In what sense it is different from sampling for variable?

3.4 POPULATION PARAMETERS BASED ON QUALITATIVE CHARACTERISTICS

As we discussed earlier, instead of getting numerical values on some quantitative characteristics, in case of qualitative characteristics, what actually we have, are the total number of levels (classes) of the characteristic under study in which the recorded facts gathered on the units of the population are classified along with the number of cases belonging to them. The examples presented in the Section 3.2 elaborate this fact regarding the attributes.

Generally, in most of the situations, there happen to be only two sub-populations of the considered attribute, one, which consists of units possessing the attribute or in which the attribute of interest is present, and the other, in which the attribute is absent. Whenever, the question of the estimation of 'population proportion' of any of the two classes arises, the only way is (i) to conduct a sample survey with the attribute in question and (ii) to select a random sample of units from the population so as to apply suitable estimation technique on the data obtained from the sampled units. Sometimes, instead of only two classes of the attribute, we come across with the attribute which have more than two levels (classes), in which the facts related to it, can be classified. In such situations, the problem remains same, that is, estimation of population proportion of one or more of the classes.

Some more situations may also arise with the attributes as study characteristic, in which we might be interested to estimate the 'Ratio of Two Characteristics', both of them varying from unit to unit. Estimation of "Sex Ratio" in a population is a good example of it. Clearly, the estimation of it is based upon the data of qualitative nature.

Another type of population parameter related to qualitative type of data is the "Rates of some events". The estimation of some vital events occurring in a human population is an example of such problems. For instance, estimation of 'Crude Death Rate', 'Crude Birth Rate', 'Infant Mortality Rate', 'Rate of

Change of Population Size' and other similar parameters require qualitative type of data.

The discussions made in this section clearly explains the fact that almost all the sampling schemes which are existing in the literature of the Sampling Theory, can be used for estimating population parameters whether these parameters are based either on quantitative characteristics or qualitative characteristics. As far as the 'selection part' of the sample survey, that is, the process of selecting a random sample, is concerned, there is no basic difference whether the characteristic under study is of quantitative nature or of qualitative nature. However, the basic difference appears at the 'stage of estimation' because of the reason that both types of surveys need its own mathematical methods and method of analysis.

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

SAQ 2

What are the parameters of a population which are computed on the basis of attributes? Mention some of them.

3.5 ESTIMATING POPULATION PROPORTIONS FOR ATTRIBUTES WITH ONLY TWO LEVELS

This section is devoted to the problem of estimation of the parameter "Population Proportion" on the basis of the qualitative data obtained under Simple Random Sampling without Replacement scheme.

To start with dealing the problem of estimation of population proportion using the data collected in a sample survey, for simplicity, we assume that the concerned attribute possesses only two classes in which it can be classified. Examples of such characteristics are many, for example, 'Sex of Persons' having only two broad sub-populations: Males and Females; 'Dietary Habit' with only two main sub-populations: Vegetarian and Non-vegetarian; 'Marital Status' with two classes: Married and Unmarried; 'Employment Status' with sub-populations: Employed and Unemployed; 'Smoking Habit' with classes: Smoker and Non-smoker and others.

3.5.1 Notations Used

Let us denote the attribute under consideration by A. Let the presence of the attribute be also denoted by A and its absence by $\bar{A}$. For instance, if the person is male, we say that attribute A is present and if the person is a female, then we may say that the attribute A is absent, and, hence, we use the notation $\bar{A}$ denote a female.

Let in the population of size N; N_1 be the number of A's and N_2, the number of $\bar{A}$, so that we have $N_1+N_2 = N$. Our aim is to estimate the proportion of A's in the population, that is, to estimate $N_1/N = P$. Obviously, P is a population parameter, which in general might not be known, so it would be estimated using the selected random sample.

Let $Q = 1 - P = \frac{N_2}{N}$. Then, we have P + Q = 1.

We see that $NP = N_1$ and $NQ = N_2$, so that NP + NQ = N.

In order to get an estimated value of the population proportion P or Q, obviously, a random sample of specific size has to be selected from the population using Simple Random Sampling without Replacement scheme.

Let the sample size be n. Further, we assume that the number of units in the sample possessing and not possessing the attribute A are respectively, n_1 and n_2 such that $n_1 + n_2 = n$.

Let us denote the respective sample proportions by p and q so that

$$p + q = \frac{n_1}{n} + \frac{n_2}{n} = 1.$$

Also, we see that $n_1 = np$ and $n_2 = nq$.

3.5.2 Estimation of Population Proportion

It is clear that if a random sample of size n is taken from the population, there would be some units possessing the attribute A and other units possessing the attribute $\bar{A}$ but their numbers would be random variables, because all the units are selected with the help of some chance mechanism, therefore, n_1 and n_2 both lie within the range $0 \le n_i \le n$ for i = 1, 2 with some specific selection probabilities. Since the sample proportions p and q are easily obtained from the sample itself, they can be taken as the direct estimators of respective population proportions P and Q.

Theorem 1: The probability of selection of n_1 units; $P(n_1)$, possessing the attribute A and n_2 units possessing the attribute $\bar{A}$ in the sample of size n follows a Hypergeometric Distribution.

Proof: We know that the number of units in the population, possessing the attribute A, is $N_1 = NP$ and the number of units possessing the attribute $\bar{A}$ is $N_2 = NQ$.

If a random sample of size n is selected from the population using Simple Random Sampling without Replacement scheme, then clearly each and every unit selected will have an equal chance of selection in any of the N independent draws. Clearly then, the total number of possible ways of selection of n_1 units out of N_1 units and n_2 units out of N_2 units will be:

$$\left({}^{N_1}C_{n_1}\right) \times \left({}^{N_2}C_{n_2}\right)$$

Similarly, the number of possible cases of selecting n units out of N units of the population will be ${}^{N}C_{n}$.

Therefore, the probability of selecting n_1 units from the class A (or equivalently, selecting n_2 units from the class $\bar{A}$) will be given by

$$p(n_1) = \frac{\left({}^{N_1}C_{n_1}\right)\left({}^{N_2}C_{n_2}\right)}{\left({}^{N}C_{n}\right)} \qquad n_1 = 0, 1, 2\ldots, N_1.$$

$$= \frac{\left({}^{NP}C_{n_1}\right)\left({}^{NQ}C_{n_2}\right)}{\left({}^{N}C_{n}\right)} \qquad \ldots (3.1)$$

Expression (3.1), in fact, provides a well-known probability distribution, namely, Hypergeometric Distribution, when the probability of selecting a random sample of size n from the population of size N, having two distinct classes, with sizes N_1 and N_2, is drawn in such a way that n_1 and n_2 units appear, respectively, out of N_1 and N_2. Since (3.1) is a probability distribution of the variable n_1, the following two results can easily be deduced:

$P(n_1) > 0$; for $n_1 = 0, 1, 2\ldots, N_1$; and $\sum_{n_1}^{N_1} P(n_1) = 1$.

As described above, a direct estimator of the population proportion, P is considered to be the corresponding sample proportion, $p = \frac{n_1}{n}$.

3.5.3 Properties of the Sample Proportion

(a) Expectation of Sample Proportion

As usual, under Estimation Theory of Statistics, we generally observe whether an estimator of a population parameter is unbiased or biased since the property of unbiasedness is always a desirable property of all the estimators. The unbiasedness of the estimator says that the average (mean) of the estimator, computed taking into consideration all the possible samples of same size, coincides with the value of the concerned parameter.

In this context, we have the following Theorem:

Theorem 2: The sample proportion p is an unbiased estimator of population proportion P.

Proof: Using the result $E(n_1) = \sum_{n_1} n_1 P(n_1)$ and expression (3.1), we have

$$E(n_1) = \sum_{n_1=0}^{N_1} n_1 \frac{\left({}^{N_1}C_{n_1}\right)\left({}^{N_2}C_{n_2}\right)}{\left({}^{N}C_{n}\right)} = \sum_{n_1=0}^{N_1} n_1 \frac{\left({}^{NP}C_{n_1}\right)\left({}^{NQ}C_{n_2}\right)}{\left({}^{N}C_{n}\right)}$$

$$= \sum_{n_1=0}^{N_1} n_1 \left\{\frac{NP!}{n_1!(NP-n_1)!}\right\}\left\{\frac{NQ!}{n_2(NQ-n_2)!}\right\}\left\{\frac{n!(N-n)!}{N!}\right\} \qquad \ldots (3.2)$$

Now, cancelling n_1 outside the first bracket with $n_1!$ inside the bracket and taking NP and $\frac{n}{N}$ common from the first and third brackets respectively, we have,

$$E(n_1) = \frac{NnP}{N} \sum_{n_1=1}^{N_1} \left\{\frac{(NP-1)!}{(n_1-1)!(NP-n_1)!}\right\}\left\{\frac{NQ!}{n_2!(NQ-n_2)!}\right\}\left\{\frac{(n-1)!(N-n)!}{(N-1)!}\right\}$$

$$= nP \sum_{n_1=1}^{NP} \frac{\left({}^{NP-1}C_{n_1-1}\right)\left({}^{NQ}C_{n_2}\right)}{\left({}^{N-1}C_{n-1}\right)} \;;$$

Here, $\sum_{n_1} \frac{\left({}^{NP-1}C_{n_1-1}\right)\left({}^{NQ}C_{n_2}\right)}{\left({}^{N-1}C_{n-1}\right)} = 1,$

since it is the sum of the probabilities that in a sample of size (n – 1), selected from a population of size (N – 1); (n_1-1) units possessing attribute A and n_2 units possessing attribute $\bar{A}$.

Therefore, we have

$$E(n_1) = nP \Rightarrow E\left(\frac{n_1}{n}\right) = P$$

$$\Rightarrow E(p) = P, \qquad \text{since } \frac{n_1}{n} = p$$

which shows that sample proportion, p, is an unbiased estimator of population proportion P. This establishes the theorem.

This Theorem states that the sample proportion p possesses the desirable property of unbiasedness.

Similarly, considering n_2 in place of n_1, we can prove that

$$E(n_2) = nQ \Rightarrow E(q) = Q.$$

We can express these facts as

Est. (P) = p and Est. (Q) = q.

(b) Sampling Variance of the Sample Proportion

The estimator, p, being the sample proportion, varies from sample to sample and, hence, is a random variable. This means that it has some variability similar to a random variable. It is, therefore, necessary to find the measure of its variability over sample to sample, which is obtained in the following theorem:

Theorem 3: The sampling variance of the estimator p is given by:

$$V(p) = \frac{N-n}{N-1}\frac{PQ}{n}$$

Proof: As per definition of variance, we have

$$V(p) = V\left(\frac{n_1}{n}\right) = \frac{1}{n^2}V(n_1) = \frac{1}{n^2}\left[E\{n_1 - E(n_1)\}^2\right]$$

But

$$E\{n_1 - E(n_1)\}^2 = E[n_1^2 + \{E(n_1)\}^2 - 2n_1E(n_1)]$$

$$= E\left[n_1^2\right] + E\left\{E(n_1)\right\}^2 - 2E(n_1)E(n_1)$$

$$= E\left[n_1^2\right] + \left\{E(n_1)\right\}^2 - 2\left\{E(n_1)\right\}^2$$

$$= E\left(n_1^2\right) - \left\{E(n_1)\right\}^2 \qquad \ldots (3.3)$$

We know that $E(n_1) = nP$,

therefore, $\{E(n_1)\}^2 = n^2P^2$, a constant value.

Therefore, $E[\{E(n_1)\}^2] = \{E(n_1)\}^2$

Also, we see that

$$E\left[2n_1E(n_1)\right] = 2E(n_1)E(n_1) = 2\{E(n_1)\}^2$$

Thus, (3.3) reduces to

$$E\{n_1 - E(n_1)\}^2 = E\left[n_1^2\right] - \{E(n_1)\}^2 = E\left[n_1^2\right] - n^2P^2. \quad \ldots (3.4)$$

We shall evaluate the value of $E\left[n_1^2\right]$ given in (3.4).

By the definition of expectation, we have

$$E\left[n_1^2\right] = \sum_{n_1} n_1^2 P(n_1)$$

$$= \sum_{n_1} \{n_1(n_1-1) + n_1\} P(n_1)$$

$$= \sum_{n_1} n_1(n_1-1)P(n_1) + \sum_{n_1} n_1 P(n_1)$$

$$= \sum_{n_1} n_1(n_1-1)P(n_1) + nP \quad \ldots (3.5)$$

Let us evaluate the term $\sum_{n_1} n_1(n_1-1)P(n_1)$ given in (3.5).

By definition, we have

$$\sum_{n_1} n_1(n_1-1)P(n_1) = \sum_{n_1=0}^{N_1} n_1(n_1-1)\frac{\binom{NP}{n_1}\binom{NQ}{n_2}}{\binom{N}{n}}$$

$$= \sum_{n_1=0}^{N_1} n_1(n_1-1)\left\{\frac{NP!}{n_1!(NP-n_1)}\right\}\left\{\frac{NQ!}{n_2!(NQ-n_2)!}\right\}\left\{\frac{n!(N-n)}{N!}\right\}$$

In the above expression, we cancel $n_1(n_1-1)$ with $n_1!$ (factorial n_1) side the first bracket and then taking common NP(NP–1) from (NP!) (Factorial NP) of the first bracket and $\left\{\frac{n(n-1)}{N(N-1)}\right\}$ from the factorials (n!) and (N!) of the third bracket, we see that,

$$\sum_{n_1} n_1(n_1-1)P(n_1) = \frac{NP(NP-1)n(n-1)}{N(N-1)}$$

$$\sum_{n_1=0}^{N_1}\left\{\frac{(NP-2)!}{(n_1-2)!(NP-n_1)}\right\}\left\{\frac{NQ!}{n_2!(N-n_2)!}\right\}\left\{\frac{(n-2)!(N-n)!}{(N-2)!}\right\}$$

$$= \frac{NP(NP-1)n(n-1)}{N(N-1)} \sum_{n_1=2}^{N_2} \frac{\binom{NP-2}{n_1-2}\binom{NQ}{n_2}}{\binom{N-2}{n-2}}$$

$$= \frac{P(NP-1)n(n-1)}{(N-1)} \qquad \text{since } \sum_{n_1=2}^{N_2} \frac{\binom{NP-2}{n_1-2}\binom{NQ}{n_2}}{\binom{N-2}{n-2}} = 1;$$

because is the sum of the probabilities that in a sample of size (n – 2), selected from a population of size (N – 2), (n_1 – 2) units are possessing attribute A and n_2 units possessing attribute $\bar{A}$.

Substituting the value of $\sum_{n_1} n_1(n_1-1)P(n_1)$ in (3.5), we have

$$E\left[n_1^2\right] = \frac{P(NP-1)n(n-1)}{(N-1)} + nP$$

Therefore, now substituting the value of $E(n_1^2)$ in (3.4), we get

$$V(p) = \frac{1}{n^2}\frac{P(NP-1)n(n-1)}{(N-1)} + \frac{nP}{n^2} - P^2$$

$$= \frac{(NP^2-P)(n-1)+P(N-1)-n(N-1)P^2}{n(N-1)}$$

$$= \frac{-nP-NP^2+NP+nP^2}{n(N-1)}$$

$$= \frac{-nP(1-P)+NP(1-P)}{n(N-1)}$$

$$= \frac{-nPQ+NPQ}{n(N-1)} = \frac{N-n}{(N-1)}\frac{PQ}{n}.$$

This completes the proof of the theorem.

Remark 3.3: The Standard Error (S.E.) of the estimator p is given by

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

Remark 3.4: In Units 1 and 2 of this Block, we derived a number of results under SRSWR and SRSWOR scheme, with a quantitative characteristic. The present unit also discussed and derived some equivalent results under the same sampling scheme, but with a qualitative characteristic. It is, therefore, interesting to see that whether the results obtained using qualitative characteristic are exactly similar to that obtained with a quantitative characteristic. In this context, a comparison of the derived results has been presented in the next sub-section.

3.5.4 Simple Random Sampling without Replacement with Attributes as a Special Case of Simple Random Sampling without Replacement with Variables

From the notations used in the Sub-section 1.6.4 of Unit 1, we know that if Y is a quantitative characteristic, then population and sample means are respectively, given by

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Further, population and sample mean squares are respectively given by

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

and

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}\left(y_i - \bar{y}\right)^2$$

Let us assume that variable Y assumes only two values, namely, 1 if it possesses the attribute A and 0 if it does not possess the attribute A; that is,

Y_i = 1, if the i^{th} unit possesses the attribute A

= 0, if the i^{th} unit possesses the attribute $\bar{A}$

If Y_i values assume value 1 for those units of the population which possess the attribute A and rest units assume value zero which do not possess the attribute A; then the population mean would be given by

$$\bar{Y} = \frac{1}{N}\left[Y_1 + Y_2 + \ldots + Y_N\right] = \frac{1}{N}\left[\left(1+1+1+1+\ldots 1\right) + \left(0+0+\ldots+0\right)\right];$$

where there would be N_1 one's and N_2 zero's above inside the bracket.

Therefore, for such a variable, we have

$$\bar{Y} = \frac{N_1}{N} = P$$

which is the population proportion if Y is a qualitative characteristic.

Thus, we see that the population proportion, P of a qualitative characteristic is equivalent to the population mean $\bar{Y}$ of a quantitative characteristic.

Similarly, the sample mean of a quantitative characteristic Y is

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i = \frac{1}{n}\left[y_1 + y_2 + \ldots + y_n\right] = \frac{1}{n}\left[\underbrace{1+1+0+1+0+0+1\ldots+1}_{\text{n times}}\right];.$$

the above bracket there will be some one's and rest zero's. Obviously, there will be n_1 y-values with 1 and n_2 y-values with zero. Therefore, we have

$$\bar{y} = \frac{n_1}{n} = p$$

the sample proportion.

Thus, we see that sample mean $\bar{y}$ of a quantitative variable is same as the sample proportion of a qualitative characteristic.

Similarly, we observe that population mean square S^2 is:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \frac{\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2}{N-1}; \quad \text{for quantitative characteristic}$$

$$= \frac{NP - NP^2}{N-1} = \frac{N}{N-1}P(1-P); \qquad \ldots (3.6)$$

for qualitative characteristic.

The sample mean square s^2

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(y_i^2) - \frac{n}{n-1}\left(\bar{y}^2\right); \quad \text{For quantitative characteristic}$$

$$= \frac{n_1 - \frac{n_1^2}{n}}{n-1} = \frac{n}{n-1}p(1-p) \qquad \ldots (3.7)$$

for qualitative characteristic

So, by virtue of the result $E(s^2) = S^2$ of the previous unit, we can write

$$E\left\{\frac{n}{n-1}p(1-p)\right\} = \frac{N}{N-1}PQ$$

$$\Rightarrow E\left\{p(1-p)\right\} = \frac{N}{N-1}\frac{n-1}{n}PQ$$

$$\Rightarrow \text{Est.}(PQ) = \frac{n}{n-1}\frac{N-1}{N}p(1-p) \qquad \ldots (3.8)$$

Now, since from the previous unit, we have

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

therefore, remembering that $\bar{y} = p$ and $S^2 = \frac{N}{N-1}P(1-P)$, we get

$$V(p) = \frac{N-n}{n(N-1)}PQ$$

the same result as obtained in **Theorem 3** of this unit.

3.5.5 Estimation of Sampling Variance of Sample Proportion

The sampling variance of the estimator p is seen to be a function of population proportions P and Q, which are mostly unknown in any sample survey. It is, therefore, not possible to know the variability of the estimator used, which must be known to the investigator for the reason that he/she may be able to rectify the estimation technique used or even change the sampling scheme too, if it is feasible in the survey; in search of obtaining more accurate estimated values in the sense that the variability of the

modified estimator is reduced to the extent possible. In this context, however, one thing which is possible is to replace the unknown parameters involved in the variance expression by their estimates as obtained from the sample values. We shall see here, if parameters are replaced by their estimated values, what would be best way to make this change. We know that,

$$V(p) = \frac{N-n}{n(N-1)}PQ$$

Therefore, in order to estimate $V(p)$, we have to find the estimate of PQ, that is,

$$\text{Est.}\,V(p) = \text{Est.}\left\{\frac{N-n}{n(N-1)}PQ\right\} = \frac{N-n}{n(N-1)}\text{Est.}(PQ) \qquad \ldots (3.9)$$

The expression (3.8) provides the result,

$$E\{p(1-p)\} = \frac{N}{N-1}\frac{n-1}{n}PQ$$

$$\Rightarrow E\left\{\frac{n(N-1)}{N(n-1)}p(1-p)\right\} = PQ$$

that is, an unbiased estimator of PQ is:

$$\frac{n(N-1)}{N(n-1)}p(1-p)$$

Therefore, one can use this unbiased estimator in place of Est.(PQ) in (3.9) in order to get an unbiased estimator of V(p).

Replacing Est. (PQ) in (3.9) by its unbiased estimator, we have

$$\text{Est.}\{V(p)\} = \hat{V}(p) = \frac{N-n}{n(N-1)}\frac{n(N-1)}{N(n-1)}p(1-p)$$

or $$\hat{V}(p) = \frac{(N-n)}{N(n-1)}p(1-p) \qquad \ldots (3.10)$$

3.6 ESTIMATING POPULATION PROPORTIONS FOR ATTRIBUTES WITH MORE THAN TWO LEVELS

This section is devoted to the problem of estimation of the parameter "Population Proportion" on the basis of the qualitative data obtained under Simple Random Sampling without Replacement scheme for attributes with more than two levels.

The case of only two classes of the concerned attribute can be extended to more than two classes of the attribute.

Let us now assume that we have such attributes under study which can be classified into more than two sub-populations. Example given under Case 3 presented in the Sub-section 3.2.1 possesses, six levels of classification. Similarly, Example 1 presented under Sub-section 3.2.1 possesses 5 levels

of classification. Some more examples may also be given. Let there be a music contest in which experts are supposed to give their opinion on the performance of the participants. The performances are decided to be categorized into four levels: 'Excellent', 'Good', 'Average 'and 'Not up to the mark'. Opinions of the experts is a qualitative characteristic which has four levels of classification in this example.

3.6.1 Notations Used

Let there be k sub-populations $(k > 2)$ classification. Let N_i be the size of the i^{th} sub-population (i = 1, 2, …, k), such that $\sum_{i=1}^{k} N_i = N,$ the size of the population. Let a random sample of size n be selected from the population in order to estimate the population proportions of the k levels of the attribute. Let in the sample, n_i $(i = 1, 2, \ldots, k)$ units possess the i^{th} level of the attribute.

Then, we have $\sum_{i=1}^{k} n_i = n.$

Let us denote the Population Proportion of the i^{th} level of the attribute by P_i.

Hence, we have $P_i = \frac{N_i}{N}; i = 1, 2, \ldots, k.$

We can see that $p_i = \frac{n_i}{n}$ represents the proportion of the units possessing the i^{th} level of the attribute in the sample.

3.6.2 Estimation of Population Proportions

We can now discuss the problem of estimation of population proportions P_i's on the basis of a random sample drawn from the population concerned using Simple Random Sampling without Replacement scheme.

For the estimation of the proportion of the i^{th} sub-population, let us consider the population consisting of only two levels of the attribute; one, the i^{th} sub-population and second, the sub-population which consists of the rest (k–1) sub-populations except the i^{th} sub-population. Thus, we convert the population having k sub-populations of the attribute under consideration into only two sub-populations: one, the i^{th} sub-population and the other, including rest (k–1) sub-populations. This would help us in applying the techniques of analysis which was used in case of two levels of the attribute under Section 3.5.

Now, we have two levels of attribute with the population sizes N_i and $(N – N_i)$. A random sample of size n was selected from the population containing two mutually exclusive classes out of which n_i units appeared with the i^{th} level of attribute and $(n – n_i)$ units appeared with rest of the levels of the attribute. With the analogy of Case - I, therefore, the probability of occurrence of n_i units from the i^{th} class and $(n – n_i)$ units from the rest (k – 1) classes together will be

$$P(n_i) = \frac{\binom{N_i}{n_i}\binom{N-N_i}{n-n_i}}{\binom{N}{n}}; \quad i = 1, 2, \ldots, k. \qquad \ldots (3.11)$$

The results derived under the Case of two levels only are now applicable to the present case where there are k classes of the attribute. The results which we derived in two level case only are equally true in this case also. We list those results in terms of the following theorems below in the context of more than two levels:

(i) **Theorem 4:** The sample proportion p_i is an unbiased estimator of population proportion P_i; $i = 1, 2, \ldots, k$.

$$E(n_i) = nP_i \Rightarrow E\left(\frac{n_i}{n}\right) = P_i \Rightarrow E(p_i) = P_i$$

which is the general result for more than two levels.

(ii) **Theorem 5:** The sampling variance of the estimator p_i is given by

$$V(p_i) = \frac{N-n}{N-1}\frac{P_iQ_i}{n}, \qquad \text{where } Q_i = 1 - P_i.$$

which is the result similar to **Theorem 3** derived for two levels.

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

SAQ 3

Show that the proportion of an attribute in a Simple Random Sampling without Replacement sample is an unbiased estimator of population proportion.

3.7 DETERMINATION OF SAMPLE SIZE

In this section we shall be discussing the determination of the optimal size of the sample from the population using simple random sampling without replacement scheme.

3.7.1 Understanding the Problem

You are familiar with the fact that every sample survey has some limitations such as,

(i) the time allowed for completing all the stages of the survey, like, survey designing stage, sample selection stage, estimation stage, report writing stage, etc.,

(ii) the manpower allowed to complete all types of works of the survey, like, designing the survey, preparing the sampling frame, selection of units of the population for constituting a sample of pre-fixed size, visiting/contacting the selected units for gathering necessary information on them, recording and coding of the gathered information

in the appropriate forms, application of mathematical/statistical techniques for the purpose of data analysis, preparing the full survey report mentioning the entire process of completing several stages of the survey, and

(iii) the total cost sanctioned for conducting the survey starting from the initial stage to the last stage.

No doubt, the most important component of a sample survey is the total cost sanctioned, since, other two components might be adjusted to some extent according to the cost of the survey.

In all types of sample surveys, mostly it is observed that the largest part of the total cost is spent on the process of selection of units and then visiting/contacting them personally with the aim to record the necessary information with utmost care and accuracy so as to get the best possible results based on them. Since the selected units, being selected randomly from the entire population, are generally scattered over a large area; the transportation cost, stationary cost and time go on increasing tremendously with increasing sample size. If anyhow, sample size is restricted to a minimum for reducing the cost, the investigator might not be getting desired accuracy of results with practically too small sample, due to increased sampling variance of the estimator. On the other hand, if he/she thinks that larger samples would produce most accurate results, he/she is not correct because then due to inclusion of a large number of units in the sample, many serious types of non-sampling errors creep up which are more dangerous than sampling error and which cannot be avoided.

We are familiar with the fact that it is the sampling variance of the estimator which provides the idea of average difference between the estimated values and the actual value of the concerned parameter, which is called **'random sampling error'**. The theory of estimation in Statistics states that among a number of estimators of the same class, the estimator having the least sampling variance is called an efficient estimator. Intuitively, it is also a fact that the larger the sample, the more accurate the results obtained. We also know that the sampling variance is a function of sample size and, therefore, it is a matter of concern, what is the roll of sample size in reducing the magnitude of sampling variance of the estimator to a minimum under the prevailing constraints, like, sanctioned cost, sanctioned manpower and time, etc. Thus, the problem of determination of sample size in sample surveys is actually the problem of finding the optimum size of the sample under certain conditions.

In sample surveys, therefore, estimations of appropriate sample size (optimum sample size) are used by researchers to determine how many units are needed in the sample with pre-defined constraints. The optimum sample size not only reduces the cost incurred in the survey and time required to finish it but also assures best prediction values of the population characteristics.

In the following texts, we shall show how the problem of obtaining optimal sample size can be tackled. We shall first discuss this for Simple Random

Sampling without Replacement scheme with quantitative characteristics and then for the same scheme with a qualitative characteristic.

3.7.2 Optimal Sample Size in Simple Random Sampling Scheme with Variables

When a quantitative characteristic is dealt with in Simple Random Sampling without Replacement scheme, the sampling variance of the sample mean estimator $\bar{y}$ are

$$V(\bar{y}) = \frac{N-1}{Nn}S^2 \text{ under SRSWR scheme.}$$

$$= \frac{N-n}{Nn}S^2 \text{ under SRSWOR scheme.}$$

Moreover, we are aware that Simple Random Sampling without Replacement scheme is always better than Simple Random Sampling with Replacement scheme in terms of efficiency of the estimator, it is not reasonable to consider the SRSWR scheme as far as the determination of sample size is concerned.

In both cases, sampling variance is a function of sample size, n; population size, N and population mean square, S^2. Being population parameters, N and S^2 are not under our control, but as the sample size is purely a decision of the investigator and, hence, is pre-fixed by him/her, to increase or decrease the size of the sample is in his/her hand. Further, from these expressions, it is evident that the sampling variance of the estimator reduces with the increase of the sample size and ultimately becomes zero for n = N, but this situation is not feasible since, complete enumeration tremendously increases the non-sampling errors which are more dangerous. In this sense, the survey statistician can think for selecting a sample as large as possible keeping in mind the budget constraint.

Because of the presence of constraints present: cost of the survey, limited time to complete the work and manpower to be applied; the investigator has to fix some goals in determining the optimum sample size. These are:

(a) Margin of Error (Confidence Interval) Permissible in the Estimate:

We know that no sample is perfect for estimating the parameter with 100% perfectness. Therefore, we have to fix in advance the error which we can allow. The confidence interval determines how much higher or lower than the population mean we are willing to let our sample mean fall.

(b) Assumption of the Probability Distribution of the Estimator:

We also need to make an assumption about the appropriate probability distribution of the estimator we are using. Generally, using the Central Limit Theorem, we may assume that if the sample size is not too small, the estimator approximately follows a normal distribution.

(c) Confidence Coefficient:

How confident do we want to be that the actual mean falls within our confidence interval? The most common confidence coefficients are

90% confident, 95% confident and 99% confident. Since our assumption about the probability distribution is normal distribution, $N(\mu, \sigma^2)$, equivalently, standard normal distribution, $Z \sim N(0,1)$, the Z-scores for the 90%, 95% and 99% are as follows:

90% – Z Score = 1.64

95% – Z Score = 1.96

99% – Z Score = 2.32

(d) Standard of Deviation:

How much variance do we expect in our responses? Since till the time we decide about it, the survey have not actually administered, the safe decision is to use 0.05; this is the most forgiving number and ensures that our sample will be large enough.

Suppose it is desired to find sample size such that the estimated value $\bar{y}$ differs from the true value $\bar{Y}$ by a quantity not exceeding a pre-assigned quantity D with a very high probability, say greater than $(1-\alpha)$. Therefore, in mathematical form the problem is to find the sample size n in such a way that

$$P\left[\left|\bar{y} - \bar{Y}\right| \le D\right] \ge (1-\alpha) \qquad \ldots (3.12)$$

Generally, in order to resolve the problem, it is assumed that the sample mean estimator $\bar{y}$ in Simple Random Sampling without Replacement scheme is approximately distributed as a normal distribution about the population mean $\bar{Y}$ with standard error $\sqrt{\frac{N-n}{Nn}}S$, whereas in case of Simple Random Sampling with Replacement scheme the standard error of the estimator $\bar{y}$ is $\sqrt{\frac{N-1}{nN}}S$.

Therefore, we may write,

$$P\left[\left|\bar{y} - \bar{Y}\right| \le Z_{\alpha,\infty} S\sqrt{\frac{N-n}{Nn}}\right] = (1-\alpha) \qquad \ldots (3.13)$$

and

$$P\left[\left|\bar{y} - \bar{Y}\right| \le Z_{\alpha,\infty} S\sqrt{\frac{N-1}{nN}}\right] = (1-\alpha)$$

respectively, for the case of Simple Random Sampling without Replacement and Simple Random Sampling with Replacement schemes, where $Z_{\alpha,\infty}$ is the value of the normal variable corresponding to the value $(1-\alpha/2)$ of the normal probability integral to hold on an average with a probability $(1-\alpha)$.

From equations (3.12) and (3.13), on comparison, we can see that

$$D = Z_{\alpha,\infty} S\sqrt{\frac{N-n}{Nn}} \quad \text{and} \quad D = Z_{\alpha,\infty} S\sqrt{\frac{(N-1)}{nN}} \qquad \ldots (3.14)$$

respectively, under Simple Random Sampling without Replacement and Simple Random Sampling with Replacement schemes. From where we get,

$$D^2 = Z^2_{\alpha,\infty} S^2 \left(\frac{1}{n} - \frac{1}{N}\right) \quad \Rightarrow \quad \frac{1}{n} - \frac{1}{N} = \frac{D^2}{Z^2_{\alpha,\infty} S^2}$$

$$\Rightarrow \qquad \frac{1}{n} = \frac{1}{N} + \frac{D^2}{Z^2_{\alpha,\infty} S^2}$$

$$\Rightarrow \qquad \frac{1}{n} = \frac{Z^2_{\alpha,\infty} S^2 + ND^2}{NZ^2_{\alpha,\infty} S^2}$$

$$\Rightarrow \qquad n = \frac{\frac{NZ^2_{\alpha,\infty} S^2}{ND^2}}{1 + \left(\frac{1}{N}\right)\left(\frac{Z^2_{\alpha,\infty} S^2}{D^2}\right)}$$

$$\text{or,} \quad n = \frac{\left\{\frac{Z_{\alpha,\infty}}{D} S\right\}^2}{1 + \left\{\frac{1}{N}\right\}\left\{\frac{Z_{\alpha,\infty}}{D} S\right\}^2} \qquad \ldots (3.15)$$

This is the value of the optimum size of the sample under the case of Simple Random Sampling without Replacement scheme which guarantees that the estimate is within the permissible margin of error with given level of accuracy of the result.

Similarly, expression of the optimum sample size under the Simple Random Sampling with Replacement scheme can easily be obtained using the equation

$$D = Z_{\alpha,\infty} S \sqrt{\frac{(N-1)}{nN}}\,.$$

Let us illustrate the computation of optimum sample size based on the following examples:

Example 3: In a population of 750 individuals, the coefficient of variation in the population was recorded as 0.226 from the past records. Due to cost constraints, the sampler wishes to go with the least sample size for estimating the population mean. If he agrees to bear with of a sample estimate which lies within 10% of the true value and a confidence coefficient of 95%; which one of the schemes Simple Random Sampling with Replacement or Simple Random Sampling without Replacement, he should adopt?

Solution: We are given that

N = 750, Coefficient of Variation $= \frac{S}{\bar{Y}} = 0.226$, $D = 0.1\bar{Y}$, $(1-\alpha) = 0.95$.

Also, from the normal table, we have Z value at 5% level of significance is 1.96.

Therefore, under Simple Random Sampling without Replacement scheme, using (3.14), we have,

$$0.1\bar{Y} = 1.96 \times 0.226\,\bar{Y}\sqrt{\frac{N-n}{Nn}}$$

$$\Rightarrow (0.1)^2 = (1.96)^2 \times (0.226)^2 \left(\frac{1}{n} - \frac{1}{N}\right)$$

$$\Rightarrow \left(\frac{1}{n} - \frac{1}{750}\right) = \frac{0.01}{3.8416 \times 0.051076}$$

$$\Rightarrow \frac{1}{n} = 0.05229 \Rightarrow n = 19.12 \approx 19$$

Thus, under the given margin of error and given confidence level, a sample of size 19 would be sufficient to provide the desired level of the accuracy of the estimate.

We know that under Simple Random Sampling with Replacement scheme the

$$V(\bar{y}) = \frac{N-1}{Nn}S^2.$$

Therefore, we have

$$0.1\bar{Y} = 1.96 \times 0.226\bar{Y}\sqrt{\frac{N-1}{Nn}}$$

$$\Rightarrow n = \frac{(1.96)^2 \times (0.226)^2}{(0.1)^2}\left(\frac{749}{750}\right)$$

Solving this equation for n, we have $n = 19.6 \cong 20$.

This shows that sample sizes needed under the two schemes are almost same. It is, therefore, advisable to use Simple Random Sampling without Replacement scheme.

Example 4: How large a sample should be taken from a population of 3000 units so that the sample estimate of the population mean differs from its true value by a quantity less than 15 with probability 0.95? It is given that an estimate of population variance is 1600.

Solution: We are given that N = 3000; S^2 =1600; D = 15; $(1-\alpha) = 0.95$

Therefore, we have the relation,

$$15 = 1.96 \times 40 \times \sqrt{\left(\frac{1}{n} - \frac{1}{3000}\right)}$$

or, $$(15)^2 = (1.96)^2 \times (40)^2 \times \left(\frac{1}{n} - \frac{1}{3000}\right).$$

$$\Rightarrow \frac{1}{n} = \frac{1}{3000} + \frac{225}{(1600 \times 1.96^2)} = 0.036938$$

$$\Rightarrow n = 27.07 \cong 27$$

3.7.3 Optimal Sample Size in Simple Random Sampling Scheme with Attributes

In such cases, we estimate population proportion on the basis of sample proportion, whose sampling variance is given by

$$\frac{N-n}{N-1}\frac{PQ}{n}$$

Then, we have equivalent expression to that given under (3.12) as

$$P\left[\left|p - P\right| \le D\right] \ge (1-\alpha) \qquad \ldots (3.16)$$

and that given in (3.13) as

$$P\left[\left|p - P\right| \le Z_{\alpha,\infty}\left(\frac{N-n}{n(N-1)}PQ\right)^{1/2}\right] = (1-\alpha) \qquad \ldots (3.17)$$

Comparing equations (3.16) and (3.17), we see that

$$Z_{\alpha,\infty}\left(\frac{N-n}{n(N-1)}PQ\right)^{1/2} = D$$

lying this equation for n, we get

$$n = \frac{\left\{\frac{Z_{\alpha,\infty}}{D}\right\}^2 PQ}{1+\left\{\frac{1}{N}\right\}\left\{\left(\frac{Z_{\alpha,\infty}}{D}\right)^2 PQ - 1\right\}}$$

This expression provides us the desired size of the sample.

Example 5: In a population of 200 employees in an office, the proportion of non-smokers is to be estimated based on a sample of appropriate size. The number of smokers in a similar population of 350 employees was found to be 45. The population proportion of non-smokers in this population can be taken to be the proportion of non-smokers in the concerned population. Find the required sample size to be selected from the population so that the sample proportion lies within 15% of the population proportion with a confidence coefficient of 95%.

Solution: We wish to estimate population proportion P in the given population. An approximate value of P is taken from the other population in which 45 smokers are observed out of 350 employees.

Therefore, the proportion of smokers in this population is 45/350 = 0.12857. This means that proportion of non- smokers in this population is 1 – 0.12857 = 0.87143. For finding the optimum sample size, therefore, we use the following equation:

$$Z_{\alpha,\infty}\left(\frac{N-n}{n(N-1)}PQ\right)^{1/2} = D \qquad \text{where } D = 0.15P \text{ (given).}$$

So, we have

$$0.15P = Z_{\alpha,\infty}\left(\frac{N-n}{n(N-1)}PQ\right)^{1/2}$$

$$\Rightarrow \quad (0.15)^2 P^2 = (1.96)^2 \frac{N-n}{n(N-1)}PQ$$

Therefore,

$$\frac{N-n}{n(N-1)} = \frac{(0.15)^2 P}{(1.96)^2 Q}$$

or

$$\frac{(200-n)}{n(200-1)} = \frac{(0.15)^2}{(1.96)^2}\frac{0.87143}{0.12857}$$

Solving this expression for n, we have n = 22.47 ≈ 23, which is the optimum sample size for estimating the proportion of non-smokers in the population with all criterions satisfied.

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

SAQ 4

In a sample survey, organized in a town with population of 6000 persons, it was required to estimate the average income of persons. For this purpose, a sample of appropriate size say, n_0 was decided to select from the population so that the sample estimate lies within 12% of the true value with a confidence coefficient of 95%. From a previous survey, it was observed that the coefficient of variation in the population was 0.70. What should be sample size n_0?

3.8 SUMMARY

In this unit we have discussed:

- Type of analysis of the gathered information on the basis of a random sample can be made if the information is purely of qualitative type, gathered using SRS scheme, both SRSWR and SRSWOR schemes
- The population parameters which are computed on the basis of qualitative type of data.
- The feasibility of using sampling schemes, such as Simple Random Sampling with Replacement and Simple Random Sampling without Replacement schemes, for the purpose of estimating parameters based on qualitative characteristics.
- The process of selecting units from the population under these sampling schemes, so as to constitute a random sample.

- The estimation process and corresponding estimator for population proportion when the qualitative characteristic under study possesses only two classes and SRSWOR scheme was followed.
- The unbiasedness property of the sample proportion estimator.
- Simple Random Sampling without Replacement scheme with qualitative characteristic is a special case of Simple Random Sampling without Replacement scheme with quantitative characteristic.
- The sampling variance of the sample proportion estimator and its unbiased estimator.
- The results of the Simple Random Sampling without Replacement scheme obtained with qualitative characteristic possessing only two levels was extended to more than two levels.
- The method of estimating the population proportion of the i^{th} class where i = 1, 2,…, k.
- Problem of determining the optimum sample size in SRS schemes with quantitative and qualitative characteristics both.
- Why a need arises to find the optimum size of the sample and what is the advantage of such sample size.
- The process of computing the optimum sample sizes in Simple Random Sampling with Replacement and Simple Random Sampling without Replacement schemes with both types of characteristics.

3.9 TERMINAL QUESTIONS

1. The record of annual examination in a college was published as follows:

Sex	Result of Students				Total
	First Division	Second Division	Third Division	Failed	
Male	18	24	9	10	**61**
Female	15	16	19	12	**62**
Total	**33**	**40**	**28**	**22**	**123**

Let M_{FD}, M_{SD}, M_{TD} and M_F be the number of male students passed with First, Second, Third division and Failed, respectively. Similarly, let F_{FD}, F_{SD}, F_{TD} and F_F respectively, have the same meaning for female students.

Select a sample of size 15 students from the population using the random number table of **Appendix - A**, Unit 1. Let a selected even number, odd number and prime number, respectively, denote a student with I division, II division and III division or failed. Find the estimates of population proportion of students under these three categories.

2. What is the utility of hypergeometric distribution in estimating the population proportion of an attribute? Using this distribution, show that

the variance of the sample proportion estimator is given by $\frac{N-n}{n(N-1)}PQ$, where the symbols have their usual meanings.

3. Let us consider the **Example 2** of Sub-section 3.3.1. Let the problem be to find an estimate of the proportion of Britishers in the population on the basis of a random sample of size 25 selected from the population of the capital city using Simple Random Sampling without Replacement scheme. You can use the Random Number Table given in the **Appendix – A** in Unit 1. Find the value of the estimate and its sampling variance of the estimator.

4. In a population of 650 workers in an organization, the proportion of female workers is to be estimated on the basis of a sample of appropriate size. The number of male workers in a similar organization of 250 workers was found to be 158. The population proportion of males in this organization can be taken to be the proportion of males in the organization concerned. Find the required sample size to be selected from the concerned organization so that the sample proportion lies within 10% of the population proportion with a confidence coefficient of 95%.

5. Find the estimator of the sampling variance of the sample proportion, p. Show that it is unbiased estimator of V(p).

3.10 ANSWERS / SOLUTIONS

Self-Assessment Questions (SAQs)

1. **Hint:** For the answer to the question, you are referred to Section 3.1.

2. **Hint:** You are referred to Section 3.4 for the answer to the question.

3. **Hint:** For the answer, you are referred to **Theorem 2** in the Sub-section 3.5.3.

4. It is given that

$$P\left[\left|\bar{y} - \bar{Y}\right| \le 0.12\bar{Y}\right] \ge (1-\alpha) \qquad \text{that is, } D = 0.12\bar{Y}$$

Therefore, we have,

$$0.12\bar{Y} = 1.96 \times 0.7\bar{Y}\sqrt{\frac{N-n}{Nn}} \qquad \text{since } S = 0.7\bar{Y}$$

Thus, we have

$$(0.12)^2 = (1.96)^2(0.7)^2\left(\frac{1}{n} - \frac{1}{N}\right)$$

$$\Rightarrow \qquad \frac{0.0144}{3.8416 \times 0.49} = \left(\frac{1}{n} - \frac{1}{600}\right)$$

$$\Rightarrow \quad \frac{1}{n} = \frac{1}{6000} + 0.007650$$

$$\Rightarrow \quad \frac{1}{n} = 0.0001667 + 0.007650 = 0.0078167$$

Therefore, $n_0 = \frac{1}{0.0078167} = 127.93 \approx 128.$

Terminal Questions (TQs)

1. A sample of size 15 has to be selected using Random Number Table given in the **Appendix - A** of the Unit 1.

Since N = 123, which is a three-digit number, in table we shall consider the last three digits of each random number. Let us start with the fifth column of the Random Number Table and move column-wise. Then you can see that the sample of size 15 consists of the following labels:

{81, 27, 113, 59 ,1, 36, 52, 10, 46, 63, 88, 90, 4, 115, 71}.

In these labels, we observe that there are 7 labels which are even; 4 labels are odd, and 4 labels are prime; therefore, the number of students with Ist division, IInd division and IIIrd division or failed are respectively, 7, 4 and 4; out of 15 students selected in the sample.

Therefore, respective sample proportions are:

$\frac{7}{15} = 0.47$, $\frac{4}{15} = 0.27$ and $\frac{4}{15} = 0.27$.

However, the respective population proportions are:

$\frac{33}{123} = 0.27$, $\frac{40}{123} = 0.32$ and $\frac{50}{123} = 0.41.$

2. **Hint:** You are referred to **Theorem 1** in the Sub-section 3.5.2 and **Theorem 3** in the Sub-section 3.5.3.

3. The Example 2 in Sub-section 3.3.1 is solved with the help of a sample of size 36. Instead of selecting a different sample here, let us assume that the first 25 units of the previous sample appeared in the sample drawn here. These 25 labels are as follows:

{154, 271, 341, 112, 177, 280, 257, 143, 73, 93, 298, 264, 248, 101, 176, 179, 150, 44, 277, 35, 188, 245, 169, 45, 162}.

Since the labels 46 to 100 belongs to Britishers, we see that there are 2 labels; i.e., 73 and 93 which correspond to Britishers.

Therefore, the estimated proportion of Britishers, $p_2 = \frac{2}{25} = 0.08$ for the actual population proportion $P_2 = \frac{55}{360} = 0.15$.

The sampling variance of the estimator is given by

$$V(p_2) = \frac{N-n}{N-1}\frac{P_2Q_2}{n}$$ (see sub-section 3.6.2).

Here, N = 360, n = 25, P_2 = 0.15 and Q_2 = 1 - 0.15 = 0.85.

Therefore,

$$V(p_2) = \frac{360-25}{359}.\frac{0.15\times 0.85}{25} = 0.0048.$$

4. Since there are some male and some female workers working in the organization, so it is a problem with two classes only. Given that in other organization the number of male workers is 158 out of 250 workers, therefore, there would be 250 – 158 = 92 females. So, the proportion of females would be

$$P = \frac{92}{250} = 0.368$$

This is taken to be the population proportion of female workers, P, in the organization. In question. It is given that D = 0.10P.

Therefore, we have the relation:

$$D = Z_{\alpha,\infty}\left(\frac{N-n}{n(N-1)}PQ\right)^{1/2}$$

$$\Rightarrow \quad (0.10x0.368)^2 = (1.96)^2.\frac{650-n}{n(649)}(0.368 \times 0.632)$$

$$\Rightarrow \quad 0.001354 = \frac{650-n}{n(649)}(0.8935)$$

$$\Rightarrow \quad 0.878746n + 0.8935n = 580.775$$

$$\Rightarrow \quad 1.772246n = 580.775$$

$$\Rightarrow \quad n \approx 328$$

5. **Hint:** You are referred to Sub-section 3.5.5 for the answer.