Chapter 11

SAMPLING AND SAMPLING DISTRIBUTIONS

11.1 INTRODUCTION

In our daily life it is quite often that we have to examine some given material. We examine fruit before we purchase it, we make a small study of the material whenever we have to purchase something. Even the children check the sweets, pencils, bats, rubbers and other items when they have to purchase them. This approach is applied in different fields of life. The products of the factories are inspected to ensure the desired quality of the products. The medicines are manufactured on commercial scale when their effects have been tested on the patients. The different fertilizers are tested on agricultural plots and different foods are tested on animals. Small dams are constructed in the laboratories to study the life and other characteristics of the big dams before they are actually constructed. Some colour may be applied on a wall, on a door or cloth etc., and the result of the colour is observed before it is applied on large scale. Cement, steel and bricks are examined before using them in different places. This process of inspection is very wide and is commonly used on various occasions. But this job is never done on very large scale. This process is carried out on a small scale. On the basis of this small study, we make an opinion about the entire material under study.

11.2 POPULATION

The word *population* or *statistical population* is used for all the individuals or objects on which we have to make some study. We may be interested to know the quality of bulbs produced in a factory. The entire product of the factory in a certain period is called a *population*. We may be interested in the level of education in primary schools. All the children in the primary schools will make a *population*. The *population* may contain living or non–living things. The entire lot of anything under study is called *population*. All the fruit trees in a garden, all the patients in a hospital and all the cattle in a cattle farm are examples of *populations* in different studies.

11.2.1 FINITE POPULATION

A *population* is called *finite* if it is possible to count its individuals. It may also be called a *countable population*. The number of vehicles crossing a bridge every day, the number of deaths per year and the number of words in a book are *finite populations*. The number of units in a *finite population* is denoted by N. Thus N is the size of the *population*.

11.2.2 INFINITE POPULATION

Sometimes it is not possible to count the units contained in the *population*. Such a *population* is called *infinite or uncountable*. Let us suppose that we want to examine whether a coin is true or not. We shall toss it a very large number of times to observe the number of heads. All the tosses will make an *infinite* or *countably infinite population*. The number of germs in the body of a patient of malaria is perhaps something which is *uncountable*.

11.2.3 TARGET AND SAMPLED POPULATION

Suppose we have to make a study about the problems of the families living in rented houses in a certain big city. All the families living in rented houses is our *target population*. The entire *target population* may not be considered for the purpose of selecting a sample from the population. Some families may not be interested to be included in the sample. We may ignore some part of the *target population* to reduce the cost of study. The *population* out of which the sample is selected is called *sampled population* or *studied population*.

11.3 SAMPLE

Any part of the population is called a *sample*. A study of the *sample* enables us to make some decisions about the properties of the population. The number of units included in the *sample* is called the size of the *sample* and is denoted by n. A good *sample* is that one which speaks about the qualities of the population. A *sample* study leads us to make some inferences about the population measures. This process is called *sampling*.

11.3.1 PARAMETER AND STATISTIC

Any measure of the population is called *parameter* and the word *statistic* is used for any value calculated from the sample. The population mean μ is a *parameter* and the sample mean $\bar{X}$ is a *statistic*. The sample mean $\bar{X}$ is used to estimate the population mean μ. Similarly the population variance σ^2 is a *parameter* and the sample variance S^2 is a *statistic*. In general the symbol θ is used for a *parameter* and the symbol $\hat{\theta}$ is used for a statistic. The value of the *parameter* is mostly unknown and the sample statistic is used to make some inferences about the unknown *parameter*.

11.3.2 SAMPLING FRACTION

If size of the population is N and size of the sample is n, the ratio $\frac{n}{N}$ is called the *sampling fraction*. If N = 100, n = 10, the ratio $\frac{n}{N} = \frac{10}{100} = \frac{1}{10}$. It means that on the average 10 units of the population will be represented by one unit in the sample. If the *sampling fraction* $\frac{n}{N}$ is multiplied with 100, we get the sampling *fraction* in percentage form. Thus $\frac{n}{N} \times 100 = \frac{10}{100} \times 100 = 10\ \%$. It means 10 % of the population is included in the sample.

11.4 COMPLETE COUNT

If we collect information about all the individuals in the population, the study is called *complete count* or *complete enumeration*. The word *census* is also used for the entire population study. In statistical studies the *complete count* is usually avoided. If size of the population is large, the *complete count* requires a lot of time and a lot of funds. The *complete count* is mostly difficult for various reasons. Suppose we want to make a study about the cattle in the cattle farms in our country. We are interested in the average cost of their food for a certain period. We want to link their cost of food with their sale price. This is of course, an important study. It is very difficult to collect and maintain the information about each and every cattle in the farms. If at all we are able to do it, the study may not be of much use. The desired information can be obtained from a reasonable sample size of the cattles.

11.4.1 POPULATION CENSUS

A complete count of the human population is called *population census*. In Pakistan, the first *population census* was conducted in 1951 and the second was conducted in 1961. The third *census of population* could not be conducted in 1971 because of agitations in the then East Pakistan. It was conducted in 1972. The 4th *census* was conducted in 1981. The fifth population census was conducted in 1998. A lot of information is collected about the human population through the *population census* conducted regularly after every 10 years. The *census* reports give information about various characteristics of the population e.g., the urban and rural population, the skilled and un-skilled labour force, the agricultural labour force and the industrial workers, level of education and illiteracy in the country, geographical distribution of the population, age and sex distribution of the population etc.

11.5 SAMPLE SURVEY

If it is not essential to conduct the complete enumeration, then a sample of some suitable size is selected from the population and the study is carried out on the sample. This study is called *sample survey*. Most of the research work is done through *sample surveys*. The opinion of the voters in favour of certain proposed election candidates is obtained through *sample surveys*

11.5.1 ADVANTAGES OF SAMPLING

Sampling has some advantages over the complete count. These are:

(i) Need for Sampling

Sometimes there is a need for sampling. Suppose we want to inspect the eggs, the bullets, the missiles and the tires of some firm. The study may be such that the objects are destroyed during the process of inspection. Obviously, we cannot afford to destroy all the eggs and the bullets etc. We have to take care that the wastage should be minimum. This is possible only in sample study. Thus sampling is essential when the units under study are destroyed.

(ii) Saves Time and Cost

As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. For complete count huge funds are required. There is always the problem of finances. A small

sample can be studied in a limited time and total cost of sample study is very small. For complete count, we need a big team of supervisors and enumerators who are to be trained and they are to be paid properly for the work they do. Thus the sample study requires less time and less of cost.

(iii) Reliability

If we collect the information about all the units of population, the collected information may be true. But we are never sure about it. We do not know whether the information is true or is completely false. Thus we cannot say anything with confidence about the quality of information. We say that the *reliability* is not possible. This is a very important advantage of sampling. The inference about the population parameters is possible only when the sample data is collected from the selected sample.

(iv) Sometimes the experiments are done on sample basis. The fertilizers, the seeds and the medicines are initially tested on samples and if found useful, then they are applied on large scale. Most of the research work is done on the samples.

(v) Sample data is also used to check the accuracy of the census data.

11.5.2 LIMITATIONS OF SAMPLING

Sometimes the information about each and every unit of the population is required. This is possible only through the complete enumeration because the sample will not serve the purpose. Some examples in which the sampling is not allowed are:

(i) To conduct the elections, we need a complete list of the voters. The candidates participating in the election will not accept the results prepared from a sample. With increase in literacy, the people may become statistical minded and they may become willing to accept the results prepared from the sample. In advanced countries the opinion polls are frequently conducted and unofficially the people accept the results of sample surveys.

(ii) Tax is collected from all the tax payers. A complete list of all the tax payers is required. The telephone, gas and electricity bills are sent to all the consumers. A complete list of the owners of land and property is always prepared to maintain the records. The position of stocks in factories requires complete entries of all the items in the stock.

11.5.3 SAMPLE DESIGN

In sample studies, we have to make a plan regarding the size of the sample, selection of the sample, collection of the sample data and preparation of the final results based on the sample study. The whole procedure involved is called the *sample design*. The term sample survey is used for a detailed study of the sample. In general, the term sample survey is used for any study conducted on the sample taken from some real world data.

11.5.4 SAMPLING FRAME

A complete list of all the units of the population is called the *sampling frame*. A *unit* of population is a relative term. If all the workers in a factory make a

population, a single worker is a unit of the population. If all the factories in a country are being studied for some purpose, a single factory is a unit of the population of factories. The *sampling frame* contains all the units of the population. It is to be defined clearly as to which units are to be included in the frame. The frame provides a base for the selection of the sample.

11.5.5 EQUAL PROBABILITY

The term equal probability is frequently used in the theory of sampling. This term is quite often not understood correctly. It is thought to be close to 'equal' in meaning. It is not true always. Suppose there is a population of 50(N = 50) students in a class. We select any one student. Every student has probability 1/50 of being selected. Then a second student is selected. Now, there are 49 students in the population and every student has 1/49 probability of being selected. When the first student is selected, all the students have equal (1/50) chance of selection and when the second student is selected, again all the students have equal (1/49) chance of selection. But 1/50 is not equal to 1/49. Thus equal probability of selection means the probability when the individual is selected from the remaining available units in the population. At the time of selecting a unit, the probability of selection is equal. It is called *equal probability* of selection.

11.5.6 KNOWN PROBABILITY

In sampling theory the term *known probability* is used in random (probability) sampling. Let us explain it by taking an example. Suppose there are 300 workers in a certain factory out of which 200 are skilled and 100 are non-skilled. We have to select one sample (sub-sample) out of skilled workers and one sample out of un-skilled workers. When the first worker out of skilled workers is selected, each worker has a probability of selection equal to 1/200. Similarly when the first worker out of un-skilled workers is selected, each worker has a probability of selection equal to 1/100. Both these probabilities are *known,* though they are not equal.

11.5.7 NON-ZERO PROBABILITY

Suppose we have a population of 500 students out of which 50 are non-intelligent. We have decided to select an intelligent student from the population. The probability of selecting an intelligent student is 1/450 which is *non-zero.* In this example, we have decided to exclude the non-intelligent students from the population for the purpose of selecting a sample. Thus probability of selecting a non-intelligent student is zero.

11.6 PROBABILITY AND NON–PROBABILITY SAMPLING

The term *probability sampling* is used when the selection of the sample is purely based on chance. The human mind has no control on the selection or non-selection of the units for the sample. Every unit of the population has known non-zero probability of being selected for the sample. The probability of selection may be equal or unequal but it should be non–zero and should be *known.* The *probability sampling* is also called the random sampling (not simple random sampling). Some examples of random sampling are:-

(i) Simple random sampling

(ii) Stratified random sampling

(iii) Systematic random sampling.

In *non-probability sampling*, the sample is not based on chance. It is rather determined by some person. We cannot assign to an element of population the probability of its being selected in the sample. Somebody may use his personal judgement in the selection of the sample. In this case the sampling is called *judgement sampling*. A drawback in *non-probability sampling* is that such a sample cannot be used to determine the error. Any statistical method cannot be used to draw inference from this sample. But it should be remembered that judgement sampling becomes essential in some situations. Suppose we have to take a small sample from a big heap of coal. We cannot make a list of all the pieces of coal. The upper part of the heap will have perhaps big pieces of coal. We have to use our judgement in selecting a sample to have an idea about the quality of coal. The *non-probability sampling* is also called non-random sampling.

11.6.1 SAMPLING WITH REPLACEMENT

Sampling is called *with replacement* when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units. A unit may be selected more than once. There is no change at all in the size of the population at any stage. We can assume that a sample of any size can be selected from the given population of any size. This is only a theoretical concept and in practical situations the sample is not selected by using this scheme of selection. Suppose the population size $N = 5$ and sample size $n = 2$, and sampling is done *with replacement*. Out of 5 elements, the first element can be selected in 5 ways. The selected unit is returned to the main lot and now the second unit can also be selected in 5 ways. Thus in total there are $5 \times 5 = 25$ samples or pairs which are possible. Suppose a container contains 3 good bulbs denoted by G_1, G_2 and G_3 and 2 defective bulbs denoted by D_1 and D_2. If any two bulbs are selected *with replacement*, there are 25 possible samples listed between in Table 11.1.

Table 11.1

	G_1	G_2	G_3	D_1	D_2
G_1	G_1G_1	G_1G_2	G_1G_3	G_1D_1	G_1D_2
G_2	G_2G_1	G_2G_2	G_2G_3	G_2D_1	G_2D_2
G_3	G_3G_1	G_3G_2	G_3G_3	G_3D_1	G_3D_2
D_1	D_1G_1	D_1G_2	D_1G_3	D_1D_1	D_1D_2
D_2	D_2G_1	D_2G_2	D_2G_3	D_2D_1	D_2D_2

The number of samples is given by $N^n = 5^2 = 25$. The selected sample will be any one of the 25 possible samples. Each sample has equal probability 1/25 of selection. A sample selected in this manner is called simple random sample.

11.6.2 SAMPLING WITHOUT REPLACEMENT

Sampling is called *without replacement* when a unit is selected at random from the population and it is not returned to the main lot. First unit is selected out of a population of size N and the second unit is selected out of the remaining population of N – 1 units and so on. Thus the size of the population goes on decreasing as the sample size n increases. The sample size n cannot exceed the population size N. The unit once selected for a sample cannot be repeated in the same sample. Thus all the units of the sample are distinct from one another. A sample *without replacement* can be selected either by using the idea of permutations or combinations. Depending upon the situation, we write all possible permutations or combinations. If the different arrangements of the units are to be considered, then the permutations (arrangements) are written to get all possible samples. If the arrangement of units is of no interest, we write the combinations to get all possible samples.

11.6.3 COMBINATIONS

Let us again consider a lot (population) of 5 bulbs with 3 good (G_1, G_2 and G_3) and 2 defective (D_1 and D_2) bulbs. Suppose we have to select two bulbs in any order there are ${}^5C_2 = \frac{5!}{2!\,3!} = 10$ possible *combinations or samples*. These *combinations* (samples) are listed as: G_1G_2, G_1G_3, G_2G_3, G_1D_1, G_1D_2, G_2D_1, G_2D_2, G_3D_1, G_3D_2, D_1D_2.

There are 10 possible samples and each of them has probability of selection equal to 1/10. The selected sample will be any one of these 10 samples. The sample selected in this manner is also called simple random sample. In general, the number of samples by *combinations* is equal to ${}^NC_n = \frac{N!}{n!(N-n)!}$.

11.6.4 PERMUTATIONS

Each combination generates a number of arrangements (*permutations*). Thus in general the number of *permutations* is greater than the number of combinations. In the previous example of bulbs, if the order of the selected bulbs is to be considered then the number of samples by *permutations* is given by ${}^5P_2 = \frac{5!}{(5-2)!} = 20$. These samples are:

G_1G_2 G_2G_1 G_1G_3 G_3G_1 G_2G_3 G_3G_2 G_1D_1 D_1G_1 G_1D_2 D_2G_1
G_2D_1 D_1G_2 G_2D_2 D_2G_2 G_3D_1 D_1G_3 G_3D_2 D_2G_3 D_1D_2 D_2D_1

Each sample has probability of selection equal to 1/20. The selected sample keeping in view the order of the bulbs will be any one of these 20 samples. A sample selected in this manner is also called simple random sample because each sample has equal probability of being selected.

11.6.5 SIMPLE RANDOM SAMPLE

Simple random sample (SRS) is a special case of a random sample. A sample is called *simple random sample* if each unit of the population has an equal chance of being selected for the sample. Whenever a unit is selected for the sample, the units

of the population are equally likely to be selected. It must be noted that the probability of selecting the first element is not to be compared with the probability of selecting the second unit. When the first unit is selected, all the units of the population have the equal chance of selection which is 1/N. When the second unit is selected, all the remaining (N – 1) units of the population have 1/(N – 1) chance of selection.

Another way of defining a *simple random sample* is that if we consider all possible samples of size n, then each possible sample has equal probability of being selected.

If sampling is done with replacement, there are N^n possible samples and each sample has probability of selection equal to $1/N^n$. If sampling is done without replacement with the help of combinations then there are ${}^{N}C_{n}$ possible samples and each sample has probability of selection equal to $1/{}^{N}C_{n}$. If samples are made with permutations, each sample has probability of selection equal to $1/{}^{N}P_{n}$. Strictly speaking, the sample selected by without replacement is called *simple random sample.*

11.6.6 DIFFERENCE BETWEEN RANDOM SAMPLE AND SIMPLE RANDOM SAMPLE

If each unit of the population has known (equal or un-equal) probability of selection in the sample, the sample is called a random sample. If each unit of the population has *equal* probability of being selected for the sample, the sample obtained is called simple random sample.

11.6.7 SELECTION OF SIMPLE RANDOM SAMPLE

A *simple random sample* is usually selected by without replacement. The following methods are used for the selection of a *simple random sample*:

(i) Lottery Method

This is an old classical method but it is a powerful technique and modern methods of selection are very close to this method. All the units of the population are numbered from 1 to N. This is called sampling frame. These numbers are written on the small slips of paper or the small round metallic balls. The paper slips or the metallic balls should be of the same size otherwise the selected sample will not be truly random. The slips or the balls are thoroughly mixed and a slip or ball is picked up. Again the population of slips is mixed and the next unit is selected. In this manner, the number of slips equal to the sample size n are selected. The units of the population which appear on the selected slips make the *simple random sample.* This method of selection is commonly used when size of the population is small. For a large population there is a big heap of paper slips and it is difficult to mix the slips properly.

(ii) Using a Random Number Table

All the units of the population are numbered from 1 to N or from 0 to N – 1. We consult the random number table to take a *simple random sample.* Suppose the size of the population is 80 and we have to select a random sample of 8 units. The units

of the population are numbered from 01 to 80. We read two-digit numbers from the table of random numbers. We can take a start from any columns or rows of the table. Let us consult *random number table* given in this book. Two-digit numbers are taken from the table. Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement. Let us read the first two columns of the table. The random number from the table are 10, 37, 08, 12, 66, 31, 63 and 73. The two numbers 99 and 85 have not been recorded because the population does not contain these numbers. The units of the population whose numbers have been selected constitute the *simple random sample*. Let us suppose that the size of the population is 100. If the units are numbered from 001 to 100, we shall have to read 3-digit random numbers. From the first 3 columns of the random number table, the random numbers are 100, 375, 084, 990, 128 and so on. We find that most of the numbers are above 100 and we are wasting our time while reading the table. We can avoid it by numbering the units of the population from 00 to 99. In this way, we shall read 2-digit numbers from the table. Thus if N is 100, 1000 or 10000, the numbering is done from 00 to 99, 000 to 999 or 0000 to 9999.

(iii) Using the Computer

The facility of selecting a *simple random sample* is available on the computers. The computer is used for selecting a sample of prize-bond winners, a sample of Haj applicants, a sample of applicants for residential plots and for various other purposes.

11.7 ERRORS

Suppose we are interested in the value of a population parameter, the true value of which is 0 but is unknown. The knowledge about 0 can be obtained either from a sample data or from the population data. In both cases, there is a possibility of not reaching the true value of the parameter. The difference between the calculated value (from sample data or from population data) and the true value of the parameter is called *error*. Thus *error* is something which cannot be determined accurately if the population is large and the units of the population are to be measured. Suppose we are interested to find the total production of wheat in Pakistan in a certain year. Sufficient funds and time are at our disposal and we want to get the 'true' figure about production of wheat. The maximum we can do is that we contact all the farmers and suppose all the farmers give maximum cooperation and supply the information as honestly as possible. But the information supplied by the farmers will have *errors* in most of the cases. Thus we may not be able to identify the 'true' figure. Inspite of all efforts, we shall be in darkness. The calculated or the observed figure may be good for all practical purposes but we can never claim that a true value of the parameter has been obtained. If the study of the units is based on 'counting' may be we can get the true figure of the population parameter. There are two kinds of *errors* (i) sampling errors or random errors (ii) non-sampling errors

11.7.1 SAMPLING ERRORS

These are the errors which occur due to the nature of sampling. The sample selected from the population is one of all possible samples. Any value calculated from the sample is based on the sample data and is called sample statistic. The sample statistic may or may not be close to the population parameter. If the statistic is $\hat{\theta}$ and the true value of the population parameter is θ, then the difference $\hat{\theta} - \theta$ is called *sampling error*. It is important to note that a statistic is a random variable and it may take any value. A particular example of *sampling error* is the difference between the sample mean $\bar{X}$ and the population mean μ. Thus sampling error is also a random term. The population parameter is usually not known, therefore the *sampling error* is estimated from the sample data. The *sampling error* is due to the reason that a certain part of the population goes to the sample. Obviously, a part of the population cannot give the true picture of the properties of the population. But one should not get the impression that a sample always gives the result which is full of errors. We can design a sample and collect the sample data in a manner so that the *sampling errors* are reduced. The *sampling errors* can be reduced by the following methods:

(i) by increasing the size of the sample (ii) by stratification.

11.7.2 REDUCING THE SAMPLING ERRORS

(i) By Increasing the size of the sample

The sampling error can be reduced by increasing the sample size. If the sample size n is equal to the population size N, then the sampling error is zero.

(ii) By Stratification

When the population contains homogeneous units, a simple random sample is likely to be representative of the population. But if the population contains dissimilar units, a simple random sample may fail to be representative of all kinds of units in the population. To improve the result of the sample, the sample design is modified. The population is divided into different groups containing similar units. These groups are called *strata*. From each group (stratum), a sub-sample is selected in a random manner. Thus all the groups are represented in the sample and sampling error is reduced. It is called stratified-random sampling. The size of the sub-sample from each stratum is frequently in proportion to the size of the stratum. Suppose a population consists of 1000 students out of which 600 are intelligent and 400 are non-intelligent. We are assuming here that we do have this much information about the population. A stratified sample of size n = 100 is to be selected. The size of the stratum is denoted by N_1 and N_2 respectively and the size of the samples from each stratum may be denoted by n_1 and n_2. It is written as under:

Stratum No.	Size of stratum	Size of sample from each stratum
1	$N_1 = 600$	$n_1 = \frac{n \times N_1}{N} = \frac{100 \times 600}{1000} = 60$
2	$N_2 = 400$	$n_2 = \frac{n \times N_2}{N} = \frac{100 \times 400}{1000} = 40$
	$N_1 + N_2 = N = 1000$	$n_1 + n_2 = n = 100$

The size of the sample from each stratum has been calculated according to the size of the stratum. This is called *proportional allocation.* In the above sample design, the sampling fraction in the population is $\frac{n}{N} = \frac{100}{1000} = \frac{1}{10}$ and the sampling fraction in both the strata is also 1/10 . Thus this design is also called *fixed sampling fraction.* This modified sample design is frequently used in sample surveys But this design requires some prior information about the units of the population. On the basis of this information, the population is divided into different strata. If the prior information is not available then the stratification is not applicable.

11.7.3 NON-SAMPLING ERRORS

There are certain sources of errors which occurs both in sample survey as well as in the complete enumeration. These errors are of common nature. Suppose we study each and every unit of the population. The population parameter under study is the population mean and the 'true' value of the parameter is μ which is unknown. We hope to get the value of μ by a complete count of all the units of the population. We get a value called 'calculated' or 'observed' value of the population mean. This observed value may be denoted by μ_{cal}. The difference between μ_{cal} and μ (true) is called *non-sampling error.* Even if we study the population units under ideal conditions, there may still be the difference between the observed value of the population mean and the true value of the population mean. *Non-sampling errors* may occur due to many reasons. Some of them are:

(i) The units of the population may not be defined properly. Suppose we have to carry out a study about skilled labour force in our country. Who is a skilled person. Some people do more than one job. Some do the secretariat jobs as well as the technical jobs. Some are skilled but they are doing the job of un-skilled worker. Thus it is important to clearly define the units of the population otherwise there will be *non-sampling errors* both in the population count and the sample study.

(ii) There may be poor response on the part of respondents. The people do not supply correct information about their income, their children, their age and property etc. These errors are likely to be of high magnitude in population study than the sample study. To reduce these errors the respondents are to be persuaded.

(iii) The things in human hand are likely to be mis-handled. The enumerators may be careless or they may not be able to maintain uniformity from place to place. The data may not be collected properly from the population or from the sample. These errors are likely to be more serious in the population data than the sample data.

(iv) Another serious error is due to 'bias'. Bias means an error on the part of the enumerator or the respondent when the data is being collected. Bias may be intentional or un-intentional. An enumerator may not be capable of reporting the correct data. If he has to report about the condition of crops in different areas after heavy rainfalls, his assessments may be biased due to lack of training or he may be inclined to give wrong reports. Bias is a serious error and cannot be reduced by increasing the sample size. Bias may be present in the sample study as well as the population study.

11.8 SAMPLING DISTRIBUTIONS

Suppose we have a finite population and we draw all possible simple random samples of size n by without replacement or with replacement. For each sample we calculate some statistic (sample mean $\bar{X}$ or proportion $\hat{p}$ etc.). All possible values of the statistic make a probability distribution which is called the *sampling distribution*. The number of all possible samples is usually very large and obviously the number of statistics (any function of the sample) will be equal to the number of sample if one and only one statistic is calculated from each sample. In fact, in practical situations, the *sampling distribution* has very large number of values. The shape of the *sampling distribution* depends upon the size of the sample and the nature of the population and the statistic which is calculated from all possible simple random samples. Some of the famous *sampling distributions* are:

(i) Binomial distribution. (ii) Normal distribution. (iii) t-distribution.
(iv) Chi-square distribution. (v) F-distribution.

These distributions are called the derived distributions because they are derived from all possible samples.

11.8.1 STANDARD ERROR

The standard deviation of some statistic is called the *standard error* of that statistic. If the statistic is $\bar{X}$, the standard deviation of all possible values of $\bar{X}$ is called *standard error* of $\bar{X}$ which may be written as S.E. $(\bar{X})$ or $\sigma_{\bar{X}}$. Similarly, if the sample statistic is proportion $\hat{p}$, the standard deviation of all possible values of $\hat{p}$ is called *standard error* of $\hat{p}$ and is denoted by $\sigma_{\hat{p}}$ or S.E.$(\hat{p})$.

11.8.2 SAMPLING DISTRIBUTION OF $\bar{X}$

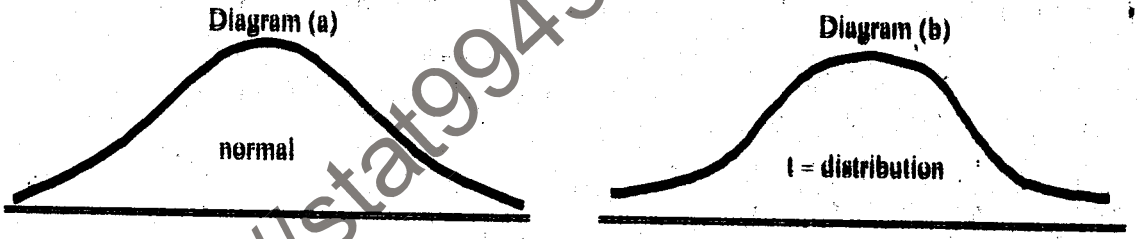
The probability distribution of all possible values of $\bar{X}$ calculated from all possible simple random samples is called the *sampling distribution* of $\bar{X}$. In brief, we shall call it distribution of $\bar{X}$. The mean of this distribution is called expected valu

of $\bar{X}$ and is written as $E(\bar{X})$ or $\mu_{\bar{X}}$. The standard deviation (standard error) of this distribution is denoted by S.E.($\bar{X}$) or $\sigma_{\bar{X}}$ and the variance of $\bar{X}$ is denoted by Var ($\bar{X}$) or $\sigma^2_{\bar{X}}$. The distribution of $\bar{X}$ has some important properties as under:

(i) An important property of the distribution of $\bar{X}$ is that it is a normal distribution when the size of the sample is large. When the sample size n is more than 30, we call it a large sample size. The shape of the population distribution does not matter. The population may be normal or non-normal, the distribution of $\bar{X}$ is normal for $n \geq 30$. But this is true when the number of samples is very large.

As the distribution of random variable $\bar{X}$ is normal, $\bar{X}$ can be transformed into standard normal variable Z where $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

The distribution of $\bar{X}$ has the t-distribution when the population is normal and $n \leq 30$. Diagram (a) shows the normal distribution and diagram (b) shows the t-distribution.



(ii) The mean of the distribution of $\bar{X}$ is equal to the mean of the population. Thus $E(\bar{X}) = \mu_{\bar{X}} = \mu$ (Population mean). This relation is true for small as well as large sample size in sampling without replacement and with replacement.

(iii) The standard error (standard deviation) of $\bar{X}$ is related with the standard deviation of population σ through the relations:

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

This is true when population is infinite which means N is very large or the sampling is done with replacement from finite or infinite population.

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

This is true when sampling is without replacement from finite population. The above two equations between $\sigma_{\bar{X}}$ and σ are true both for small as well as large sample sizes.

Example 11.1.

Draw all possible samples of size 2 without replacement from a population consisting of 3, 6, 9, 12, 15. Form the sampling distribution of sample means and verify the results:

(i) $E(\bar{X}) = \mu$ (ii) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

Solution:

We have population values 3, 6, 9, 12, 15, population size N = 5 and sample size n = 2. Thus, the number of possible samples which can be drawn without replacement is

$\binom{N}{n} = \binom{5}{2} = 10.$

Sample No.	Sample Values	Sample Mean ($\bar{X}$)	Sample No.	Sample Values	Sample Mean ($\bar{X}$)
1	3, 6	4.5	6	6, 12	9.0
2	3, 9	6.0	7	6, 15	10.5
3	3, 12	7.5	8	9, 12	10.5
4	3, 15	9.0	9	9, 15	12.0
5	6, 9	7.5	10	12, 15	13.5

The sampling distribution of the sample mean $\bar{X}$ and its mean and standard deviation are:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}\, f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
4.5	1	1/10	4.5/10	20.25/10
6.0	1	1/10	6.0/10	36.00/10
7.5	2	2/10	15.0/10	112.50/10
9.0	2	2/10	18.0/10	162.00/10
10.5	2	2/10	21.0/10	220.50/10
12.0	1	1/10	12.0/10	144.00/10
13.5	1	1/10	13.5/10	182.25/10
Total	10	1	90/10	877.5/10

$$E(\bar{X}) = \Sigma \bar{X} f(\bar{X}) = \frac{90}{10} = 9$$

$$\text{Var}(\bar{X}) = \Sigma \bar{X}^2 f(\bar{X}) - \left[\Sigma \bar{X} f(\bar{X})\right]^2 = \frac{877.5}{10} - \left(\frac{90}{10}\right)^2 = 87.75 - 81 = 6.75$$

The mean and variance of the population are:

X	3	6	9	12	15	$\Sigma X = 45$
X^2	9	36	81	144	225	$\Sigma X^2 = 495$

$$\mu = \frac{\Sigma X}{N} = \frac{45}{5} = 9 \text{ and } \sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{495}{5} - \left(\frac{45}{5}\right)^2 = 99 - 81 = 18$$

Verification:

(i) $E(\bar{X}) = \mu = 9$ (ii) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{18}{2}\left(\frac{5-2}{5-1}\right) = 6.75$

Example 11.2

If random samples of size three are drawn without replacement from the population consisting of four numbers 4, 5, 5, 7. Find sample mean $\bar{X}$ for each sample and make sampling distribution of $\bar{X}$. Calculate the mean and standard deviation of this sampling distribution. Compare your calculations with population parameters.

Solution:

We have population values 4, 5, 5, 7, population size N = 4 and sample size n = 3. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{4}{3} = 4$.

Sample No.	Sample Values	Sample Mean ($\bar{X}$)
1	4, 5, 5	14/3
2	4, 5, 7	16/3
3	4, 5, 7	16/3
4	5, 5, 7	17/3

The sampling distribution of the sample mean $\bar{X}$ and its mean and standard deviation are:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}\, f(\bar{X})$	$\bar{X}^2\, f(\bar{X})$
14/3	1	1/4	14/12	196/36
16/3	2	2/4	32/12	512/36
17/3	1	1/4	17/12	289/36
Total	4	1	63/12	997/36

$$\mu_{\bar{x}} = \Sigma \bar{X}\, f(\bar{X}) = \frac{63}{12} = 5.25$$

$$\sigma_{\bar{x}} = \sqrt{\Sigma \bar{X}^2\, f(\bar{X}) - [\Sigma \bar{X}\, f(\bar{X})]^2} = \sqrt{\frac{997}{36} - \left(\frac{63}{12}\right)^2} = 0.3632$$

The mean and standard deviation of the population are:

X	4	5	5	7	$\Sigma X = 21$
X^2	16	25	25	49	$\Sigma X^2 = 115$

$$\mu = \frac{\Sigma X}{N} = \frac{21}{4} = 5.25 \text{ and } \sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{115}{4} - \left(\frac{21}{4}\right)^2} = 1.0897$$

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1.0897}{\sqrt{2}}\sqrt{\frac{4-2}{4-1}} = 0.6532$$

Hence $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Example 11.9

Take all possible samples of size two with replacement from the population 2, 2, 8. Show that the population mean is equal to the mean of means of all samples and population variance is twice the variance of sample means.

Solution:

We have population values 2, 2, 8, population size N = 3 and sample size n = 2. Thus, the number of possible samples which can be drawn with replacement is $N^n = 3^2 = 9$.

Sample No.	Sample Values	Sample Mean ($\bar{X}$)	Sample No.	Sample Values	Sample Mean ($\bar{X}$)
1	2, 2	2	6	2, 8	5
2	2, 2	2	7	8, 2	5
3	2, 8	5	8	8, 2	5
4	2, 2	2	9	8, 8	8
5	2, 2	2			

The sampling distribution of the sample mean $\bar{X}$ and its mean and variance are:

$\bar{X}$	Tally	f	$f(\bar{X})$	$\bar{X}\, f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
2	||||	4	4/9	8/9	16/9
5	||||	4	4/9	20/9	100/9
8	|	1	1/9	8/9	64/9
Total		9	1	36/9	180/9

$$E(\bar{X}) = \Sigma \bar{X} f(\bar{X}) = \frac{36}{9} = 4$$

$$Var(\bar{X}) = \Sigma \bar{X}^2 f(\bar{X}) - [\Sigma \bar{X} f(\bar{X})]^2 = \frac{180}{9} - \left(\frac{36}{9}\right)^2 = 4$$

$$2Var(\bar{X}) = 2(4) = 8$$

The mean and variance of the population are:

X	2	2	8	$\Sigma X = 12$
X^2	4	4	64	$\Sigma X^2 = 72$

$$\mu = \frac{\Sigma X}{N} = \frac{12}{3} = 4 \text{ and } \sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{72}{3} - \left(\frac{12}{3}\right)^2 = 8$$

Hence $E(\bar{X}) = \mu = 4$ and $\sigma^2 = 2$ Var $(\bar{X}) = 8$.

Example 11.4.

A population has the values 10, 12, 14, 16, 18 and 20. Draw all possible samples of size 2 without replacement and calculate the sample mean $\bar{X}$ for each sample. Write the sampling distribution of $\bar{X}$. Find the following probabilities:

(i) $\bar{X}$ will be greater than 16. (ii) $\bar{X}$ will differ from μ by less than 3 units.

(iii) Sampling error will be less than 2. (iv) $\bar{X}$ will be equal to μ.

Solution:

All possible samples of size 2 will be equal to ${}^6C_2 = \frac{6!}{2!\,4!} = 15$

The samples, their means and necessary calculations are as under:

Sample No.	Sample Values	Sample Mean ($\bar{X}$)	Sample No.	Sample Values	Sample Mean ($\bar{X}$)
1	10, 12	11	9	12, 20	16
2	10, 14	12	10	14, 16	15
3	10, 16	13	11	14, 18	16
4	10, 18	14	12	14, 20	17
5	10, 20	15	13	16, 18	17
6	12, 14	13	14	16, 20	18
7	12, 16	14	15	18, 20	19
8	12, 18	15			

Sampling Distribution of $\bar{X}$		
$\bar{X}$	f	$f(\bar{X})$
11	1	1/15
12	1	1/15
13	2	2/15
14	2	2/15
15	3	3/15
16	2	2/15
17	2	2/15
18	1	1/15
19	1	1/15
Total	15	1

Population mean $\mu = \frac{10 + 12 + 14 + 16 + 18 + 20}{6} = \frac{90}{6} = 15$

(i) $P(\bar{X} > 16) = \frac{2}{15} + \frac{1}{15} + \frac{1}{15} = \frac{4}{15}$

(ii) $\bar{X}$ will differ from μ by less than 3 units if $\bar{X}$ is greater than 12 and is less than 18.

Thus $P[|\bar{X} - \mu| < 3] = P(12 < \bar{X} < 18) = \frac{2}{15} + \frac{2}{15} + \frac{3}{15} + \frac{2}{15} + \frac{2}{15} = \frac{11}{15}$

(iii) The sampling error will be less than 2 if the random variable $\bar{X}$ is greater than 13 and less than 17. Thus $P(13 < \bar{X} < 17) = P(14 \le \bar{X} \le 16) = P[|S.E.| < 2]$
$= \frac{2}{15} + \frac{3}{15} + \frac{2}{15} = \frac{7}{15}$

(iv) $P(\bar{X} = \mu) = P(\bar{X} = 15) = \frac{3}{15}$

Example 11.5

Certain tubes produced by a company have a mean lifetime of 900 hours and a standard deviation of 100 hours. The company sends out 2000 lots of 100 tubes each.

Compute the mean and standard deviation of the sampling distribution of the sample mean $\bar{X}$ if sampling is done: (i) with replacement (ii) without replacement.

Solution:

Here $N = 2000$, $n = 100$, $\mu = 900$, $\sigma = 100$

(i) Sampling with replacement

$$\mu_{\bar{x}} = \mu = 900 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10$$

(ii) Sampling without replacement

$$\mu_{\bar{x}} = \mu = 900 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{100}{\sqrt{100}}\sqrt{\frac{2000-100}{2000-1}} = 9.75$$

11.8.3 SAMPLING DISTRIBUTION OF s^2 and S^2

Suppose we draw all possible samples of size n from a finite population and calculate the sample variance $s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$ for each sample. The mean of the sampling distribution of s^2 is denoted by $E(s^2)$ or μ_{s^2}. It can be shown that if sampling is with replacement, then $E(s^2) = \mu_{s^2} = \sigma^2$. Thus s^2 is an unbiased estimator of σ^2. The sample variance S^2 is defined as: $S^2 = \frac{\Sigma(X-\bar{X})^2}{n}$. If samples are drawn with replacement, it can be shown that: $E(S^2)\frac{n}{n-1} = \sigma^2$ $[E(S^2) \neq \sigma^2]$.

Thus S^2 is a biased estimator of σ^2. In case of sampling without replacement, we have the following relations:

$$E(s^2)\frac{N-1}{N} = \sigma^2 \quad \text{or} \quad E(s^2) = \left(\frac{N}{N-1}\right)\sigma^2$$

$$E(S^2)\frac{n}{n-1}\cdot\frac{N-1}{N} = \sigma^2 \quad \text{or} \quad E(S^2) = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2$$

Example 11.6

A population consists of three numbers 10, 12, 14. Take all possible samples of size two with replacement from this population. Find the mean and the unbiased variance for each sample. Show that $E(s^2) = \sigma^2$ where $s^2 = \Sigma(X-\bar{X})^2/(n-1)$

Solution:

We have population values 10, 12, 14, population size $N = 3$ and sample size $n = 2$. Thus, the number of possible samples which can be drawn with replacement is $N^n = 3^2 = 9$.

Sample No.	Sample Values	Sample Mean $\bar{X} = \Sigma X/n$	Sample Variance $s^2 = \Sigma(X - \bar{X})^2/(n-1)$
1	10, 10	$\frac{10+10}{2} = 10$	$\frac{(10-10)^2 + (10-10)^2}{2-1} = 0$
2	10, 12	$\frac{10+12}{2} = 11$	$\frac{(10-11)^2 + (12-11)^2}{2-1} = 2$
3	10, 14	$\frac{10+14}{2} = 12$	$\frac{(10-12)^2 + (14-12)^2}{2-1} = 8$
4	12, 10	$\frac{12+10}{2} = 11$	$\frac{(12-11)^2 + (10-11)^2}{2-1} = 2$
5	12, 12	$\frac{12+12}{2} = 12$	$\frac{(12-12)^2 + (12-12)^2}{2-1} = 0$
6	12, 14	$\frac{12+14}{2} = 13$	$\frac{(12-13)^2 + (14-13)^2}{2-1} = 2$
7	14, 10	$\frac{14+10}{2} = 12$	$\frac{(14-12)^2 + (10-12)^2}{2-1} = 8$
8	14, 12	$\frac{14+12}{2} = 13$	$\frac{(14-13)^2 + (12-13)^2}{2-1} = 2$
9	14, 14	$\frac{14+14}{2} = 14$	$\frac{(14-14)^2 + (14-14)^2}{2-1} = 0$

The sampling distribution of the sample variance s^2 and its mean is:

s^2	Tally	f	$f(s^2)$	$s^2 f(s^2)$
0	|||	3	3/9	0
2	||||	4	4/9	8/9
8	||	2	2/9	16/9
Total		9	1	24/9

$$E(s^2) = \Sigma s^2 f(s^2) = \frac{24}{9} = 2.67$$

The variance of the population is:

X	10	12	14	$\Sigma X = 36$
X^2	100	144	196	$\Sigma X^2 = 440$

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{440}{3} - \left(\frac{36}{3}\right)^2 = 2.67$$

Hence $E(s^2) = \sigma^2 = 2.67$

Example 11.7.

A population consists of five values 4, 6, 8, 10, 12. Take all possible samples of size two without replacement from this population and verify that

$$E(S^2) = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2.$$

Solution:

We have population values 4, 6, 8, 10, 12, population size N = 5 and sample size n = 2. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{5}{2} = 10$.

Sample No.	Sample Values	Sample Mean $\bar{X} = \frac{\Sigma X}{n}$	Sample Variance $S^2 = \frac{\Sigma(X-\bar{X})^2}{n}$
1	4, 6	$\frac{4+6}{2} = 5$	$\frac{(4-5)^2+(6-5)^2}{2} = 1$
2	4, 8	$\frac{4+8}{2} = 6$	$\frac{(4-6)^2+(8-6)^2}{2} = 4$
3	4, 10	$\frac{4+10}{2} = 7$	$\frac{(4-7)^2+(10-7)^2}{2} = 9$
4	4, 12	$\frac{4+12}{2} = 8$	$\frac{(4-8)^2+(12-8)^2}{2} = 16$
5	6, 8	$\frac{6+8}{2} = 7$	$\frac{(6-7)^2+(8-7)^2}{2} = 1$
6	6, 10	$\frac{6+10}{2} = 8$	$\frac{(6-8)^2+(10-8)^2}{2} = 4$
7	6, 12	$\frac{6+12}{2} = 9$	$\frac{(6-9)^2+(12-9)^2}{2} = 9$
8	8, 10	$\frac{8+10}{2} = 9$	$\frac{(8-9)^2+(10-9)^2}{2} = 1$
9	8, 12	$\frac{8+12}{2} = 10$	$\frac{(8-10)^2+(12-10)^2}{2} = 4$
10	10, 12	$\frac{10+12}{2} = 11$	$\frac{(10-11)^2+(12-11)^2}{2} = 1$

The sampling distribution of the sample variance S^2 and its mean is:

S^2	f	$f(S^2)$	$S^2 f(S^2)$
1	4	4/10	4/10
4	3	3/10	12/10
9	2	2/10	18/10
16	1	1/10	16/10
Total	10	1	50/10

$$E(S^2) = \mu_{S^2} = \Sigma S^2 f(S^2) = \frac{50}{10} = 5$$

The variance of the population is:

X	4	6	8	10	12	$\Sigma X = 40$
X^2	16	36	64	100	144	$\Sigma X^2 = 360$

$$\sigma^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{360}{5} - \left(\frac{40}{5}\right)^2 = 72 - 64 = 8$$

$$\left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{5}{5-1}\right)\left(\frac{2-1}{2}\right)8 = \left(\frac{5}{4}\right)\left(\frac{8}{2}\right) = \frac{40}{8} = 5$$

Hence $E(S^2) = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = 5$

Example 11.8

A population of 10 numbers has a mean of 100 and a standard deviation of 10. If samples of size 5 are drawn from this population, find the mean of the sampling distribution of variances when sampling is done

(i) with replacement (ii) without replacement.

Solution:

Here $N = 10, \mu = 100, \sigma = 10, \sigma^2 = 100, n = 5$

(i) Sampling with replacement

$$E(S^2) = \mu_{S^2} = \left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{5-1}{5}\right)100 = 80$$

(ii) Sampling without replacement

$$E(S^2) = \mu_{S^2} = \left(\frac{N}{N-1}\right)\left(\frac{n-1}{n}\right)\sigma^2 = \left(\frac{10}{10-1}\right)\left(\frac{5-1}{5}\right)100 = 88.89$$

11.8.4 SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN TWO MEANS

Suppose there is a population with mean μ_1 and variance σ_1^2. Another population has the mean μ_2 and variance σ_2^2. All possible simple random samples of size n_1 are selected from the first population and the sample means $\bar{X}_1$ for each sample are calculated. Similarly, all possible simple random samples of size n_2 are selected from the second population and the sample means $\bar{X}_2$ are calculated. The

difference $(\bar{X}_1 - \bar{X}_2)$ is another random variable and its distribution is called sampling distribution of $\bar{X}_1 - \bar{X}_2$. Some properties of this distribution are:

(i) The mean of the distribution of $\bar{X}_1 - \bar{X}_2$ is equal to the difference $\mu_1 - \mu_2$. Thus

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

Similarly the distribution of $\bar{X}_2 - \bar{X}_1$ has the mean $\mu_{\bar{X}_2 - \bar{X}_1} = \mu_2 - \mu_1$.

If $\mu_1 = \mu_2$, then $E(\bar{X}_1 - \bar{X}_2) = 0$

The above relations are true for any type of population with any sample size, small or large and the samples may be drawn by without replacement or with replacement.

(ii) When samples are selected by without replacement from a finite population, the standard error of $\bar{X}_1 - \bar{X}_2$ has the following relation with σ_1^2 and σ_2^2.

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

When samples are drawn with replacement or they are drawn from infinite populations (N_1 and N_2 are very large), the relation becomes:

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

It may be noted that in practical life, N_1 and N_2 are usually very large and the fractions $\frac{N_1 - n_1}{N_1 - 1}$ and $\frac{N_2 - n_2}{N_2 - 1}$ are almost equal to unity. Thus in the subsequent chapter, we shall frequently use the relation $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(iii) The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is a normal distribution when $n_1 > 30$ and $n_2 > 30$. The sample sizes n_1 and n_2 may be equal or unequal but both should be large in size. The difference $(\bar{X}_1 - \bar{X}_2)$ is a random variable with normal distribution and the standard normal variable Z can be written as

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The distribution of $\bar{X}_1 - \bar{X}_2$ has the t-distribution when both n_1 and n_2 are small in size.

Example 11.9.

Draw all possible random samples of size $n_1 = 2$ without replacement from the finite population 2, 2, 6. Similarly, draw all possible random samples of size $n_2 = 2$ without replacement from the population 1, 1, 2, 4.

(i) Find the possible differences between the sample means of the two populations.

(ii) Construct the sampling distribution of $\bar{X}_1 - \bar{X}_2$ and compute its mean and variance.

(iii) Verify that: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ and $Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$

Solution:

Population I: 2, 2, 6

Population size $N_1 = 3$

Sample size $n_1 = 2$

The number of possible samples which can be drawn without replacement

$= \binom{N_1}{n_1} = \binom{3}{2} = 3$

Population II: 1, 1, 2, 4

Population size $N_2 = 4$

Sample size $n_2 = 2$

The number of possible samples which can be drawn without replacement

$= \binom{N_2}{n_2} = \binom{4}{2} = 6$

From Population I			From Population II		
Sample No.	Sample Values	Sample Mean ($\bar{X}_1$)	Sample No.	Sample Values	Sample Mean ($\bar{X}_2$)
1	2, 2	2	1	1, 1	1.0
2	2, 6	4	2	1, 2	1.5
3	2, 6	4	3	1, 4	2.5
			4	1, 2	1.5
			5	1, 4	2.5
			6	2, 4	3.0

(i) The 18 possible differences $\bar{X}_1 - \bar{X}_2$ are shown in the following table.

	$\bar{X}_1$		
$\bar{X}_2$	2	4	4
1.0	1.0	3.0	3.0
1.5	0.5	2.5	2.5
2.5	– 0.5	1.5	1.5
1.5	0.5	2.5	2.5
2.5	– 0.5	1.5	1.5
3.0	– 1.0	1.0	1.0

(ii) The sampling distribution of differences between sample means $\bar{X}_1 - \bar{X}_2$ and its mean and variance are computed below:

$\bar{X}_1 - \bar{X}_2 = d$	f	f(d)	d f(d)	d^2 f(d)
– 1.0	1	1/18	– 1/18	1.0/18
– 0.5	2	2/18	– 1/18	0.5/18
0.5	2	2/18	1/18	0.5/18
1.0	3	3/18	3/18	3.0/18
1.5	4	4/18	6/18	9.0/18
2.5	4	4/18	10/18	25.0/18
3.0	2	2/18	6/18	18.0/18
Total	18	1	24/18	57/18

$$E(\bar{X}_1 - \bar{X}_2) = E(d) = \Sigma d\, f(d) = \frac{24}{18} = \frac{4}{3}$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(d) = \Sigma d^2 f(d) - [\Sigma d\, f(d)]^2 = \frac{57}{18} - \left(\frac{4}{3}\right)^2 = \frac{57 - 32}{18} = \frac{25}{18}$$

(iii) The mean and variance of the first population are:

X_1	2	2	6	$\Sigma X_1 = 10$
X_1^2	4	4	36	$\Sigma X_1^2 = 44$

$$\mu_1 = \frac{\Sigma X_1}{N_1} = \frac{10}{3} \text{ and } \sigma_1^2 = \frac{\Sigma X_1^2}{N_1} - \left(\frac{\Sigma X_1}{N_1}\right)^2 = \frac{44}{3} - \left(\frac{10}{3}\right)^2 = \frac{44}{3} - \frac{100}{9} = \frac{132 - 100}{9} = \frac{32}{9}$$

The mean and variance of the second population are;

X_2	1	1	2	4	$\Sigma X_2 = 8$
X_2^2	1	1	4	16	$\Sigma X_2^2 = 22$

$$\mu_2 = \frac{\Sigma X_2}{N_2} = \frac{8}{4} = 2 \text{ and } \sigma_2^2 = \frac{\Sigma X_2^2}{N_2} - \left(\frac{\Sigma X_2}{N_2}\right)^2 = \frac{22}{4} - \left(\frac{8}{4}\right)^2 = \frac{22}{4} - 4 = \frac{22-16}{4} = \frac{3}{2}$$

$$\mu_1 - \mu_2 = \frac{10}{3} - 2 = \frac{10-6}{3} = \frac{4}{3}$$

$$\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right) = \frac{32}{18}\left(\frac{3-2}{3-1}\right) + \frac{3}{4}\left(\frac{4-2}{4-1}\right) = \frac{16}{18} + \frac{1}{2} = \frac{16+9}{18} = \frac{25}{18}$$

Hence $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = \frac{4}{3}$ and $Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right) = \frac{25}{18}$

Example 11.10

Given $N_1 = 800$, $N_2 = 600$, $n_1 = 200$, $n_2 = 124$, $\mu_1 = 1800$, $\mu_2 = 1600$, $\sigma_1 = 200$ and $\sigma_2 = 124$. Compute the mean and standard error of the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ if sampling is done (i) with replacement (ii) without replacement.

Solution:

(i) Sampling with replacement

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 1800 - 1600 = 200$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{200} + \frac{(124)^2}{124}} = 18$$

(ii) Sampling without replacement

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 1800 - 1600 = 200$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)} = \sqrt{\frac{(200)^2}{200}\left(\frac{800-200}{800-1}\right) + \frac{(124)^2}{124}\left(\frac{600-124}{600-1}\right)}$$

$$= 15.77$$

11.8.5 PROPORTION

What is a proportion? Suppose there are 1000 students in a school out of which 600 are male and 400 are female. The ratio of 600 to the total is called the proportion of males and is denoted by p. Thus proportion of males $= p = \frac{600}{1000} = 0.6$

and proportion of females $= q = \frac{400}{1000} = 0.4$

Let us denote male by success and female by a failure. If the male students are assigned the number 1 and females are assigned the number 0, then the population contains 600 ones and 400 zeros. This can be written as below in the form of a distribution called the Bernoulli distribution. Let us calculate the mean of this distribution.

Random Variable (X)	f	f(X)	X f(X)
0	400	400/1000 = 0.4	0
1	600	600/1000 = 0.6	0.6
Total	1000	1	0.6

$E(X) = \text{Mean} = \Sigma X\, f(X) = 0.6$

Thus the *proportion* p of the population called the binomial population is equal to the mean of the population containing 0's and 1's.

11.8.6 SAMPLING DISTRIBUTION OF PROPORTION

Suppose there is a finite population in which the proportion of successes is p and the proportion of failures is q. Suppose we draw all possible samples of size n from the population and calculate the sample proportion $\hat{p}$ for each sample. The sampling distribution of $\hat{p}$ has the following properties.

(i) The mean of the sampling distribution of $\hat{p}$ is equal to the population proportion p. Thus $E(\hat{p}) = \mu_{\hat{p}} = p$

This relation is true in sampling with replacement and without replacement for any sample size.

(ii) The standard error of $\hat{p}$ is related to the population parameters p and q through the equations:

$$S.E.(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\left(\frac{N-n}{N-1}\right)} \qquad \text{(True for sampling without replacement)}$$

and $$S.E.(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \qquad \left(\begin{array}{c}\text{True for sampling with replacement}\\ \text{or when N is very large}\end{array}\right)$$

(iii) The shape of the distribution of $\hat{p}$ is normal when $n > 30$. The value of Z can be calculated from $\hat{p}$, where $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$.

Warning: It is important to note that when n is small, the distribution of $\hat{p}$ is not the t-distribution.

Example 11.11

A population consists of five numbers 2, 5, 6, 7, 9. Take all possible samples of size 3 from this population without replacement and compute the proportion of odd numbers for each sample. Verify that: (i) $\mu_{\hat{p}} = p$ (ii) $\sigma^2_{\hat{p}} = \dfrac{pq}{n}\left(\dfrac{N-n}{N-1}\right)$

Solution:

We have population values 2, 5, 6, 7, 9, population size N = 5 and sample size n = 3. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{5}{3} = 10$. Let $\hat{p}$ represent the proportion of odd numbers in the sample.

Sample No.	Sample Values	Sample Proportion ($\hat{p}$)	Sample No.	Sample Values	Sample Proportion ($\hat{p}$)
1	2, 5, 6	1/3	6	2, 7, 9	2/3
2	2, 5, 7	2/3	7	5, 6, 7	2/3
3	2, 5, 9	2/3	8	5, 6, 9	2/3
4	2, 6, 7	1/3	9	5, 7, 9	3/3
5	2, 6, 9	1/3	10	6, 7, 9	2/3

The sampling distribution of the sample proportion $\hat{p}$ and its mean and variance are:

$\hat{p}$	Tally	f	$f(\hat{p})$	$\hat{p}\, f(\hat{p})$	$\hat{p}^2 f(\hat{p})$
1/3	|||	3	3/10	3/30	3/90
2/3	++++ |	6	6/10	12/30	24/90
3/3	|	1	1/10	3/30	9/90
Total		10	1	18/30	36/90

$$\mu_{\hat{p}} = \Sigma \hat{p}\, f(\hat{p}) = \frac{18}{30} = 0.6$$

$$\sigma^2_{\hat{p}} = \Sigma \hat{p}^2 f(\hat{p}) - [\Sigma \hat{p}\, f(\hat{p})]^2 = \frac{36}{90} - \left(\frac{18}{30}\right)^2 = 0.40 - 0.36 = 0.04$$

Population proportion $p = \frac{X}{N} = \frac{3}{5} = 0.6$, $q = 1 - p = 0.4$

where X represents the number of odd digits in the population.

$$\frac{pq}{n}\left(\frac{N-n}{N-1}\right) = \frac{(0.6)(0.4)}{3}\left(\frac{5-3}{5-1}\right) = 0.04$$

Hence (i) $\mu_{\hat{p}} = p = 0.6$ (ii) $\sigma^2_{\hat{p}} = \frac{pq}{n}\left(\frac{N-n}{N-1}\right) = 0.04$

Example 11.12.

A finite population contains 4 smokers denoted by S_1, S_2, S_3 and S_4 and 2 non-smokers denoted by N_1 and N_2. Draw all possible random samples of size 2 without replacement from the population and calculate the proportion of smokers $\hat{p}$ in each

ample. Write the probability distribution (sampling distribution) of $\hat{p}$ and find the following probabilities:

(i) $\hat{p}$ is more than p (ii) $\hat{p}$ is equal to p (iii) $\hat{p} = \frac{1}{2}$ (iv) that both are smokers.

Solution:

We have population values $S_1, S_2, S_3, S_4, N_1, N_2$, population size N = 6 and sample size n = 2. Thus, the number of possible samples which can be drawn without replacement is $\binom{N}{n} = \binom{6}{2} = 15$.

Sample No.	Sample Values	Sample proportion ($\hat{p}$)	Sample No.	Sample Values	Sample proportion ($\hat{p}$)
1	S_1, S_2	2/2	9	S_2, N_2	1/2
2	S_1, S_3	2/2	10	S_3, S_4	2/2
3	S_1, S_4	2/2	11	S_3, N_1	1/2
4	S_1, N_1	1/2	12	S_3, N_2	1/2
5	S_1, N_2	1/2	13	S_4, N_1	1/2
6	S_2, S_3	2/2	14	S_4, N_2	1/2
7	S_2, S_4	2/2	15	N_1, N_2	0
8	S_2, N_1	1/2			

The sampling distribution of the sample proportion $\hat{p}$ is:

$\hat{p}$	f	$f(\hat{p})$
0	1	1/15
1/2	8	8/15
2/2	6	6/15
Total	15	1

Population proportion $p = \frac{4}{6} = \frac{2}{3}$

(i) $P(\hat{p} > p) = \frac{6}{15}$ (ii) $P(\hat{p} = p) = 0$

(iii) $P\left(\hat{p} = \frac{1}{2}\right) = \frac{8}{15}$ (iv) P(both are smokers) $= \frac{6}{15}$

Example 11.13

If samples of n = 200 observations are to be drawn from a large population N = 2500 in which the population proportion is 20 %. Determine the expected mean and standard deviation of the sampling distribution of proportions when sampling is done (i) with replacement (ii) without replacement.

Solution:

Here $N = 2500$, $n = 200$, $p = 0.20$, $q = 1 - p = 1 - 0.20 = 0.80$

(i) Sampling with replacement

$$E(\hat{p}) = p = 0.20 \text{ and } S.E.(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{200}} = 0.0283$$

(ii) Sampling without replacement

$$E(\hat{p}) = p = 0.20 \text{ and } S.E.(\hat{p}) = \sqrt{\frac{pq}{n}\left(\frac{N-n}{N-1}\right)} = \sqrt{\frac{(0.20)(0.80)}{200}\left(\frac{2500-200}{2500-1}\right)}$$
$$= 0.0271$$

11.8.7 SAMPLING DISTRIBUTION OF DIFFERENCE BETWEEN $\hat{p}_1$ and $\hat{p}_2$

Suppose there are two populations with proportions p_1 and p_2 and all possible simple random samples of size n_1 and n_2 are selected from the populations respectively. The sample proportions calculated from the samples are $\hat{p}_1$ and $\hat{p}_2$. The difference $\hat{p}_1 - \hat{p}_2$ is a random variable and its distribution is called the sampling distribution of $\hat{p}_1 - \hat{p}_2$. The properties of this distribution are:

(i) The mean of the distribution of $\hat{p}_1 - \hat{p}_2$ is equal to the difference between p_1 and p_2. Thus $\mu_{\hat{p}_1 - \hat{p}_2} = E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

This relation is true for any sample size and for sampling with and without replacement.

(ii) The standard error of the distribution of $(\hat{p}_1 - \hat{p}_2)$ has the following relation with population parameters

$$S.E.(\hat{p}_1 - \hat{p}_2) = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1}\left(\frac{N_1-n_1}{N_1-1}\right) + \frac{p_2q_2}{n_2}\left(\frac{N_2-n_2}{N_2-1}\right)}$$

(True for sampling without replacement)

$$\text{and } S.E.(\hat{p}_1 - \hat{p}_2) = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

(True for sampling with replacement or when N is very large)

(iii) The distribution of $\hat{p}_1 - \hat{p}_2$ has the normal distribution when both n_1 and n_2 are large in size, when n_1 and n_2 are small in size, the distribution of $\hat{p}_1 - \hat{p}_2$ does not form any standard distribution. The random difference $(\hat{p}_1 - \hat{p}_2)$ can be transformed into standard normal variable Z where $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1q_1}{n_1} + \dfrac{p_2q_2}{n_2}}}$

Example 11.14

Given the data: $N_1 = 6$, $n_1 = 3$, $X_1 = 3$, $N_2 = 5$, $n_2 = 2$, $X_2 = 2$.

Find $E(\hat{p}_1 - \hat{p}_2)$ and $Var(\hat{p}_1 - \hat{p}_2)$ if sampling is done

(i) with replacement (ii) without replacement

Solution:

Here $N_1 = 6,\ n_1 = 3,\ X_1 = 3,\ p_1 = \frac{X_1}{N_1} = \frac{3}{6} = 0.5,\ q_1 = 1 - p_1 = 0.5$

$N_2 = 5,\ n_2 = 2,\ X_2 = 2,\ p_2 = \frac{X_2}{N_2} = \frac{2}{5} = 0.4,\ q_2 = 1 - p_2 = 0.6$

(i) Sampling with replacement

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0.5 - 0.4 = 0.1$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} = \frac{(0.5)(0.5)}{3} + \frac{(0.4)(0.6)}{2}$$

$$= 0.0833 + 0.12 = 0.2033$$

(ii) Sampling without replacement

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0.5 - 0.4 = 0.1$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{p_2q_2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)$$

$$= \frac{(0.5)(0.5)}{3}\left(\frac{6-3}{6-1}\right) + \frac{(0.4)(0.6)}{2}\left(\frac{5-2}{5-1}\right)$$

$$= 0.05 + 0.09 = 0.14$$

SHORT DEFINITIONS

Population whole of aggregate of items is called.

A population is the total set of measurements of interest in a particular problem.

or

The population is a set of data that characterizes some phenomenon.

Finite Population

If a population has finite number of elements, it is called as finite population. For example human population, number of chairs in a college.

Infinite Population

If a population has infinite number of elements, it is called as infinite population. For example number of points on line, number of stars in the sky.

Target Population

A population about which we want to get some information is called target population.

Sampled Population

A population from which a sample is drawn is called sampled population.

Sample

A sample is a subset of data selected from a population.

or

A sample is a subset of the population that contains measurements obtained by an experiment.

Random Sample

A sample obtained by random sampling is called a random sample.

or

If a sample is selected from such a population whose sampling units have known probability that may be equal or unequal, the sample is said to be a random sample.

Sampling

Sampling is the process of drawing sample from the population.

Random Sampling

Any procedure for selecting members (samples) from a group on the basis of chance or luck is called a random sampling.

or

A method of selecting samples so that each sample of a given size in a population has an equal or unequal chance of being selected.

Sampling Units

Sampling units are nonoverlapping collections of elements from the population.

or

The basic elements that constitutes a population are known as sampling units.

Simple Random Sample

A simple random sample is one in which every item from a population has the same chance of selection as any other item.

or

A sample selected in such a manner that each possible sample of a specified size has an equal chance of being selected.