Chapter ... 7

SAMPLING AND HYPOTHESIS

♦ LEARNING OBJECTIVES ♦

After completing this chapter, students will be able to:

- Know the importance of sampling in research.
- Select best method for sampling.
- Select proper sample size for analysis.
- Know the hypothesis testing and errors in sampling.

7.1 INTRODUCTION

Sampling is one of important part of all scientific research and it is well developed in biological, physical, chemical and social sciences. Sampling is a process of obtaining information about entire population by examining only part of it. A sample is a part of the population which is studied in order to make inferences or conclusion about the whole population.

The total or the aggregate of all units that confirm to some designated set of specifications is called the universe or population. A population may be a group of people, students, workers, customers, voters, houses etc. The specific nature of the population depends on the purpose of investigation. Each entity (student, person, family) from the population about which information is collected is called a sampling element. Sampling frame is the complete list of all elements or units from which the sample is collected e.g. list of patients, list of students, etc.

Sampling is based on the law of statistical regularity and the law of inertia of large numbers. The first law states that the small numbers of items (sample) are picked from a large number of items (universe or population), the sample will tend to possess the same characteristics as that of whole group of items. The second law of inertia states that the sample should be large enough to represent truly the entire population of the universe.

While collecting primary data, the information may be obtained either by census method or sample method. In census method, every functional unit of the population or group is studied. This method will provide more reliable and accurate results but, it requires more time and money compared to sample method. In sample method, only a small part of whole population is to be studied and the conclusions are applicable to the whole population.

If the sample results are to be worthwhile, it is necessary that a sample possesses the following essentials. The characteristics of a good sample are:

- An ideal sample must be the representative of the population corresponding to its properties. It should not lack in any characteristic of the population.
- It must be unbiased and must be obtained by a probability process or random method.
- It must make the research work more feasible and has the practicability for research situation.
- It must yield an accurate result and does not involve errors. The probability of error can be estimated.
- Sample must be adequate in order to ensure reliability. A sample having 10% of the whole population is generally adequate.
- Sample must be comprehensive in nature. It is a quality of sample which is controlled by specific purpose of the investigation.
- Sample units must be chosen in a systematic and objective manner.

The important advantages of sampling are as follows:

- Sampling techniques brings about cost control of research project. It is less expensive to collect data from a portion of the population than from the entire population.
- It is possible for the researcher to collect more elaborate information from the few sample units than from the large population. Multiple approaches can be applied to a sample for an intensive analysis.
- Sampling technique has a greater adaptability, precision and accuracy in the observation.
- Sampling techniques is a scientific method because the conclusion derived from the study of certain units can be verified from other units. The deviations can be seen from average.
- Sampling technique helps the researcher to complete the project in short duration. Researcher can save the time and utilize for other research studies.
- The examination of few sample units is sufficient. Hence, it has wide applicability in most situations or research studies.

The limitations of sampling techniques are as follows:

- Sampling is not feasible in the situation where precise knowledge about each unit of population is needed.
- The sampling technique can be successful only if the researchers have the basic and specialized knowledge, understanding and experience.

- Sampling does not give the accuracy available from census. It is less accurate than the census method.
- The accuracy and reliability of sample data is affected by sampling errors and data collection errors.
- If the units in the field of survey are liable to change then the sampling technique will be very hazardous. It is not scientific to extend the conclusions derived from one set of sample to other sets which are dissimilar.
- If sampling is done on partial basis, then it can not expect the results true, neutral and impartial.
- Selection of representative sampling from social and political units may differ. If sample will not give correct representation, then its finding will be less reliable and authentic.

7.2 SAMPLING DESIGN

A sampling design is a definite plan for obtaining a sample from the sampling frame. Population frame or sampling frame is the listing of all items in the population with proper identification. The list of names and addresses of workers are called sampling frame if we want to find out the capital invested and number of workers working in chemical industries in Pimpri. Sampling design is a technique of selecting items for the sample. There are many sample designs from which a researcher can choose a specific and most appropriate design. While developing a sampling design, the researcher must consider the following parameters.

Universe or Population:

The first step in developing any sample design is to clearly define the set of objects or universe. The universe can be finite or infinite. The population contains fixed number of items so that it is possible to list it in its totality, it is called finite population. e.g. total population of village, workers working in small unit, etc. The population where we cannot make out the total number of items is called an infinite population. e.g. number of stars in the sky, observers of cricket match, etc.

Sampling Unit:

A final decision has to be taken concerning a sampling unit before selecting a sample. Sampling unit may be geographical units (state, district, town etc.), social units (schools, family, etc.) or construction units (flat, house etc.). The researcher has to select one or more such units for his study.

Sampling Frame or Source List:

The sampling frame is a list of elements in the population from which the sample is actually drawn. It contains the names of all items of a universe. Researcher has to prepare comprehensive, correct and reliable list if source list is not available. The source list may be polling list, telephone directory, maps, publications of the government, chamber of industries, etc.

Sample Size:

Size of sample refers to number of items to be selected from the universe to constitute a sample. The size of sample should be optimum, neither be excessively large, nor too small. An optimum sample is one which fulfills the requirements of efficiency, representativeness and reliability. Researcher must consider the different parameters before deciding the sample size such as size population, homogeneity or heterogeneity of the universe, nature of study, type of sampling, standard of accuracy, cost, etc.

Sampling Method:

Sampling design is based on method of sampling such as probability or random sampling and non-probability or purposive sampling. The choice of method depends upon the nature of problem, standard of accuracy, availability of sampling frame, time, cost, and other constraints.

Budgetary Constraint:

The researcher has to keep the cost in mind while preparing a sample design. Budget for sampling have a major impact not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.

Sampling Plan:

It involves specifying how the sampling plan is going to be put into operation. It consists of instructions for actual implementation of the sampling plan and instructions to the interviewers regarding how to select a random sample.

7.3 METHODS OF SAMPLING

There are basically two types of sampling such as probability sampling and non-probability sampling (Fig. 7.1). Probability sampling is one in which every unit of the population has an equal probability of being selected for the sample. Non-probability sampling makes no claim for representativeness, as every unit does not get the chance of being selected. Probability sampling is based on the concept of random selection, whereas non-probability sampling is 'non-random' sampling.

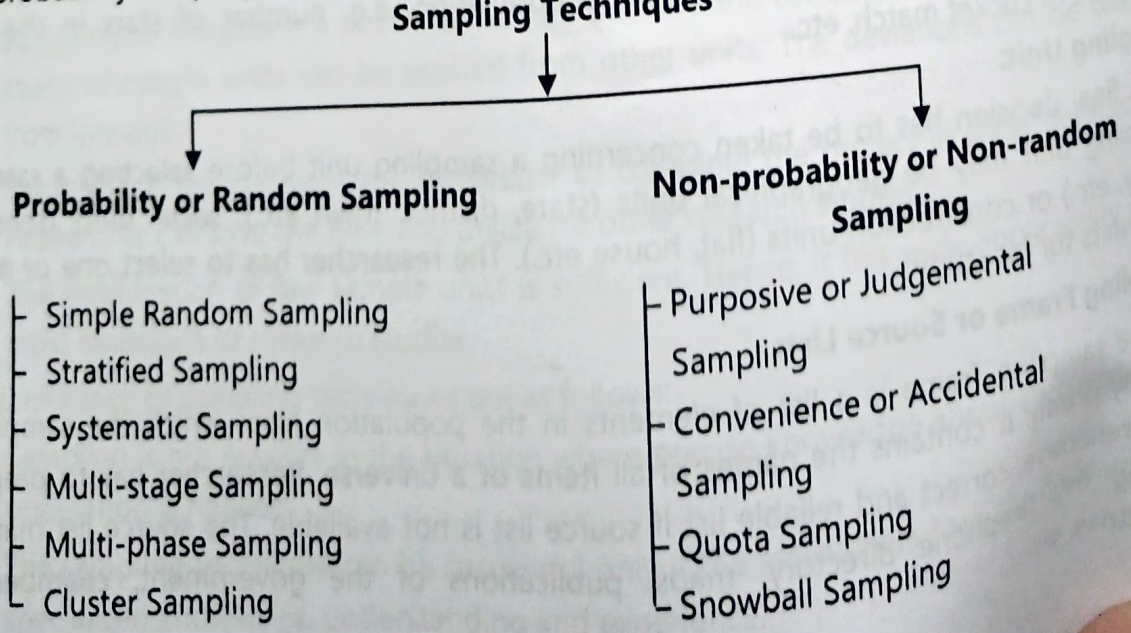


Fig. 7.1: Types of Sampling Techniques

7.3.1 Probability or Random Sampling

Probability sampling methods are those in which every item in the population has a known chance or probability of being chosen for the sample. This method is a primary method for selecting large, representative samples for social science and business researchers. This sampling design, the population must be clearly defined and list of target population must be available. It is lottery method in which individual units are picked up from the whole group by some mechanical process. Probability sampling is also known as chance sampling or random sampling. This sampling technique provides estimates which can be measured precisely and are inherently unbiased. It is possible to evaluate the relative efficiency of various sample designs in probability sampling. It requires a high degree of skill level, expertise, and a lot of time to plan and determine a probability sample. The cost involved in probability sampling is higher as compared to non-probability sampling. There are a number of methods used for probability sampling such as simple random sampling, systematic sampling, stratified sampling, multiple sampling, multi-stage sampling, cluster sampling etc.

Simple Random Sampling:

This method is most common technique for selecting a probability sample. Random sampling method of selection assures each element of the population has an equal and independent chance of being included in the sample. Items in the sample are selected completely independent of each other. The selection of one unit does not influence the selection of other unit. There is no chance of bias of any kind in the selection of sample units. In simple random sampling method, the investigator can keep himself away from prejudices, bias and other elements of subjectivity.

It has the following advantages:

- This method is free from bias, prejudices and free from personal error.
- It is quite simple method, follows mathematical procedures and there is no need of advance knowledge of population.
- It is less expensive and easily practicable procedure if the population is not large.
- It is representative of the universe, each unit has equal chance of being selected.
- Assessment of the accuracy of the results is possible by simple error estimation.

The disadvantages of this method are:

- The representativeness of the sample cannot be ensured by this method.
- Random sampling method is not useful for units which are heterogeneous in nature.
- It requires study of a whole population and individual characteristics of each item are difficult to study.

A random sample can be selected by lottery method, Tippet's number method, grid system and selection from sequential method.

(a) Lottery Method: In this method, a lottery is drawn by writing the numbers or the name of various units and putting them in a box. Instead of pieces of paper, plastic discs or coins can also be used. They are thoroughly mixed and certain numbers are picked up from the box as a sample. This method is suitable for drawing a small number of samples from a small universe.

(b) Tippet's Number Method: Tippet has constructed a four digits random list of 10,400 institutions. These numbers are the results of combining 41,600 population statistics reports. The random numbers are quite suitable if the size of the population is large. A researcher wants to select 500 supervisors from a sugar cane industry out of total 10,000. A supervisor from the list should be numbered and then random numbers should be used to make the selection. For framing the random numbers from 1 to 10,000, open a page of Tippett's numbers and then first 500 number's are selected. The selection of 500 supervisors will be included in the sample.

(c) Grid System: The map of the entire area is prepared and then a screen with a square is placed upon the map. The squares are selected by random process. The areas falling within the selected squares are taken as samples.

(d) Selection from Sequential Method: In this method, units are arranged serially according to some particular orders such as alphabetical, numerical or geographical sequence. Then out of the list, every 5^{th}, 10^{th} or any other number may be taken up. For example, if we want to select 15 persons, then start from 15^{th} and select 15^{th}, 30^{th}, 45^{th}, 60^{th}, 75^{th} and so on.

Stratified Sampling:

This method of selecting samples is a mixture of the deliberate and random sampling technique. In this method, the units or data in a domain are split into various classes or groups (strata) on the basis of their characteristics and then certain items are selected from these classes by the random sampling technique. This method is also called as mixed technique of sampling.

According to this method, first the population is divided into homogeneous groups or strata and sample is taken from each stratum. The basis of classification may be sex, age, class, religion, income, education, land holding etc. which is known as the stratification factor. The stratification factor is selected in correlation with the problem under study, which becomes a logical basis of stratification.

Each stratum in the universe should be large enough in the size so that selection of items may be done on random basis.

The advantages of this method are:

- It is good representative of the population and it has greater accuracy.
- It is more precise and to a great extent avoid bias.

- Sample size may be small in this method that saves time and cost of data collection.
- This method is mostly used in geographical purpose.

The disadvantages of stratified sampling method are:

- It requires the knowledge of the traits of the population.
- Stratification or classification sometimes becomes difficult and assignment of each unit of the population to a particular stratum can be even more difficult.
- Attainment of proportion becomes particularly difficult when there is wide variance is size of different strata.
- The result sometimes has to be weighted according to the size of strata, which is difficult task.

There are two types of stratified sampling method:

(a) Proportionate Stratified Sampling:

In this method, selection from each sampling unit of a sample that is proportionate to the size of the unit or stratum. It gives more representativeness with respect to variables used as the basis of classifying categories and increased chances of being able to make comparisons between strata.

(b) Disproportionate Stratified Sampling:

In this method, an equal number of cases are taken from each stratum without any consideration to the size of strata in proportion to universe. This method of sampling is more effective for comparing strata which have different error possibilities.

(c) Systematic Sampling:

In this systematic sampling, only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available. A voter list, a telephone directory or a card index system mainly satisfy the conditions of systematic sampling. The researcher first has to arrange the units of the universe on same basis such as alphabetical, chronological, geographical, numerical, ascending order of capital employed or labour employed etc.

The first item is selected at random by lottery method. Subsequent items are selected by taking every K^{th} items from the list.

$$K = \frac{N}{n}$$

where, K is sampling interval, N is universe size and n is sample size.

For example, let the size of the population 500 and the sample size 50 that indicates 50 samples are to be drawn from the population 500.

The sample interval is $\frac{500}{50} = 10$.

Suppose, the first sample selected at random from the first interval is 7, then subsequent samples are 17, 27, 37, 47, 57, 67 and so on.

The advantages of systematic sampling are:

- This is simple method of selecting a sample and convenient to adopt.
- The time and work involved in sampling by this method are relatively less.
- Samples may be comprehensive and representative of population.

The disadvantages of systematic sampling are:

- This method is not suitable for heterogeneous population.
- It is not probability random sampling, since each element has no chance of being selected.
- Any hidden periodicity in the list will adversely affect the representativeness of the sample.
- Knowledge of population and information of each individual is essential.

Multi-stage Sampling:

In this method, the selection of sample is made in different stages. The original units into which the universe is divided are primary units. Each primary unit that falls into the sample is subdivided into secondary units in preparation for the second stage sampling. In three stage sampling, there may be primary, secondary and tertiary units. For example, for studying the panchayat system in villages, our country is divided into zones (North, South, East and West). One state is selected from each zone and then one district is selected from each state. For each district one block is selected and then three villages are selected from each block. This sampling system helps in comparing the functioning of panchayats in different parts of India. Sampling in each stage may be random but it can also be deliberate or purposive. The individuals are selected from different stages for constituting the multi-stage sampling. The variability of estimates yielded by multi-stage sampling may be greater than that of estimates yielded by simple random sampling for equal size. The variability of estimates is depended on the composition of primary units.

The advantages of multi-stage sampling are:

- A complete listing of population or universe is not required. Sampling lists, identification and numbering are required only for sampling units which are selected as sample.
- It is a good representative of the population and easy to administer.
- A large number of units can be sampled for a given cost under multistage sampling because of sequential clustering.

The disadvantages of multi-stage sampling are:

- It is a more difficult and complex method of sampling compared to other methods.
- It involves more errors compared to simple random sampling or systematic sampling for similar sample size.

Multi-phase Sampling:

Multi-phase sampling in which some information is collected from the whole sample and additional information is collected either at the same time or later from a sub-sample of the full sample. It increases the precision of sub-sample results. The cost of first phase sample is cheaper than the second phase. The first phase information is collected from some records or by mail while the second phase information is obtained by personal interviews. In leprosy survey, test like Lepromin is performed for all samples in the first phase. Those who are tested positive in Lepromin test are screened in the second phase of culture of organism. In this method, the information collected at each phase helps the researcher to select a more relevant sample.

Cluster Sampling:

In cluster sampling, population may be divided into groups called clusters and drawing a sample of clusters to represent the population. A cluster may consist of either the primary sample units or elementary secondary sample units. In a selected cluster either all the sample units are selected or a few of them are chosen by other sampling method.

Many national surveys are based on cluster sampling where villages, schools, colonies and corporation areas are considered as a cluster. A plan to make a selection of clusters within clusters is called multi-stage cluster sampling. From each of the selected clusters, select the sample of elements instead of including the entire cluster. This procedure of sampling is called as two-stage sampling or sub-sampling. If clusters are geographic units then, it is known as area sampling. In the area sampling, the basis of selection is the map rather than a list of the population. Maximum accuracy in cluster sampling is obtained if differences or variability within a cluster is large (heterogeneity) or variability among or between cluster is small (homogeneity).

The advantages of cluster sampling are:

- Cluster sampling is an simple, economical and good representative of the population.
- It is commonly used method when no other satisfactory sampling frame for the whole population exists.
- It is much easier to apply this sampling method when large populations or geographical area is studied.
- It is flexible method and characteristics of clusters can be estimated.

The disadvantages of cluster sampling are:

- Each cluster is not of equal size in selection of one district from one state or one village from one block. The district or the village can be small, intermediate or large-sized.
- It may homogeneity in one cluster but heterogeneity in other.
- Cluster sampling is not free from error and it is not comprehensive.

7.3.2 Non-probability or Non-random Sampling

Non-probability sampling or non-parametric sampling methods are those which do not provide every component in the population or universe with an equal chance of being included in the sample. It is the process of sampling without use of randomization. This sampling technique does not follow the rules of probability theory as well as do not claim representativeness. In this sampling, personal element has a great chance of entering into the selection of the sample. The researcher may select a sample which may get results favourable to him. There is always the bias entering into the sampling methods and hence, no assurance that every element has specifiable chance of being selected. Sampling error in non-probability sampling cannot be estimated. There are four non-random designs, which are commonly used in qualitative and quantitative research such as purposive sampling, convenience sampling, quota sampling and snowball sampling.

(a) Purposive or Judgmental Sampling:

This sampling method is also known as deliberate sampling. In this sampling, the choice of sample items depends primarily on the judgment of the investigator. The investigator includes those elements in the sample which he or she likes in the universe with regard to the characteristics of research topic. This technique also depends on the sampling design and purpose of representativeness. Hence, it is also called purposive sampling.

In this method, the researcher has complete freedom to choose his sample according to this wishes and desires. The researcher selects some elements from the whole data as a sample. This is simple method for selecting the samples and is useful in cases where the entire data is homogeneous and the researcher has knowledge of the various aspects.

The advantages of purposive sampling are:

- This technique of sampling is simple and economical.
- It is a practical method when randomization is not possible. It is commonly used in solving many types of economic and business problems.
- Knowledge of the researcher can be best used in this method of sampling.

The disadvantages of purposive sampling are:

- This sampling is not free from error and it includes uncontrolled variation.
- Researcher can select sample as per his wish and desire. Hence, sample may be biased.
- The investigator is unable to understand the whole group and knowledge of the whole group or population is necessary.

(b) Convenience or Accidental Sampling:

This sampling is also known as haphazard sampling or incidental sampling. In this sampling, the investigator studies all those persons who are most conveniently available or who accidentally come in his contact during investigation. This method of sampling is most

common among market research and newspaper reporters. For example, the investigator engaged in the study of college or university students might visit the canteen, library, play-grounds, departments and interview certain number of students. During election period, media personnel often present man-on-the-street interviews that are presumed to reflect public opinion. Convenience sampling is commonly used when universe is not well defined or sampling unit is not clear or complete list of the source is not available.

The advantages of convenience sampling are:

- This sampling is quick, easy and economical method.
- This sampling is frequently used in behavioural sciences and exploratory research.

The disadvantages of convenience sampling are:

- This sampling is not free from error and parametric statistics cannot be used.
- It is not a representative of the population and it may be a very biased sample.

(c) Quota Sampling:

This sampling method is a non-random form of stratified sampling method. The universe is divided into strata on the basis of certain characteristics and then the quota is fixed for each stratum in proportion to its size. It is most commonly used method of sampling in market surveys, opinion polls and municipal surveys.

Quota can be fixed according to their proportion in the entire population. For studying the attitudes of persons towards use of loudspeakers in religions places with 1000 males and 500 females belonging to different religions, quota can be fixed in the ratio of one female for every two males. Quota may be fixed on the basis of number of persons in each of the three religious groups such as Hindu, Muslim or others.

The advantages of quota sampling are:

- This is simple method of sampling and most frequently used in social surveys.
- It is less costly method than other techniques. There is no need of any information of sampling frame, total number of elements, their location or other information about the sampling population.
- It is practical method as well as completed in a very short period of time.

The disadvantages of quota sampling are:

- It is not a truly representative of the total sampling population.
- It is not possible to estimate the sampling error and it has interviewer's bias in the selection.
- It has the influence of regional geographical and social factors.

(d) Snowball Sampling:

This sampling is the process of selecting a sample using networks. In snowball sampling, the investigator starts the research with the few respondents who are known to him. These

respondents give other names who meet the criteria of research, who in turn give more new names. This process of sample selection is continued until adequate number of persons is interviewed. In this manner, the investigator accumulates more and more respondents. For example, if the investigator is able to find a few bonded advertisers willing to talk, he might ask them for details of other advertisers who might also be willing to interact.

Snowball sampling method is useful if you unknown about the group or organization under study. This sampling is useful for studying communication patterns or diffusion of knowledge within a group. Main advantages of this sampling method are less cost and less sample size.

The disadvantages of snowball sampling are:

- It is difficult to use this method when the sample becomes fairly large.
- The choice of the entire sample depends upon the choice of individuals at the first stage. If they belong to a particular faction, the study may be biased.
- Snowball sampling may have serious problems, if there are major differences between those who are widely known by others and those who are not.

7.4 SAMPLE SIZE DETERMINATION AND POWER OF A STUDY

Sample size determination is a major step in the design of a research study. It is very important to have proper sample size for study or justification of the problem. The size of sample depends upon the precision, the researcher desires in estimating the population parameter at a particular confidence level. There is no specific rule that can be used to determine sample size. The sample size should be small enough to avoid unnecessary expenses and large enough to avoid intolerable sampling error. Sample size which fulfills the requirements of efficiency representativeness, reliability and flexibility is called optimum sample size. The sample size required to reject or accept a study hypothesis is determined by the power of test.

The sample size is based on the following parameters:

- **Size of the Population:** If the total population to be studied is very large then, the sample size also more.
- **Nature of Population:** Population may be homogeneous or heterogeneous. If the population is homogeneous, a small sample may serve the purpose but if the population is heterogeneous, a large sample required to serve the purpose.
- **Nature of Study:** In some investigation, the items need to be intensively and continuously studied. In such cases, sample size should be small. Studies which are not likely to be repeated and are quite extensive in nature then the sample size may be large.
- **Type of Sampling:** Sampling methods play an important role in determining the size of a sample. In simple random sampling require larger sample but properly drawn stratified sampling plan, small sample gives better results.

- **Desired Accuracy or Confidence Level:** If sample size is larger then degree of accuracy is more. Researcher has to think of the level at which he will be confident that his sample is representative. The 95% confidence level is chosen meaning that one anticipates that there is a 95% chance that the sample and the population will look alike and a 5% chance that it will not.
- **Sampling Error or Desired Risk Level:** If the sample size is more then minimum sample error. The study of parents who want to send their children to English medium private schools or government schools. If the average annual family income of parents in the area is ₹ 10,00,000 then, the investigator should make sure that his sample's average income is a close to ₹ 10,00,000.
- **Purpose of Study:** Sample size depends on study whether it is descriptive, exploratory or explanatory. Qualitative or quantitative studies may be considered for sampling size.
- **Availability of Funds:** The size of sample also depends upon the availability of funds for the research process. Financial sources should be kept in view while determining the size of a sample.

It is very difficult task for the researcher to determine the size of sample. In experimental study, it is essential to equate the control and experimental groups, but in survey study sample should be representative of population. Therefore, size of sample is an important parameter for the representativeness. The precision of data is determined primarily by the size of the sample, rather than by the percentage of the population represented in the sample.

A number of formulae have been devised for determining the sample size depending upon the available of information.

$$n = \left(\frac{z \cdot \sigma}{d}\right)^2$$

where,

n = Sample size

z = Specific level of confidence or desired degree of precision

σ = Standard deviation of the population

d = Difference between population mean and sample mean

Specific level of confidence (z) at 1% level of significance or 99% confidence level the value of 'z' is 2.58 and 5% level of significance or 95% confidence level the value of 'z' is 1.96.

If standard deviation of population is 12, population mean is 36, sample mean is 30 and the confidence level of 99% then the sample size is calculated as follows:

- The confidence level of 99% i.e. $z = 2.58$.
- Standard deviation of population i.e. $\sigma = 12$.
- Difference between population mean (36) and sample mean (30) i.e. $d = 36 - 30 = 6$.

$$n = \left(\frac{z\sigma}{d}\right)^2$$

$$n = \left(\frac{2.58 \times 12}{6}\right)^2 = (5.16)^2 = 26.62.$$

In studies where the plan is to estimate the proportion of successes in a dichotomous outcome variable (yes/no) in a single population, the formula used for determining sample size is:

$$n = p(1 - p)\left(\frac{Z}{E}\right)^2$$

Where,

n = Sample size

Z = Value from the standard normal distribution reflecting the confidence level

E = Desired margin of error.

p = Proportion of successes in the population.

The equation to determine the sample size for determining p seems to require knowledge of p. The range of p is 0 to 1, and therefore the range of p(1 − p) is 0 to 1. The value of p that maximizes p(1 − p) is p = 0.5. Consequently, if there is no information available to approximate p, then p = 0.5 can be used to generate the most conservative, or largest, sample size.

In studies where the plan is to estimate the difference in means between two independent populations, the formula for determining the sample sizes required in each comparison group is given below:

$$n_i = 2\left(\frac{Z\sigma}{ES}\right)^2$$

Where,

n_i = Sample size required in each group (i = 1, 2),

Z = Value from the standard normal distribution reflecting the confidence level

E = Desired margin of error.

σ = Standard deviation of the outcome variable.

In studies where the plan is to estimate the difference in proportions between two independent populations (i.e., to estimate the risk difference), the formula for determining the sample sizes required in each comparison group is:

$$n_i = \{p_1(1 - p_1) + p_2(1 - p_2)\}\left(\frac{Z}{E}\right)^2$$

Where,

n_i = Sample size required in each group (i = 1, 2),

Z = Value from the standard normal distribution reflecting the confidence level (e.g., Z = 1.96 for 95%)

E = Desired margin of error.

p_1 and p_2 = The proportions of successes in each comparison group.

The difference between two groups in a study will be explored in terms of estimate of effect, appropriate confidence interval, and P value. The confidence interval indicates the likely range of values for the true effect in a population while P value determines how likely it is that the observed effect in the sample is due to chance. A related quantity is the statistical power of the study and it is the probability of detecting a predefined clinical significance. High power ideal study means that the study has a high chance of detecting a difference between groups if it exists. The ideal power for any study is considered to be 80%. For example, if study has 80% power, it has an 80% chance of detecting an effect that exists. Generally, 90% power or more is recommended to calculate sample size to achieve real effect of the experiment.

Statistical power is generally calculated with major two objectives given as follows:

- Power can be calculated before data collection based on information from previous studies to decide the sample size needed for the current study.
- It can also be calculated after data analysis.

The second situation occurs when the result turns out to be non-significant. In this case, statistical power is calculated to verify whether the non-significance result is due to lack of relationship between the groups or due to lack of statistical power.

Power is the probability that the statistical test results in rejection of H_o when a specified alternative is true. The 'stronger' the power, the better the chance that the null hypothesis will be rejected when, in fact, H_o is false. The larger the power, the more sensitive is the test. Power is defined as $1 - \beta$. The larger the error, the weaker is the power.

There are different parameters that can affect the power of a test such as sample size (n), significance level of test (α), "true" value of tested parameter, etc. The power of a study or statistical test is the probability that it correctly rejects the null hypothesis (H_o) when the null hypothesis is false (i.e., the probability of not committing a Type II error, or β).

$$\text{Power} = \text{Smallest difference} \times \sqrt{\frac{n}{\text{Variance}}}$$

Statistical power is positively correlated with the sample size as a larger sample size gives greater power. However, researchers should be clear to find a difference between statistical difference and scientific difference. A larger sample size enables statistician to find smaller difference statistically significant, the difference found may not be scientifically meaningful.

7.5 RESEARCH HYPOTHESIS

The second important step in the formulation of a research problem is the construction of hypothesis. The hypothesis is a tentative solution of a problem. It is a specific, testable prediction about research study. It is very essential to a scientist to understand the meaning and nature of hypothesis. Hypothesis need to be clear, precise and capable of being tested.

Hypothesis is a tentative statement about the solution of the problem. The term hypothesis has been defined in several ways. A hypothesis is a provisional formulation or possible solution or tentative explanation or suggested answer to the problem being faced by the researcher.

Hypothesis is an important part of scientific research. The importance of hypothesis is generally recognised more in the studies which aim to make predictions about some outcome. In experimental research, the scientists are interested in making predictions about the outcome of the experiment and hence, the role of hypothesis is most important. In the historical or descriptive research, the researcher is investigating the history of nation or a village and thus may not have a basis for making a prediction of results. Therefore, a hypothesis may not be required in such fact-finding studies. If a researcher is tracing the history of an university or making a study about the results of a coming Loksabha elections, the facts or data he gathers will prove useful only if he is able to draw generalizations from them. Hypothesis is recommended for all major studies to explain observed facts, conditions or behaviour and to serve as a guide in the research studies. Working hypothesis or a tentative hypothesis is described as the best guess or statement derivable from known or available evidence. The amount of evidence and quality, of it, determine other forms of hypothesis.

7.5.1 Null Hypothesis

A null hypothesis is a hypothesis that shows there is no statistical significance between the two variables in the hypothesis. Null hypothesis is a non-directional hypothesis that proposes no relationship between two variables. For example, there is no significant difference in academic performance of college students who participate in sports and sports non-participating students. Since, a null hypothesis can be statistically tested then it is called 'statistical hypothesis'. They are also called the testing hypothesis by converting them into null form. The proponents of null hypothesis emphasize that the researcher must remain unbiased throughout the study. Researcher may reject the null hypothesis by showing that the outcome mentioned in the declarative hypothesis does occur. The quantum of it is such that it cannot be easily dismissed as having occurred by chance. It is the hypothesis that the researcher tries to disprove.

If the hypothesis is that "the consumption of a particular medicine reduces the chances of heart arrest", the null hypothesis will be "the consumption of the medicine doesn't reduce the chances of heart arrest. "If the hypothesis is that, "If random test scores are collected from men and women, does the score of one group differ from the other?" a possible null hypothesis will be that the mean test score of men is the same as that of the women.

$$H_0: \mu_1 = \mu_2$$

Where,

H_0 = Null hypothesis,

μ_1 = Mean of population 1, and

μ_2 = Mean of population 2.

A stronger null hypothesis is that the two samples are drawn from the same population, such that the variances and shapes of the distributions are also equal.

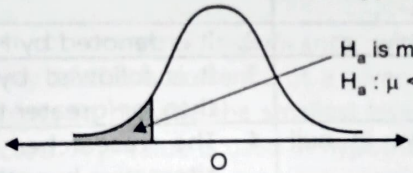
Statistical hypotheses are tested using a four-step process. The first step is for the analyst to state the two hypotheses so that only one can be right. The next step is to formulate an analysis plan, which outlines how the data will be evaluated. The third step is to carry out the plan and physically analyze the sample data. The fourth step is to analyze the results and either reject the null hypothesis, or claim that the observed differences are explainable by chance alone.

The principle of the null hypothesis is collecting the data and determining the chances of the collected data in the study of a random sample, proving that the null hypothesis is true. In situations or studies where the collected data doesn't complete the expectation of the null hypothesis, it is concluded that the data doesn't provide sufficient or reliable pieces of evidence to support the null hypothesis and thus, it is rejected.

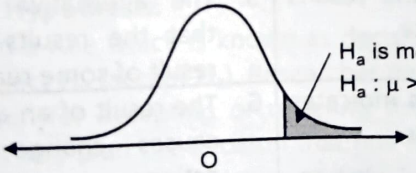
7.5.2 Alternative Hypothesis

Alternative hypothesis defines there is a statistically important relationship between two variables. The alternative or experimental hypothesis reflects that there will be an observed effect for our experiment. It is contradictory to the null hypothesis and denoted by H_a or H_1. In many cases, the alternate hypothesis will just be the opposite of the null hypothesis. For example, the null hypothesis might be "There was no change in the water level this spring," and the alternative hypothesis would be "There was a change in the water level this spring". The alternative hypothesis is the hypothesis that is to be proved that indicates that the results of a study are significant and that the sample observation is not results just from chance but from some non-random cause. It is a hypothesis that the researcher tries to prove.

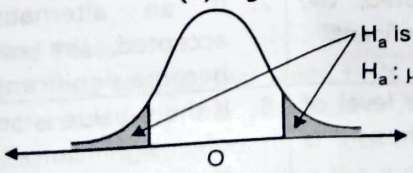
Basically, there are three types of the alternative hypothesis (Fig. 7.2).



(a) Left-tail Test



(b) Right-tail Test



(c) Two-tail Test

Fig. 7.2: Types of Hypothesis Tests

(a) Left-Tailed: Here, it is expected that the sample proportion (μ_1) is less than a specified value which is denoted by μ_2, such that;

$$H_1 : \mu_1 < \mu_2$$

(b) Right-Tailed: It represents that the sample proportion (π) is greater than some value, denoted by π_0.

$$H_1 : \mu_1 > \mu_2$$

(c) Two-Tailed: According to this hypothesis, the sample proportion (denoted by π) is not equal to a specific value which is represented by π_0.

$$H_1 : \mu_1 \neq \mu_2$$

The null hypothesis for all the three alternative hypotheses, would be $H_1 : \mu_1 = \mu_2$. If the null hypothesis is rejected, then we accept the alternative hypothesis. If the null hypothesis is not rejected, then we do not accept the alternative hypothesis. The difference between null hypothesis and alternative hypothesis is given in Table 7.1.

An alternative hypothesis provides the researchers with some specific restatements and clarifications of the research problem. The alternative hypothesis is important as they prove that a relationship exists between two variables selected and that the results of the study conducted are relevant and significant.

Table 7.1: Difference between null hypothesis and alternative hypothesis

Null hypothesis	Alternative hypothesis
1. This hypothesis states that there is no relationship between two phenomenon under consideration or that there is no association between two groups.	1. Alternative hypothesis states that there is a relationship between two selected variables in a study.
2. It is denoted by H_0.	2. It is denoted by H_1 or H_a.
3. It is followed by 'equals to' sign.	3. It is followed by not equals to, 'less than' or 'greater than' sign.
4. The researcher tries to disprove in null hypothesis.	4. The researcher tries to prove in alternative hypothesis.
5. This hypothesis believes that the results are observed as a result of chance.	5. The alternative hypothesis believes that the results are observed as a result of some real causes.
6. The result of the null hypothesis indicates no changes in opinions or actions.	6. The result of an alternative hypothesis causes change in opinions and actions.
7. If the null hypothesis is accepted, the results of the study become insignificant.	7. If an alternative hypothesis is accepted, the results of the study become significant.
8. If the p-value is greater than the level of significance, the null hypothesis is accepted.	8. If the p-value is smaller than the level of significance, an alternative hypothesis is accepted.

7.5.3 Hypothesis Testing

Hypothesis testing is a process of deciding statistically whether the findings of a research show chance or real effects at a given level of probability. Hypothesis testing is depending on probability theory and sampling. It consists of stating the hypothesis (null or alternative), construction of data gathering tools, collection of data, statistical analysis and drawing

inferences from the results. Research in which the independent variable is manipulated is called 'experimental hypothesis-testing research' and a research in which an independent variable is not manipulated is called 'non-experimental hypothesis testing research'. Some of important concepts in the context of testing of hypothesis are as follows:

Null Hypothesis and Alternative Hypothesis:

If two methods are compared about its superiority and proceed on the assumption that both methods are equally good, then this assumption is called as null hypothesis (H_o). It may conclude that method A is superior than method B then it is called as alternative hypothesis (H_a). Both hypothesis are chosen before the sample is drawn. Generally, in hypothesis testing, researcher can proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Researcher can assign the probabilities to different possible sample results if the null hypothesis is true, but this cannot be done if proceed with the alternative hypothesis. Hence, the use of null hypothesis is quite frequent.

The Level of Significance:

This is the essential concepts of hypothesis testing and is always considered in percentages (normally 5%). Significance level is the maximum values of the probability of rejecting a null hypothesis when it is true. It is usually determined in advance before testing the hypothesis. For example, if you assume the significance level to be 5%, it means that the researcher is ready to take 5% risk to reject the null hypothesis when it happens to be true.

P-Value:

The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A p-value is used in hypothesis testing to help you support or reject the null hypothesis.

Decision Rule or Test of Hypothesis:

Researcher can design a rule which is known as decision rule according to which it may accept H_o (reject H_a) or reject H_o (accept H_a). Researcher must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis. For example, if five items are tested in the lot and plan the decision that null hypothesis will be accepted only if out of those five items, either none is defective or only 1 is defective; otherwise alternative hypothesis will be accepted.

Two-tailed and One-tailed Test:

In the context of hypothesis testing, two tailed tests and one-tailed tests are important and must be clearly understood by the researcher. A two tailed test rejects the null hypothesis if the sample mean is either more or less than the hypothesized value of the mean of the population. Test is appropriate when the null hypothesis is some specific value and the alternative hypothesis is a value not equal to the specified value of null hypothesis. In a two-tailed curve, there are two rejection regions also called critical regions. When the population means is either lower or higher than some hypothesized value, the one tailed test is considered to be appropriate. If the rejection area is only on the left tail of the curve, then this is known as left-tailed test.

7.6 SAMPLING AND NON-SAMPLING ERRORS

A sample survey requires study of small portions of a population, as there may be certain amount of inaccuracy in the information collected during the sampling analysis. This inaccuracy is called the sampling error or error variance. On the other hand, non-sampling errors (systematic errors) mainly occurs due to errors of computation at the state of classification and processing of data. Errors in sampling can be classified as random error and systematic error or sampling errors and error of measurement. Sampling error is a function of sampling size and systematic error is the result of non-sampling factors like study design, correctness of execution sample frame errors, random sampling error and non-response error. Random sampling error and systematic error associated with the sampling process may combine to yield a sample that is less than perfectly representative of the population.

Sampling Errors:

These errors are not a measurement error not a systematic bias in sample. It is the error which depends on the representativeness of the sample. The precision of the sample is greater when sampling error is less. Sampling errors are also classified as biased errors and unbiased errors. The process of selection and estimation of samples may have some bias which leads to biased errors. Judgment sampling is used in a research survey instead of simple random sampling; some bias is introduced in the result due to judgment of the investigator in selecting the sample. These errors are biased sampling errors. Unbiased errors are caused due to disagreement between the population units selected in the sample and those not selected. Errors occur in final result due to fact of difference in the unit.

Bias may arise due to faulty selection of sample or substitution. The easiest method of avoiding bias is by selecting the sample randomly. Sampling errors are reduced by increasing the sample size.

Non-sampling Errors:

Non-sampling errors mainly occur due to incorrect method of interviews, lack of experience of investigators, inadequate data specification, errors in data processing operations, errors in classification of data etc.

These sampling errors can be reduced by controlling all the above factors. Generally, non-sampling errors increases with increase in sample size. Therefore, size of sample should be optimum so as to minimize a sum of sampling and non-sampling errors.

7.7 TYPES OF ERRORS

A statistically significant result cannot prove that a research hypothesis is correct. Because a p-value is based on probabilities, there is always a chance of making an incorrect conclusion regarding accepting or rejecting the null hypothesis (H_o). Researcher generally make a decision using statistics there are four possible outcomes, with two representing correct decisions and two representing errors (Table 7.2).

Table 7.2: Errors in testing of hypothesis

	Accept H_o	Reject H_o
H_o (true)	Correct decision	Type-I error (α-error)
H_o (false)	Type-II error (β-error)	Correct decision

Two types of errors may occur while testing a hypothesis. These are called as Type-I error and the Type-II error.

- **Type-I error:** Type-I error means rejection of null hypothesis which should have been accepted. In other words, this is the error of accepting an alternative hypothesis when the results can be attributed to chance. Type-I error is known as alpha (α) error or level of significance of test. Type I errors have a probability of "α" correlated to the level of confidence that you set. A test with a 95% confidence level means that there is a 5% chance of getting a type I error. Type I errors can happen due to bad luck (the 5% chance has played against you) or because you didn't respect the test duration and sample size initially set for your experiment. Consequently, a type I error will bring in a false positive. This means that you will wrongfully assume that your hypothesis testing has worked even though it hasn't. Type I error is caused when something other than the variable affects the other variable, which results in an outcome that supports the rejection of the null hypothesis. The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis.
- **Type-II error:** Type-II errors are referred to as "false negatives" and this error means accepting the null hypothesis which should have been rejected. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. This error is known as beta (β) error. The difference between type I and type II errors are given in table 7.3. Type II errors happen when you inaccurately assume that no winner has been declared between a control version and a variation although there actually is a winner. A type II error would occur if we accepted that the drug had no effect on a disease, but in reality that drug is effective. In more statistically accurate terms, type II errors happen when the null hypothesis is false and you subsequently fail to reject it. The probability of committing a Type II error is calculated by subtracting the power of the test from 1.

Table 7.3: Difference between type I and Type II errors.

Type I error	Type II error
1. Type I error is the error caused by rejecting a null hypothesis when it is true.	1. Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
2. Type I error is denoted by α.	2. Type II error is denoted by β.
3. Type I error is refers as false positive.	3. Type II error is refers as a false negative.
4. It is a false rejection of a true hypothesis.	4. It is the false acceptance of an incorrect hypothesis.
5. It is caused by luck or chance.	5. It is caused by a smaller sample size or a less powerful test.
6. The probability of this error is equal to the level of significance.	6. The probability of this error is equal to one minus the power of the test.
7. Type-I error can be reduced by decreasing the level of significance.	7. Type II error can be reduced by increasing the level of significance.
8. It is associated with rejecting the null hypothesis.	8. It is associated with rejecting the alternative hypothesis.

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If Type I error is fixed at 4%, it means that there are about 4 chances in 100 that will reject H_o when H_o is true. When we try to reduce Type I error then the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. The only solution to this problem is to set an appropriate level for the errors by considering the costs of penalties attached to them or to strike a proper balance between both types of errors.

P (reject H_o when it is true) = P (reject H_o/H_o) = α

and

P (accept H_o when it is wrong) = P (reject H_o/H_1) = β

The α and β are called the size of Type I Error and size of Type II Error respectively.

7.8 STANDARD ERROR OF MEAN (SEM)

Standard error of mean is also called the standard deviation of the mean. It is a method used to estimate the standard deviation of a sampling distribution. In the context of statistical data analysis, the mean and standard deviation of sample population data is used to estimate the degree of dispersion of the individual data within the sample but the standard error of mean (SEM) is used to estimate the sample mean dispersion from the population mean. The standard error (SE) along with sample mean is used to estimate the approximate confidence intervals for the mean. The estimation with lower standard error of mean (SEM) indicates that it has more precise measurement. The standard error of mean is always smaller than the standard deviation. SEM gives the accuracy of sample mean by measuring the sample-to-sample variability of the sample means. It describes how precise the mean of the sample is an estimate of the true mean of the population. The standard error of mean is calculated by using the following formula :

$$SE_x = \frac{S.D.}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

where,

σ = Standard deviation of the population

n = Size of the sample

Example 7.1: Calculate the standard error of mean of tablets not passes test of dissolution which are collected from 10 different batches out of 20 tablets.

Number of tablets (Not passes test) 3, 4, 4, 6, 5, 3, 3, 4, 3, 5

Solution:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{40}{10} = 4$$

First we have to calculate standard deviation.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
3	-1	1
4	0	0
4	0	0
6	2	4
5	1	1
3	-1	1

x	$(x - \bar{x})$	$(x - \bar{x})^2$
3	−1	1
4	0	0
3	−1	1
5	1	1
$\Sigma x = 50$		$\Sigma (x - \bar{x})^2 = 10$

Now, standard deviation, $(\sigma) = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$

$= \sqrt{\frac{10}{10}}$

$= \sqrt{1} = 1$

Standard error of mean $(SE_{\bar{x}}) = \frac{S.D.}{\sqrt{n}} = \frac{1}{\sqrt{10}} = \frac{1}{3.16}$

$= 0.316$

Example 7.2: If standard deviation of population is 15, population mean is 40, sample mean is 30 and the confidence level of 95%. Then calculate the sample size.

Solution: Sample size $(n) = \left(\frac{z \cdot \sigma}{d}\right)^2$

z = Specific level of confidence (95%) i.e. $z = 1.96$

σ = Standard deviation of population i.e. $\sigma = 15$

d = Difference between population mean and sample mean i.e. $d = 40 - 30 = 10$

$n = \left(\frac{1.96 \times 15}{10}\right)^2$

$= (2.94)^2 = 8.64$

QUESTIONS

(A) Objective Type Questions:

1. List different probability sampling techniques.
2. What is sampling?

(B) Short Answer Questions:

1. Write the importance of sampling in research.
2. Write the ideal characteristics of sampling.
3. Write advantages and disadvantages of sampling.
4. Differentiate between:
   (a) Sample and Population
   (b) Null hypothesis and Alternative hypothesis
   (c) Type-I error and type-II error.

5. Write advantages and disadvantages of:
   (a) Multi-stage sampling (b) Quota sampling
6. Write notes on:
   (a) SEM (b) Snowball sampling
   (c) Sampling and non-sampling errors.

C) Long Answer Questions:

1. Explain different parameters involved in sampling design.
2. How will you determine the size of sample?

D) Multiple Choice Questions:

1. To test null hypothesis, a researcher uses .....
   (a) Karl Pearson Coefficient Test (b) t test
   (c) ANOVA (d) factorial analysis
2. A subset of the population is called .....
   (a) Element (b) Sampling unit
   (c) Sample (d) Sampling frame
3. Which one is called non-probability sampling?
   (a) Quota sampling (b) Cluster sampling
   (c) Systematic sampling (d) Stratified random sampling
4. The following is true about sampling except .....
   (a) Sample is a part of population
   (b) Sampling saves time, money and energy
   (c) Sampling helps in estimating sampling error
   (d) Sampling increases cost of research
5. Which of the following would generally require the largest sample size?
   (a) Cluster sampling (b) Simple random sampling
   (c) Systematic sampling (d) Proportional stratified sampling
6. In the process of conducting research 'Formulation of Hypothesis" is followed by
   (a) Statement of Objectives (b) Analysis of Data
   (c) Selection of Research Tools (d) Collection of Data
7. Which of the following is not a non-probability sampling?
   (a) Judgmental sampling (b) Convenience sampling
   (c) Extensive sampling (d) Cluster sampling
8. Sampling in qualitative research is similar to which type of sampling in quantitative research?
   (a) Probability sampling (b) Quota sampling
   (c) Stratified sampling (d) Purposive sampling
9. The following are steps in sampling process except .....
   (a) Defining target population
   (b) Selecting and identifying the sample method
   (c) Choosing sampling frame
   (d) Selection of research problem
10. Sample value is called .....
    (a) Parameter (b) Statistic
    (c) Variable (d) Data
11. Sampling which provides for a known non-zero chance of selection is .....
    (a) Probability sampling (b) Non probability sampling
    (c) Quota sampling (d) Extensive sampling

***