Directorate of Distance and Continuing Education

Manonmaniam Sundaranar University

Tirunelveli-627 012, Tamil Nadu.

# M.A. ECONOMICS

## (Second Year)

# RESEARCH METHODOLOGY

## SECM33

Compiled by

**Dr. G. Monikanda Prasad**
Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli - 627 012

# RESEARCH METHODOLOGY

## Unit – I Research in Economics

Research Methods in Economics Nature of Social Science Research, Research Methods in Social Science – Formulating the problem- Types and Sources of Hypothesis – Characteristics of a Good hypothesis- Components and types of research design – Collection of data – sources and methods-Presentation of results – Format of a Report.

## Unit – II Data Collection

Sampling and Data collection Sampling; Need, types, Probability sampling, random, systematic, stratified, multistage or cluster sampling, Non-Probability sampling; Purposive Judgment, quota and snowball sampling-Data collection; Primary and Secondary data; NSS and censes data Methods of data collection- Tools of data collection; schedule and questionnaire.

## Unit – III Research Design

Data Processing and Presentation Processing and analysis of data: Editing, coding and tabulation; use of computers in social science research-Diagrammatic and graphic representation of data- Interpretation of results and Report writing – Preparation of Project Proposals.

## Unit – IV Data Analysis – I

Statistical Inferences Census Versus sampling -Random and Non-Random sampling Techniques Estimation – Point and interval estimation – Statistics and Parameter – Standard Error – Confidence interval- Null and Alternative hypothesis – Type I Error and Type II Error, Level of Significance – Critical region – Steps in Testing of Hypothesis.

## Unit – V Data Analysis – II

Large and small Sample Tests Properties and uses of Normal Distribution – Standard normal ‘Z’ Statistic Z-Test of Significance of proportions, means and Correlation- ‘t’ Test for sample mean and Equality of mean – Paired ‘t’ Test- Chi-Square Test for Association of Attributes.

## Text Books

C.R. Kothari (2002), Research Methodology Vikas publishing House, New Delhi.

Goode W. J and Hatt(1952), Methods in Social Research, Mcgraw Hill Book Co, Tokyo.

# UNIT-I
# RESEARCH IN ECONOMICS

## Introduction

Statistical enquiry/survey is conducted only to collect information pertaining to the topic of study. The information may be quantitative or qualitative, but in statistical enquiry, we expert numerical information. Information may also be collected from sources other than enquiry. It means that there are two sources of information or data. Simply speaking, the two sources of data are:

Primary sources and Secondary sources.

To collect first-hand information, it is necessary to conduct enquiry or survey and for second hand information enquiry is not necessary, but there must be enough secondary sources. The primary data are original in character while secondary data have already been collected by someone for some other purpose and are now available for the present study. The census data are primary data to the census department, but to researchers and other people, they are secondary.

## Nature and scope of statistics

Statistics is the study of collection, organization, analysis, interpretation and presentation of data with the use of quantified models. In short, it is a mathematical tool that is used to collect and summarize data. Scope of Statistics: It presents the facts in numerical figures

## Nature of Statistics

Statistics is both science and art. Statistical methods are systematic and have a general application which makes it a science. Further, the successful application of these methods requires skills and experience of using the statistical tools. These aspects make it an art.

## Scope of Statistics

The scope of Statistics is very immense, the application of statistics goes into diverse fields such as solving social problems, industrial and scientific problems. Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.

## Uses and limitations of Statistics

Its primary scope is descriptive and inferential statistics, which includes summarizing and analyzing data, making predictions, and generalizations about a population. However, statistics also has its limitations, which include being limited to numerical data, sampling bias, and the inability to establish causation.

Statistics provide the information to educate how things work. They're used to conduct research, evaluate outcomes, develop critical thinking, and make informed decisions.

## Hypothesis

Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

There are three types of hypothesis tests: right-tailed, left-tailed, and two-tailed. When the null and alternative hypotheses are stated, it is observed that the null hypothesis is a neutral statement against which the alternative hypothesis is tested.

## Characteristics of the hypothesis:

- The hypothesis should be clear and precise to consider it to be reliable.
- If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
- The hypothesis must be specific and should have scope for conducting more test.

## Components and Types of Research Design

### Meaning of Research Design

The formidable problem that follows the task of defining the research problem is the preparation of the design if the research project, popularly known as the "research design". Decisions reading what, where, when, how much, by what means concerning an inquiry or a research study constitute a research design.

A research design is the arrangement of conditions for collection and analysis if data in a manner that aims to combine relevance to the research purpose with economy in procedure.

As such the design includes an outline of what the researcher will do from writing the hypothesis and its operational implications to the final analysis of data. More explicitly, the design decisions happen to be in respect of:

(i) What is the study about?

(ii) Why is the study being made?

(iii) Where will the study be carried out?

(iv) What type of data is required?

(v) Where can the required data be found?

(vi) What periods of time will the study include?

(vii) What will be the sample design?

(viii) What techniques of data collection will be used?

(ix) How will the data be analysed?

(x) In what style will the report be prepared?

Keeping in view the above stated design decisions, one may split the overall research design into the following parts:

(a) the sampling design which deals with the method of selecting items to be observed for the given study;

the observational design which relates to the conditions under which the observations

are to be made;

(c) the statistical design which concerns with the question of how many items are to be

observed and how the information and data gathered are to be analysed; and

(d) the operational design which deals with the techniques by which the procedures specified

in the sampling, statistical and observational designs can be carried out.

## Need for Research Design

Research design is needed because it facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money. Just as for better, economical and attractive construction of a house, we need a blueprint (or what is commonly called the map of the house) well thought out and prepared by an expert architect, similarly we need a research design or a plan in advance of data collection and analysis for our research project. Research design stands for advance planning of the methods to be adopted for collecting the relevant data and the techniques to be used in their analysis, keeping in view the objective of the research and the availability of staff, time and money. Preparation of the research design should be done with great care as any error in it may upset the entire project.

Research design, in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work. Even then the need for a well thought out research design is at times not realized. It is, therefore, imperative that an efficient and appropriate design must be prepared before starting research operations. The design helps the researcher to organize his ideas in a form whereby it will be possible for him to look for flaws and inadequacies. Such a design can even be given to others for their comments and critical evaluation. In the absence of such a course of action, it will be difficult for the critic to provide a comprehensive review of the proposed study.

## Features of a Good Design

A good design is often characterised by adjectives like flexible, appropriate, efficient, economical and so on.

The design which minimises bias and maximises the reliability of the data collected and analysed is considered a good design. The design which gives the smallest experimental error is supposed to be the best design in many investigations.

Similarly, a design which yields maximal information and provides an opportunity for considering many different aspects of a problem is considered most appropriate and efficient design in respect of many research problems. Thus, the question of good design is related to the purpose or objective of the research problem and also with the nature of the problem to be studied. A design may be quite suitable in one case, but may be found wanting in one respect or the other in the context of some other research problem. One single design cannot serve the purpose of all types of research problems.

A research design appropriate for a particular research problem, usually involves the consideration of the following factors:

(i) the means of obtaining information;

(ii) the availability and skills of the researcher and his staff, if any;

(iii) the objective of the problem to be studied;

(iv) the nature of the problem to be studied; and

(v) the availability of time and money for the research work.

## Characteristics of a Good Sample Design

From what has been stated above, we can list down the characteristics of a good sample design as under:

(a) Sample design must result in a truly representative sample.

(b) Sample design must be such which results in a small sampling error.

(c) Sample design must be viable in the context of funds available for the research study.

(d) Sample design must be such so that systematic bias can be controlled in a better way.

(e) Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

**Collection of Data**

The task of data collection begins after a research problem has been defined and research design/plan chalked out. While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz., primary and secondary. The primary data are those which are collected afresh and for the first time, and thus happen to be original in character.

The secondary data, on the other hand, are those which have already been collected by someone else and which have already been passed through the statistical process. The researcher would have to decide which sort of data he would be using (thus collecting) for his study and accordingly he will have to select one or the other method of data collection.

The methods of collecting primary and secondary data differ since primary data are to be originally collected, while in case of secondary data the nature of data collection work is merely that of compilation. We describe the different methods of data collection, with the pros and cons of each method.

**Collection of Primary Data**

Primary data during the course of doing experiments in an experimental research but in case we do research of the descriptive type and perform surveys, whether sample surveys or census surveys, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews.

Methods of collecting primary data, particularly in surveys and descriptive researches. important ones are: (i) observation method, (ii) interview method, (iii) through questionnaires,(iv) through schedules, and (v) other methods which include (a) warranty cards; (b) distributor audits; (c) pantry audits; (d) consumer panels; (e) using mechanical devices; (f) through projective techniques; (g) depth interviews, and (h) content analysis.

**Observation Method**

The observation method is the most commonly used method specially in studies relating to behavioural sciences. In a way we all observe things around us, but this sort of observation is not scientific observation. Observation becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability.

Under the observation method, the information is sought by way of investigator's own direct observation without asking from the respondent.

The observation is characterised by a careful definition of the units to be observed, the style of recording the observed information, standardised conditions of observation and the selection of pertinent data of observation, then the observation is called as structured observation.

When observation is to take place without these characteristics to be thought of in advance, the same is termed as unstructured observation. Structured observation is considered appropriate in descriptive studies,

If the observer observes by making himself, more or less, a member of the group he is observing so that he can experience what the members of the group experience, the observation is called as the participant observation.

But when the observer observes as a detached emissary without any attempt on his part to experience through participation what others feel, the observation of this type is often termed as non-participant observation.

**Interview Method**

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews.

(a) Personal interviews: Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons.

This method is particularly suitable for intensive investigations. The method of collecting information through personal interviews is usually carried out in a structured way. As such we call the interviews as structured interviews. Such interviews involve the use of a set of predetermined questions and of highly standardised techniques of recording. Thus form and order prescribed.

The unstructured interviews are characterised by a flexibility of approach to questioning. Unstructured interviews do not follow a system of pre-determined questions and standardised techniques of recording information. In a non-structured interview, the interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires. He may even change the sequence of questions. He has relatively greater freedom while recording the responses to include some aspects and exclude others. Unstructured interviews also demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, however, happens to be the central technique of collecting information in case of exploratory or Focussed interview is meant to focus attention on the given experience of the respondent and its effects. Under it the interviewer has the freedom to decide the manner and sequence in which the questions would be asked and has also the freedom to explore reasons and motives. The main task of the interviewer in case of a focussed interview is to confine the respondent to a discussion of issues with which he seeks conversance. Such interviews are used generally in the development of hypotheses and constitute a major type of unstructured interviews.

The clinical interview is concerned with broad underlying feelings or motivations or with the course of individual’s life experience. The method of eliciting information under it is generally left to the interviewer’s discretion.

Non-directive interview, the interviewer’s function is simply to encourage the respondent to talk about the given topic with a bare minimum of direct questioning. The interviewer often acts as a catalyst to a comprehensive expression of the respondents’ feelings and beliefs.

(b) *Telephone interviews:* This method of collecting information consists in contacting respondents on telephone itself. It is not a very widely used method, but plays important part in industrial surveys, particularly in developed regions.

**Format of a Report**

A report typically includes the following sections:

Title page: Includes the title, author, class, section, and date

Table of contents: An index page for the report

Executive summary: A summary of the report's main points, including the topic, data, analysis, and recommendations

Introduction: Provides the origin and essentials of the main subject

Body: The main report, which may include methods, findings, research, and results

Conclusions: Includes inferences, measures taken, and projections

References: A list of sources cited in the text

Appendices: Additional information

Research report, must necessarily be conveyed enough about the study so that he can place it in its general scientific context, judge the adequacy of its methods and thus form an opinion of how seriously the findings are to be taken. For this purpose there is the need of proper layout of the report. The layout of the report means as to what the research report should contain. A comprehensive layout of the research report should comprise

(A) Preliminary pages - title and date, acknowledgements, table of contents, list of tables and illustrations.

(B) The main text; - Introduction, Statement of the Problem, Methodology, Statement of findings and recommendations, Results, Implications of the results.

(C) The end matter – Bibilography, Appendices

# UNIT –II
# DATA COLLECTION

## Data Collection

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

## Sampling Design

A complete enumeration of all items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry, when all items are covered, no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observation increases. Moreover, there is no way of checking the element of bias or its extent except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy.

## Need

However, it needs to be emphasized that when the universe is a small one, it is no use resorting to a sample survey. When fields studies are undertaken in practical life, considerations of time and cost almost invariably lead to a selection of respondents i.e., selection of only a few items. The respondents selected should be as representative of the total population as possible in order to produce a miniature cross-section. The selected respondents constitute what is technically called a 'sample' and the selection process is called 'sampling technique'. The survey so conducted is known as 'sample survey'. Algebraically, let the population size be N and if a part of size n (which is <N) of this population is selected according to some rule for studying some characteristic of the population, the group consisting of these *n* units is known as 'sample'. Researcher must

prepare a sample design for his study I.e., he must plan how a sample should be selected and of what size such a sample would be.

Sampling is used practice for a variety of reasons such as:

(1) Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.

(2) Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.

(3) Sampling remains the only way when population contains infinitely many members.

(4) Sampling remains the only choice when a test involves the destruction of the item under study.

(5) Sampling usually enables to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population.

**Characteristics of a goods sample design**

From what has been stated above, we can list down the characteristics of a good sample design as under:

(a) Sample design must result in a truly representative sample.

(b) Sample design must be such which results in a small sampling error.

(c) Sample design must be viable in the context of funds available for the research study.

(d) Sample design must be such so that systematic bias can be controlled in a better way.

(e) Sample should be such that the results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

**Different types of sample designs**

There or different types of sample designs based on two factors viz., the representation basis and the element selection technique. On the representation basis, the sample may be probability sampling is it may be non-probability sampling. Probability sampling based on the concept of random selection basis, the sample may be either unrestricted or restricted. When each sample element is drawn individually from the population at large, then the sample so drawn is known as 'unrestricted sample', whereas all other forms of sampling are covered under the term 'restricted sampling'. The following chart exhibits the sample design as explained above.

Thus, sample designs are basically of two types' viz., non-probability sample and probability sampling. We take up these two designs separately.

Chart showing basic sampling designs

| | Representation basis | |
|---|---|---|
| | ↓ Probability sampling | ↓ Non-probability sampling |
| Element selection technique<br>↓<br>Unrestricted sampling | Simple random sampling | Haphazard sampling or convenience sampling |
| Restricted sampling | Complex random sampling | Purposive sampling |

Non-probability sampling: Non-probability sampling is that sampling procedure which does not afford any basis for estimating the probability that each item in the population has of being included in the sample. Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgments sampling. In this type of sampling, items for the sample are selected deliberately by the researcher; his choice concerning the item remains supreme. In other words, under non-probability sampling the organizers of the inquiry purposively choose the particular units of the universe for constituting a sample on the basis that the small mass that they so select out of a huge one will be typical or representative of the whole.

Quota sampling is also example of non-probability sampling. Under quota sampling the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled.

**Probability sampling:**

Probability sampling is also known as 'random sampling' or 'chance sampling'. Under this sampling design, every item of the universe has an equal chance of inclusion in the sample. It is, so to say, a lottery method in which individual units are picked up from the whole group not deliberately but by some mechanical process. Here it is blind chance alone that determines whether one item or the other is selected. The results obtained from probability of random sampling van are assured in terms of probability.

(a) It gives each element in the population an equal probability of getting into the sample; and all choices are independent of one another.

(b) It gives each possible sample combination an equal probability of being chosen.

**Universe population:**

From a statistical point of view, the term 'Universe ' refers to the total of the items or units in any field of inquiry, whereas the term 'population' refers to the total of items about which information is desired. The attributes that are the object of study are referred to as characteristics and the units possessing them are caked as elementary units. The aggregate of such units is generally described as population. Thus, all units in any field of inquiry constitute universe and all elementary units constitute population. Quit often, we do not find any difference between population and universe, and as such the two terms are taken as interchangeable. However, a researcher must necessarily define these terms precisely.

**Sampling frame:**

The elementary units or the group or cluster of such units may form the basis of sampling process in which case they are called as sampling units. A list containing all such sampling units is known as sampling frame. Thus sampling frame consists of a list of items from which the sample is to be drawn. If the population is finite and the time frame is in the present or past, then it is possible for the frame to be identical with the population. In most cases they are not identical because it is often impossible to draw a sample directly from population. As such this frame is either constructed by a researcher for the purpose of his study or may consist of some existing list of the population. For instance, one can use telephone directory as a frame for conducting opinion survey in a city. Whatever the frame may be, it should be a good representative of the population.

**Sample design:**

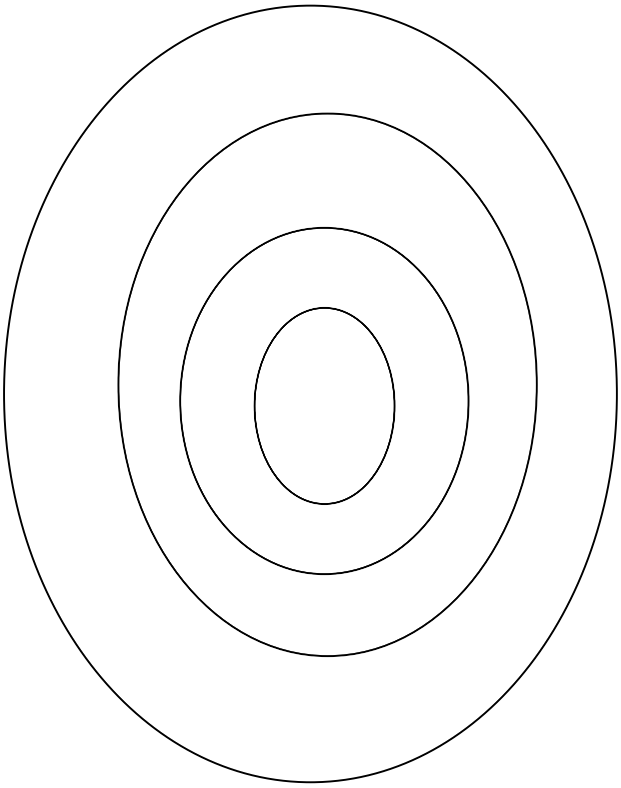
A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data are collected. Various sampling designs have already been explained earlier in the book.

**Statistic(s) and parameter(s):**

A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population. Thus, when we work out certain measures such as mean, median, mode or the like ones from samples, then they are called statistic(s) for they describe the characteristics of sample. But when such measures describe the characteristics of a population, they are known as parameter(s).

**Sampling error:**

Sample surveys do imply the study if a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be termed as sampling error or error variance. In other words sampling errors are those errors which arise on account of sampling and they generally happen to be random variations in the sample estimates around the true population values. The meaning of sampling error can be easily understood from the following diagram.



**Precision:**

Precision is the range within which the population average will lie in accordance with the reliability specified in the confidence level as a percentage if the estimate ± or as a numerical quantity. For instance, if the estimate is Rs 4000 and the precision desired is 4%, then the true value will be no less than Rs 3840 and no more than Rs 4160. This is the range within which the true answer should lie. But if we desire that the estimate should not deviate from the actual value by more than Rs 200 in either direction, in that case the range would be Rs 3800 to Rs 4200.

**Confidence level and significance level:**

The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits. Thus, if we take a confidence level of 95%, then we mean that there are 95 chances in 100 that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100 that is does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within the range, and the significance level indicates the likelihood that the answer will fall outside that range. We should also remember that if the confidence level is 95%, then the significance level will be (100-95) i.e., 5% if the confidence level is 99%, the significance level is (100-99) i.e., 1%, and so on.

**Sampling distribution:**

We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say ,mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution if any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sample distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact the mean of a sampling distribution is the same as the mean the universe.

## Theoretical basis of sampling:

On the basis of sample study we can predict and generalize the behavior of mass phenomena. This is possible because there is no statistical population whose elements would vary from each other without limit. For example, wheat varies to a limited extent in color, protein content, length, weight, etc., it can always be identified as wheat. Similarly, apples of the same tree may vary in size, color, taste, weight, etc., but they can

Always be identified as apples. Thus we find that although diversity is a universal quality of mass data, every population has characteristic properties with limited variation. This makes possible to select a relatively small unbiased random sample that can portray fairly well the traits of the population.

There are two important laws on which the theory of sampling is based:

- Law of 'Statistical Regularity', and
- Law of 'Inertia of Large Numbers'.

### Law of Statistical Regularity

This law is derived from the mathematical theory of probability. In the words of King: "The law if statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group". In other words, this law points out that if a sample is taken at random from a population, it is likely to possess almost the same characteristics as that of the population. This law directs our attention to one very important point, that is, the desirability of choosing the sample at random.

By random selection we mean a selection where each and every item of the population has an equal chance of being selected in the sample.

**Law of Inertia of Large Number:**

This law is a corollary of the law of statistical regularity. It is of great significance in the theory of sampling. It states that, other things being equal, larger the size if the sample, more accurate the results are likely to be. This is because large numbers are more stable as compared to small ones. The difference in the aggregate result is likely to be insignificant, when the number in the sample is large, because when large numbers are considered the variations in the component parts tend to balance each other and, therefore, the variation in the aggregate is insignificant.

**Essential of sampling:**

If the sample results are to have any worthwhile meaning, it is necessary that a sample possesses the following essentials:

(1) **Representativeness:** A sample should be so selected that it truly represents the universe otherwise the results obtained may be misleading. To ensure representativeness the random method of selection should be used.

(2) **Adequacy:** the size of sample should be adequate, otherwise it may not represent the characteristics of the universe.

(3) **Independence:** All items of the sample should be selected independently of one another and all items of the universe should have the same chance of being selected in the sample of a particular item in one draw has influence on the probabilities of selection in any other draw.

(4) **Homogeneity:** When we talk of homogeneity we mean that there is no basic difference in the nature if units of the universe and that of the sample or less the same unit.

## Methods of Sampling:

The various methods of sampling can be grouped under two broad heads: probability sampling and non-probability sampling. Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being chosen for the sample.

Non-probability sampling methods are those which do not provide every item in the universe with a known chance of being included in the sample. The selection process is, at least, partially subjective.

## Advantages of Probability Sampling:

The following are the basic advantages of probability sampling methods:

(1) Probability sampling does not depend upon the existence of detailed information about the universe for its effectiveness.

(2) Probability sampling provides estimates which are essentially unbiased and have measurable precision.

(3) It is possible to evaluate the relative efficiency of various sample designs only when probability sampling is used.

## Limitation of Probability Sampling:

Despite the great advantages of probability sampling techniques mentioned above, it has certain limitations because of which non-probability sampling is quite often used in practice. These limitations are:

- Probability sampling requires a very high level of skill and experience for its use.
- It requires a lot of time to plan and execute a probability sample.
- The costs involved in probability sampling are generally large as compared to non-probability sampling.

Non-random sampling is a process of sample selection without the use of randomization. In other words, a non-random sample is selected on a basis other than the probability consideration such as convenience, judgment, etc.

**Non-probability sampling methods:**

(i) Judgment sampling;

(ii) Convenience sampling; and

(iii) Quota sampling.

**Probability sampling methods**

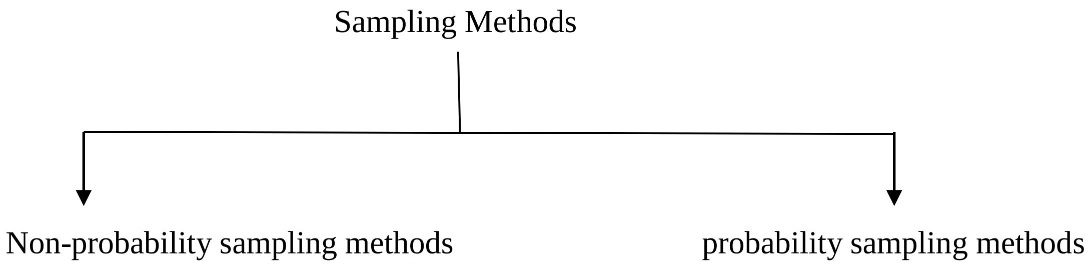
(a) Simple or unrestricted random sampling; and

(b) Restricted random sampling:

(i) Stratified sampling,

(ii) Systematic sampling, and

(iii) Cluster sampling.


Sampling Methods
Non-probability sampling methods
probability sampling methods

**Non-probability sampling methods**

**Judgment sampling**

In this method of sampling the choice of sample items depends exclusively on the judgment if the investigator. In other words, the investigator exercises his judgment in the choice and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristic under investigation. For example, if sample of ten students is to be selected form a class of sixty for analysing the spending habits of students, the investigator would select 10 students who, in his opinion, are representative if the class.

**Merits:**

Though the principles of sampling theory are not applicable to judgment sampling, the method is sometimes used in solving many types of economic and business problems. The use of judgment sampling is justified under a variety of circumstances:

(i) When only a small number of sampling units are in the universe, simple random selection may miss the more important elements, whereas judgment selection would certainly include them in the sample.

(ii) When we what to study some unknown traits of a population, some of whose characteristics are known, we may then stratify the population according to these known properties and select sampling units from each stratum on the basis of judgment. This method is used to obtain a more representative sample.

(iii) In solving everyday business problems and making public policy decisions, executives and public officials are often pressed for time and cannot wait for probability sample designs. Judgment sampling is then the only practical method to arrive at solutions to their urgent problems.

**Limitations:**

Judgment sampling method is however associated with the following limitations:

(i) This method is not scientific because the population units to be sampled may be affected by the personal prejudice or bias of the investigator. Thus, judgment sampling involves the risk that the investigator may establish foregone conclusions by including those items in the sample which conform to his preconceived notions. For example, if an investigator holds the view that the wages of workers in a certain establishment are very low, and if he adopts the judgment sampling method, he may include only those workers in the sample whose wages are low and thereby establish his point of view which may be far from the truth. Since an element of subjectiveness is possible, this method cannot be recommended for general use.

(ii) There is no objective way of evaluating the reliability of sample results. The success of this method depends upon the excellence in judgment. If the individual making decisions is knowledgeable about the population and has good judgment, then the resulting sample may be representative, otherwise the inferences based on the sample may be erroneous. It may be noted that even if a judgment sample is reasonably representative, there is no objective method for determining the size or likelihood of sampling error. This is a big defect of the method.

**Quota Sampling:**

Quota sampling is a type of judgment sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quotas are set up according to some specified characteristics such as so many in each of several income groups, so many in each age, so many with certain political or religious affiliations, and so on. Each interviewer is then told to interview a certain number of persons which constitute his quota. Within the quota, the selection of sample items depends on personal judgment. For example, in a radio listening, survey, the interviewers may be told to interview 500 people living in a certain area and that out of every 100 persons interviewed 60 are to be housewives, 25 farmers and 15 children under the age of 15, within these quotas the interviewer is for to select the people to be interviewed. The cost per person interviewed may be relatively small for a quota sample but there are numbers opportunities for bias which may invalidate the results. For example, interviewers may miss farmer working in the fields or talk with those housewives who are at home. If a person refuses to respond, the interviewer simply selects someone else. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is not widely used in practical work.

**Convenience Sampling:**

A convenience sample is obtained by selecting ‘convenient’ population units. The method of convenience sampling is also called the chunk. A chunk refers to that fraction of the population being investigated which is selected neither by probability nor by judgment but by convenience. A sample obtained from readily available lists such as automobile registrations; telephone directories, etc., is a convenience sample and not a random sample even if the sample is drawn at random from the lists. If a person is to submit a project report on labour-management relations in textile industry and he takes a textile mill close to his office and interviews some people over there, he is following the convenience sampling method. Convenience samples are prone to bias by their very nature-selecting population elements which are convenient to choose almost always make them special or different from the best of the elements in the population in some way.

Hence the results obtained by following convenience sampling method can hardly be representative of the population they are generally biased and unsatisfactory. However, convenience sampling is often used for making pilot studies. Questions may be tested and preliminary information may be obtained by the chunk before the final sampling design is decided upon.

**Probability Sampling Method**

**Simple or Unrestricted Random Sampling**

Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter if chance personal bias if the investigator does not influence the selection.

**Lottery Method**

This is a very popular method if taking a random sample. Under this method, all item if the universe are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container or drum. A blindfold selection is then made of the number of slips required to constitute the desired sample size. The selection of items thus depends entirely on chance. The method would be quite clear with the help of an example. If we want to take a sample of 10 persons out of a population of 100, the procedure is to write the names of the 100 persons on separate slips of paper, fold these slips, mix them thoroughly and then make a blindfold selection of 10 slips.

The above method is very popular in lottery draws there a decision about prizes is to be made. However, while adopting lottery method it is absolutely essential to see that the slips are of identical six, shape and color, otherwise there is a lot of possibility of personal prejudice and bias affecting the results.

**Table of Random Numbers:**

The lottery method discussed above becomes quite cumbersome as the size of population increases. An alternative method of random selection is that of using the table of random numbers. The random numbers are generally obtained by some mechanism which, when repeated a larger numbers from 0 to 9 and also proper frequencies for various combinations of numbers that could be expected in a random sequence of the digits 0 to 9. Several standard tables of random numbers are available, among which the following may be specially mentioned, as they have been tested extensively for randomness.

**Merits:**

Simple random sampling method has the following advantages:

- Since the selection of items in the sample depends entirely on chance there is no possibility of personal bias affecting the results.
- As compared to judgment sampling a random sample represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population.
- The analyst can easily assess the accuracy of this estimate because sampling errors follow the principles of chance. The theory of random sampling is further developed than that of any other type of sampling which enables the analyst to provide the most reliable information at the least cost.

**Limitation:**

This method is however associated with following limitations:

- The use of simple random sampling necessitates a completely catalogued universe from which to draw the sample. But it is often difficult for the investigator to have up-to-date lists of all the items of the population to be sampled. This restricts the use of this method in economic and business data where very often we have to employ restricted random sampling designs.
- The size of the sample required to ensure statistical reliability is usually larger under random sampling than stratified sampling.
- From the point of view of field survey it has been claimed that cased selected by random sampling time to be too widely dispersed geographically and that the time and cost of collecting data become too large.
- Random sampling may produce the most non-random looking results, for example, thirteen cards from a well-shuffled pack of playing cards may consist of one suit. But the probability of this type of occurrence is very, very low.

## Restricted Random Sampling

### Stratified sampling:

Stratified random sampling or simply or simply stratified sampling is one of the random methods which, by using the available information concerning the population, attempt to design a more efficient sample than obtained by the simple random procedure.

While applying stratified random sampling technique, the procedure followed is given below:

(a) The universe to be sampled is sub-divided into groups which are mutually exclusive and include all items in the universe.

(b) A simple random sample is then chosen independently from each group.

### How to select stratified random sample?

Some of the issues involved in setting up a stratified random sample are:

(i) **Base of Stratification:** what characteristic should be used to be sub divide the universe into different strata? As a general rule, strata are created on the basis of a variable known to be correlated with the variable of interest and for which information on each universe element is known. Strata should be constructed in a way which will minimize differences among sampling units within strata, and maximize difference among strata.

(ii) **Number of Strata:** how many strata should be constructed? The practical considerations limit the number of strata that is feasible, costs of adding more strata may soon outrun benefits. As a generalization more than six strata may be undesirable.

**(iii)** **Sample size within Strata:** How many observations should be taken from each stratum? When deciding this question we can use either a proportional or a disproportional allocation. In proportional allocation, one samples each stratum in proportion to its relative weight. In disproportional allocation this is not the case. It may be pointed out that proportional allocation approach is simple and if all one knows about each stratum is the number of items in that stratum; it is generally also the preferred procedure. In disproportional sampling, the different strata are sampled at different rates. As a general rule when variability among observations within a stratum is high, one samples that stratum at a higher rate than for strata with less internal variation.

**Merits**

Stratified sampling methods have the following advantages:

- **More representatives:** Since the population is first divided into various strata and then a sample is drawn from each stratum there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured. C.J Grohmann has rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.
- **Greater accuracy:** Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items.
- **Greater geographical concentration:** as compared with random sample, stratified samples can be more concentrated geographically, i.e., the units from the different strata may be selected in such a way that all of them are localized on one geographical area. This would greatly reduce the time and expenses of interviewing.

**Limitations:**

The limitation if these methods are:

- Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous items as otherwise the results may not be reliable. If proper stratification of the population is not done, the sample may have the effect of bias.

- The item from each stratum should be selected at random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.
- Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random sample cost per observation may quite high.

**Systematic Sampling:**

A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available.

**Merits:**

The systematic sampling design is simple and convenient to adopt. The time and work involved in sampling by this method are relatively less. The results obtained are also found to be generally satisfactory provided care is taken to see that there are no periodic features associated with the sampling interval. If populations are sufficiently large, systematic sampling can often be expected to yield results similar to those obtained by proportional stratified sampling.

**Limitations:**

The main limitation of the method is that it becomes less representative if we are dealing with populations having "hidden periodicities". Also if the population is ordered in a systematic way with respect to the characteristics the investigator is interested in, then it is possible that only certain types of items will be included in the population, or at least more of certain types than others. For instance, in a study of workers' wages the list may be such that every tenth worker on the list gets wages above Rs.750 per month.

**Multi stage Sampling or Cluster Sampling:**

Under this method, the random selection is made of primary, intermediate and final units from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method. Such as simple random sampling. Then, a sample if second stage units is selected from each of the selected first stage units, again by some suitable method which may be the same as or different from the method employed for the first stage units.

**Merits:**

Multi stage sampling introduced flexibility in the sampling method which lacking in the other methods. It enables existing divisions and sub-divisions of the field work to be concentrated and yet large area to be covered. Another advantage of the method is that subdivision into second stage units need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in surveys of underdeveloped areas where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

**Limitation:**

However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units which have been selected by some single stage process. We have discussed above the various random procedures in independent designs. In practice we often combine two or more of these methods into a single design.

**Merits and Limitations of Sampling**

Merits: the sampling technique has the following merits over the complete enumeration survey:

(i) **Less Time-consuming:** Since the sample is a study if a part of the population, considerable time and labour are saved when a sample survey is carried out. Time is saved not only in collecting data but also in processing it. For these reasons a sample provides more timely data in practice than a census.

(ii) **Less Cost:** Although the amount of effort and expense involved in collecting information is always greater per unit of the sample than a complete census, the total financial burden of a sample survey is generally less than that of a complete census. This is because of the fact that in sampling, we study only a part of population and the total expense of collecting data is less than that required when the census method is adopted. This is great advantage particularly in an underdeveloped economy where much of the information would be difficult to collect by the census method for lack of adequate resources.

(iii) **More Reliable Results:** Although the sampling technique involves certain inaccuracies owing to sampling errors, the result obtained is generally more reliable than that obtained from a complete count. There are several reasons for it. First, it is always possible to determine the extent of sampling errors. Secondly, other types of errors to which a survey is subject, such as inaccuracy of information, incompleteness of returns.

(iv) **More detail information:** Since the sampling technique saves time and money, it is possible to collect more detailed information in a sample survey. For example, if the population consists of 1,000 persons in a survey of the consumption pattern of the people, the two alternative techniques available are as follows:

(a) We may collect the necessary data from each one of the 1,000 people through a questionnaire containing, say, 100 questions(census method); or

(b) We may take a sample or 100 persons and prepare questionnaire containing as many as 10 questions. The expenses involved in the latter case would almost be the same as in the former but it will enable ni9ne times more information to be obtained.

(v) **Sampling Method is the only Method that can be used in Certain Cases:** There are some cases in which the census method is inapplicable and the only practicable means is provide by the sample method. For example, if one is interested in testing the breaking strength of chalks manufactured in a factory under the census method all the chalks would be broken in the possess of testing. Hence, census method is impracticable and resort must be had to the sample method. Similarly, if the producer wants to find out whether the tensile strength of a lot of steel wires meets the specified standard, he must resort to sample method because census would mean complete destruction of all the wires. Also if the population under investigation is infinite, sampling is the only possible solution.

(vi) **The sample Method is often used to Judge the Accuracy of the Information Obtained on a Census Basis:** For example, in the population census which is conducted very often the field officers employ the sample method to determine the accuracy of information obtained by the enumerators on the census basis.

**Limitation**

Despite the various advantages of sampling, it is not completely free from limitations. Some of the difficulties involved in sampling are stated below:

- A sample must be carefully planned and executed otherwise the results obtained may be inaccurate and misleading. Of course, even for a complete count care must be taken but serious errors may arise in sampling, if the sampling procedure is not perfect.
- Sampling generally requires the services of experts, if only for consultation purposed. In the absence of qualified and experienced person, the information obtained from sample surveys cannot be relied upon. In India, shortage of experts in the sampling field is serious hurdle in the way of reliable statistics.
- At times the sampling plan may be so complicated that it requires more time, labour and money than a complete count. This is so if the size of the sample is a large proportion of the total population and if complicated weighted procedures are used. With each additional complication in the survey, the chances of error multiply and greater care has to be taken, which in turn, means more time and labour.
- Of the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary.

**Sampling and Non-Sampling Errors**

To appreciate the need for sample surveys, it is necessary to understand clearly the role of sampling and non-sampling errors in complete enumeration and sample surveys. The error arising due to drawing inferences about the population on the basis of few observations is termed sampling error. Clearly, the sampling error in this sense is non-existent in complete enumeration survey, since the whole population is surveyed. However, the error mainly arising at the stage of ascertainment and processing of data, which are termed non-sampling errors, are common both in complete enumeration and sample surveys.

**Sampling Error:**

Even if utmost care has been taken in selecting a sample, the results derived from a sample study may not be exactly equal to the true value in the population. The reason is those estimates are based on a part and not in the whole and samples are seldom, if ever, perfect miniature of the population. Hence sampling gives rise to certain errors known as Sampling errors. These errors would not be present on a modern sampling theory helps in designing the errors can be controlled. The modern sampling theory helps in designing thr survey in such a manner that the sampling errors can be made small.

Sampling errors are of two types: biased and unbiased.

**(i)** **Biased Errors:** These errors arise from any bias in selection, estimation, etc. For example, if in place of simple random sampling, deliberate sampling has been used in a particular case some bias is introduced in the result and hence such errors are called biased sampling errors.

**(ii)** **Unbiased Errors:** This error arises due to chance differences between the members of population included in the sample and those not included. An error in statistics is the difference between the value of a statistic and that of the corresponding parameter.

**Causes of Bias:** Bias may arise due to:

**(i)** Faulty process of selection;

**(ii)** Faculty work during the collection; and

**(iii)** Faculty methods of analysis

(i) Faulty Selection:

Faulty selection of the sample may give rise to bias in a number of ways, such as:

(a) Deliberate selection of a ‘representative’ sample.

(b) Conscious or unconscious bias in the selection of a ‘random’ sample. The randomness of selection may not really exist, even though the investigator claims that he had a random sample if he allows his desire to obtain a certain result to influence his selection.

(c) Substitution. Substitution of an item in place of one chosen in random sample sometimes leads to bias. Thus, if it were decided to interview every 50th household in the street, it would be inappropriate to interview the 51st or any other number in his place as the characteristics possessed by them differ from those who were originally to be included in the sample.

(d) Non-response. If all the items to be included in the sample are not covered there will be bias even though no substitution has been attempted. This fault particularly occurs in mailed questionnaires, which are incompletely returned. Moreover, the information supplied by the informants may also be biased.

(e) An appeal to the vanity of the person questioned may give rise to yet another kind of bias. For example, the question ‘Are you a good student?’ is such that most of the students would succumb to vanity and answer “yes’.

(ii) Bias due to faulty collection of data any consistent error in measurement will give rise to bias whether the measurements are carried out on a sample or on all the units of the population. The danger of error is, however, likely to be greater in sampling work. Since the units measured are often smaller. Bias may arise due to improper formulation of the decision, problem or wrongly defining the population, specifying the wrong decisions, securing an inadequate frame, and so on. Biased observations may result from a poorly designed questionnaire, an ill-trained interviewer, failure of a respondent’s memory, etc. bias in the flow of data may be due to unorganised collection procedure, faulty editing or coding of responses.

### Primary Data

As discussed about, the data collected for the first time by the investigator himself or by his agents are called primary data. They are original in nature. They have to be presented, analyzed and interpreted by the researcher. Generally the data collected are purpose oriented.

### Merits of Primary data

1. If the investigation is good, the data collected will be accurate and reliable.
2. Adequate and topic specific data can be collected.
3. They are original in nature and so primary data make the knowledge world prosperous.
4. As there is personal contact between the informants and the investigator, misinterpretation can be avoided and accuracy can be enhanced.
5. Questionnaire method of data collection makes the coverage of wide areas easier.

### Demerits

1. The collection of primary data is more expensive and time consuming.
2. If requires a lot of men and material.
3. If the agents appointed for collecting information are inefficient, the data collected may be inaccurate.
4. There are changes for personal bias are prejudice.
5. If there is non- response from informants, then there will be undue delay in completing the survey.

## Tools for Collecting Primary Data

As is well known, gathering primary data is costly and time intensive. The main techniques for gathering data are observation, interviews, questionnaires, schedules, and surveys.

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc.

## Requisites of Good Questionnaire

Primary data refers to the first hand data gathered by the researcher himself. Secondary data means data collected by someone else earlier. Surveys, observations, experiments, questionnaire, personal interview, etc. Government publications, websites, books, journal articles, internal records etc.

## Sources of Secondary data

Secondary data is research data that has previously been gathered and can be accessed by researchers. The term contrasts with primary data, which is data collected directly from its source.

Secondary data is usually gathered from the published (printed) sources. A few major sources of published information are as follows: Published articles of local bodies, and central and state governments. Statistical synopses, census records, and other reports issued by the different departments of the government.

Secondary data in research methodology is any information or statistics that researchers have already collected through their primary resources. Secondary data is readily available for other individuals to reference as they conduct their own primary research, allowing them to gain insights into different processes that contribute to a research process. So, what can be primary data for one researcher may be secondary for another, depending on how they sourced it. Secondary researchers can gather data from various sources and summarise it into a new document that is easier to understand.

Secondary data can be a direct by-product of someone else's research procedures, and likely took the initial researcher significant time to develop and publish before it became readily available for other people to use. While primary data tends to be more time-consuming to gather, secondary data often requires minimal research, especially when using resources from the Internet of other digital mediums. The use of search engines and online databases has reduced the level of effort that was previously necessary for gathering large amounts of secondary data.

**Advantages of Secondary Data**

Listed are a few advantages of secondary data.

- **Easy to access**: Data is available anywhere and anytime it can be in the form of periodicals, magazines, or can be accessed anytime through the internet. People generally use secondary data maximum nowadays to evaluate their studies. A very small example is the students who depend on books, internet sites, and teachers to access information and prepare for exams.
- **Low cost or cost-effective**: The secondary data is of low cost as data are available at cheap rates, for example, the internet access, newspaper, or periodicals are available at cheaper rates and available in large quantities, so there is no non-availability of data to its users. Thus it is cost-effective.
- **Less time taking**: Data is available quickly and readily while primary data need to be collected first and then only after summarization data are used. Time taken to collect and analyze data is less than secondary data that is quickly available. Therefore it takes less time to take the source of data.

- **Various sources are available to collect data:** Secondary data is not only available through one source, but there are multiple sources like books, magazines, the internet, periodicals, and many more. Therefore various sources are available to collect data for analysis for its users. These sources are easily accessible and readily available to their users.
- **Data can be collected by anyone**: Anyone can collect data whether he /she is specialized in collecting it or not, depending upon the use. Also, there is no ownership of data that can be claimed by its user as data has already been shared by its owner, who was a primary collector of data.
- **The study is based on longitudinal analysis:** Since the data has been collected over years, thus a longitudinal analysis is done by the researchers with the help of secondary data. The data collected is more reliable and valid for users.

**Disadvantages of Secondary Data**

Listed are a few disadvantages of secondary data.

1. **Inaccuracy**: It is a limitation of secondary data that the data collected over the past few years may be inaccurate. The basis of data collected may not be correct or the analysis or interpretation made may not be accurate or relevant.
2. **Data may be sometimes outdated:** The data provided through different sources may also be outdated as it has been stored and managed for many years. Therefore it may also sometimes be outdated and may not be relevant for today’s scenario.
3. **Not compatible with the needs of the user:** Since data is related to past surveys and according to the needs of the researchers of that time. It may happen that the present user of this data may not need or not have topics relevant to his study or research. Therefore here instead of outdated data, the data becomes irrelevant for the user to be used in research.

4. **Anyone can access data**: There is no privatization of data by its owner, data can be accessed by anyone willing to research on that topic. There is no secrecy of data but the user of data cannot appeal their possession or ownership of the data they accessed.
5. **Data quality cannot be controlled:** The researchers have no control over the quality of data. As data is already surveyed by researchers according to their relevant basis and there may be changes in the surroundings and other factors that may lead to the change in the data provided thus no proper quality can be controlled.
6. **Data can be biased**: Since data collected by the researcher is based on his/her opinion, therefore data is biased. And it may also have an impact on the data collected by the user of the secondary data.

### Tools for Data Collection

Methods of collecting **primary data**, particularly in surveys and descriptive researches. important ones are: (i) observation method, (ii) interview method, (iii) through questionnaires,(iv) through schedules, and (v) other methods which include (a) warranty cards; (b) distributor audits; (c) pantry audits; (d) consumer panels; (e) using mechanical devices; (f) through projective techniques; (g) depth interviews, and (h) content analysis.

**Secondary data** means data that are already available i.e., they refer to the data which have already been collected and analysed by someone else. When the researcher utilises secondary data, then he has to look into various sources from where he can obtain them. In this case he is certainly not confronted with the problems that are usually associated with the collection of original data. Secondary data may either be published data or unpublished data. Usually published data are available in: (a) various publications of the central, state are local governments; (b) various publications of foreign governments or of international bodies and their subsidiary organisations; (c) technical and trade journals; (d) books, magazines and newspapers; (e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (f) reports prepared by research scholars, universities, economists, etc. in different fields; and (g) public records and statistics, historical documents, and other sources of published information. The sources of unpublished data are many; they may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/ private individuals and organisations.

**Questionnaires and Schedules**

Both questionnaire and schedule are popularly used methods of collecting data in research surveys. The important points of difference are as under:

1. The questionnaire is generally sent through mail to informants to be answered as specified in a covering letter, but otherwise without further assistance from the sender. The schedule is generally filled out by the research worker or the enumerator, who can interpret questions when necessary.

2. To collect data through questionnaire is relatively cheap and economical since we have to spend money only in preparing the questionnaire and in mailing the same to respondents. To collect data through schedules is relatively more expensive.

3. Non-response is usually high in case of questionnaire as many people do not respond and many return the questionnaire without answering all questions. Bias due to non-response often remains indeterminate. As against this, non-response is generally very low in case of schedules because these are filled by enumerators who are able to get answers to all questions.

4. In case of questionnaire, it is not always clear as to who replies, but in case of schedule the identity of respondent is known.

5. The questionnaire method is likely to be very slow since many respondents do not return the questionnaire in time despite several reminders, but in case of schedules the information is collected well in time as they are filled in by enumerators.

6. Personal contact is generally not possible in case of the questionnaire method as questionnaires are sent to respondents by post who also in turn return the same by post. But in case of schedules direct personal contact is established with respondents.

7. Questionnaire method can be used only when respondents are literate and cooperative, but in case of schedules the information can be gathered even when the respondents happen to be illiterate.

8. Wider and more representative distribution of sample is possible under the questionnaire method, but in respect of schedules there usually remains the difficulty in sending enumerators over a relatively wider area.

9. Risk of collecting incomplete and wrong information is relatively more under the questionnaire method, particularly when people are unable to understand questions properly. But in case of schedules, the information collected is generally complete and accurate as enumerators can remove the difficulties, if any, faced by respondents in correctly understanding the questions.

10. The success of questionnaire method lies more on the quality of the questionnaire itself, but in the case of schedules much depends upon the honesty and competence of enumerators.

11. In order to attract the attention of respondents, the physical appearance of questionnaire must be quite attractive, but this may not be so in case of schedules as they are to be filled in by enumerators and not by respondents.

12. Along with schedules, observation method can also be used but such a thing is not possible while collecting data through questionnaires.

# UNIT III
# RESEARCH DESIGN

## Data processing

The data, after collection, has to be processed and analyzed in accordance with the outline laid down for the purpose at the time of developing the research plan. This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis. Technically speaking, processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. The term analysis refers to the computation of certain measures along with searching for patterns of relationship that exist among data-groups. Thus, "in the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions".

## Editing :

Editing of data is a process or examination the collected raw data to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and or schedules. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as completed as possible and have been well arranged to facilitate coding and tabulation./

Central editing should take place when all forms of schedules have been completed and returned to the office. This type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a large inquiry. Editor may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule. At times, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final results, especially in the context of mail surveys.

**Coding:**

Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to be observed is that of unidimensionality by which is meant that every class is defined in terms of only one concept.

Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis. Coding decisions should usually be taken at the designing stage of the questionnaire. This makes it possible to precode the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch form the original questionnaires.

**Classification:**

Most research studies result in a large volume of raw data which must be reduced into homogeneous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes.

Classification can be one of the following two types, depending upon the phenomenon involved.

**Classification:**

Classification is used to group similar data into categories, while tabulation is used to present data in a structured and organized manner. Method: Classification involves grouping data based on certain criteria, while tabulation involves organizing data into rows and columns.

**Types**

Spatial Classification of Data and Tabular Presentation
Classification of Data and Tabular Presentation
Quantitative Classification of Data and Tabular Presentation
Temporal Classification of Data and Tabular Presentation
Qualitative Classification of Data and Tabular Presentation

Classification of data is also used in tabular presentation and is of four types; viz., Geographical or Spatial Classification, Chronological or Temporal Classification, Qualitative Classification, and Quantitative Classification.

**Frequency Distribution**

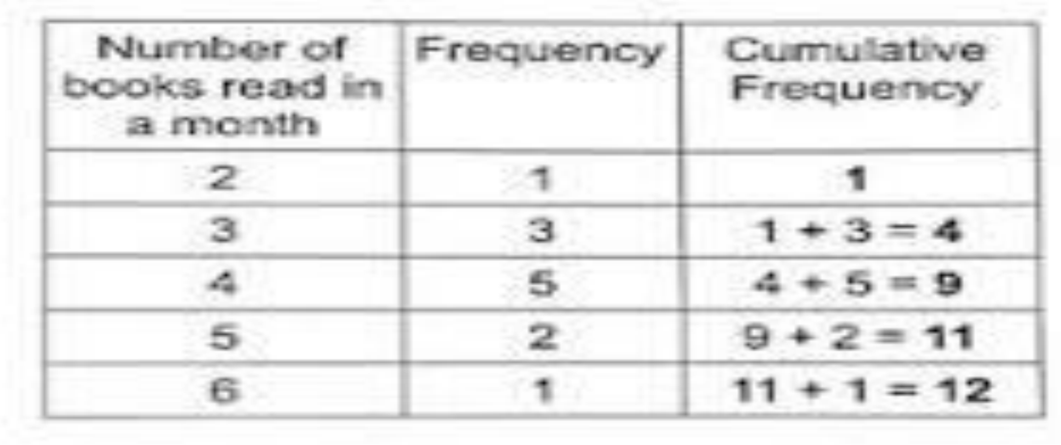
A frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The frequency is how often a value occurs in an interval while the distribution is the pattern of frequency of the variable.

**Cumulative Frequency Distribution**

What is Cumulative Frequency Distribution? Cumulative frequency distribution is a form of frequency distribution that represents the sum of a class and all classes below it.

Remember that frequency distribution is an overview of all distinct values (or classes of values) and their respective number of occurrence.

**Example of cumulative frequency**



| Number of books read in a month | Frequency | Cumulative Frequency |
|---|---|---|
| 2 | 1 | 1 |
| 3 | 3 | 1 + 3 = 4 |
| 4 | 5 | 4 + 5 = 9 |
| 5 | 2 | 9 + 2 = 11 |
| 6 | 1 | 11 + 1 = 12 |

Here is an example of a cumulative frequency table. The cumulative frequency table shows how many books a student read each month over a one-year period. Look at the table to see the student read 2 books in one month, 3 books in three of the months, 4 books in five of the months, and so on.

**Class Interval**

Class interval refers to the numerical width of any class in a particular distribution. In maths, class interval is defined as the difference between the upper class limit and the lower class limit. The size of the class into which a particular data is divided. Eg. divisions on a histogram or bar graph.

**Tabulation:**

When a mass of data has been assembled, it become necessary for the researcher to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarizing raw data and displaying the same in compact form for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows.

Tabulation is essential because of the following reasons.

(1) It conserves space and reduces explanatory and descriptive statement to a minimum.

(2) It facilitates the process of comparison.

(3) It facilitates the summation of items and the detection of errors and omissions.

(4) It provides a basis for various statistical computations.

Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulation machine or computers. In relatively large inquiries, we may use machines or computer tabulation if other factors are favourable and necessary facilities are available. Under this method, the codes are written on a sheet of paper, called tally sheet, and for each response a stroke is marked against the code in which it falls. Usually after every four strokes against a particular code, the fifth response is indicated by drawing a diagonal or horizontal line through the stroke. Tabulation may also be classified as simple and complex tabulation.

**Generally accepted principles of tabulation:**

Such principles of tabulation, particularly of construction statistical tables, can be briefly states as follows:

(1) Every table should have a clear, concise and adequate title so as to make that table intelligible without reference it the text and this title should always be placed just above the body of the table.

(2) Every table should be given a distinct number to facilities easy reference.

(3) The column headings and the row headings of the table should be clear and brief.

(4) The units of measurement under each heading or sub-heading must always be indicated.

(5) Explanatory footnotes, if any, concerning the table should be placed directly beneath the table, along with the reference symbols used in the table.

(6) Source or source from where the data in the table have been obtained must be indicated just below the table.

(7) Usually the columns are separated from one another by lines which make the table more readable and attractive. Lines are always drawn at the top and bottom of the table and below the captions.

(8) There should be thick lines to separate the data under one class from the data under another class and the lines separating the sub-divisions of the classes should be comparatively thin lines.

(9) The column may be numbered to facilitate reference.

(10) Those columns whose data are to be compared should be kept side by side. Similarly, percentages and\or average must also be kept close to the data.

(11) It is generally considered better to approximate figures before tabulation as the same would reduce unnecessary details in the table itself.

(12) In order to emphasize the relative significance of certain categories, different kinds of type, spacing and indentations may be used.

(13) It is important that all column figures be properly aligned. Decimal points and (+) or (-) signs should be in perfect alignment.

(14) Abbreviations should be avoided to the extent possible and ditto marks should not be used in the table.

(15) Miscellaneous and exceptional items, if any, should be usually placed in the last row of the table.

(16) Table should be made as logical, clear, accurate and simple as possible. If the data happen to be very large, they should not be crowded in a single table for that would make that table unwieldy and inconvenient.

(17) Total of rows should normally be placed in the extreme right column and that of columns should be placed at the bottom.

(18) The arrangement of the categories in a table may be chronological, geographical, alphabetical or according to magnitude to facilitate comparison. Above all, the table must suit the needs and requirements of an investigation.

**Diagram**

(a) One – dimensional diagram
(b) Two – dimensional diagram
(c) Three – dimensional diagram
(d) Pictograms
(e) Cartograms

**Graphical Representation**

Graphical representation is a form of visually displaying data through various methods like graphs, diagrams, charts, and plots. It helps in sorting, visualizing, and presenting data in a clear manner through different types of graphs. Statistics mainly use graphical representation to show data.

**The four different types of graphical representation**

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot.

**Graphical data display**

Graphical displays communicate comparisons, relationships, and trends. They emphasize and clarify numbers. To choose the appropriate type of display, first define the purpose of the report, and then identify the most effective display to suit that purpose.

## Histogram

Histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears).

## Difference between Bar Graph and Histogram

A histogram is one of the most commonly used graphs to show the frequency distribution. As we know that the frequency distribution defines how often each different value occurs in the data set. The histogram looks more similar to the bar graph, but there is a difference between them. The list of differences between the bar graph and the histogram is given below:

| Histogram | Bar Graph |
|---|---|
| It is a two-dimensional figure | It is a one-dimensional figure |
| The frequency is shown by the area of each rectangle | The height shows the frequency and the width has no significance. |
| It shows rectangles touching each other | It consists of rectangles separated from each other with equal spaces. |

## Types of Histogram

The histogram can be classified into different types based on the frequency distribution of the data. There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, and so on. The histogram can be used to represent these different types of distributions. The different types of a histogram are:

- Uniform histogram
- Symmetric histogram

- Bimodal histogram
- Probability histogram

**Uniform Histogram**

A uniform distribution reveals that the number of classes is too small, and each class has the same number of elements. It may involve distribution that has several peaks.

**Symmetric Histogram**

A symmetric histogram is also called a bell-shaped histogram. When you draw the vertical line down the center of the histogram, and the two sides are identical in size and shape, the histogram is said to be symmetric. The diagram is perfectly symmetric if the right half portion of the image is similar to the left half. The histograms that are not symmetric are known as skewed

**Bimodal Histogram**

If a histogram has two peaks, it is said to be bimodal. Bimodality occurs when the data set has observations on two different kinds of individuals or combined groups if the centers of the two separate histograms are far enough to the variability in both the data sets.

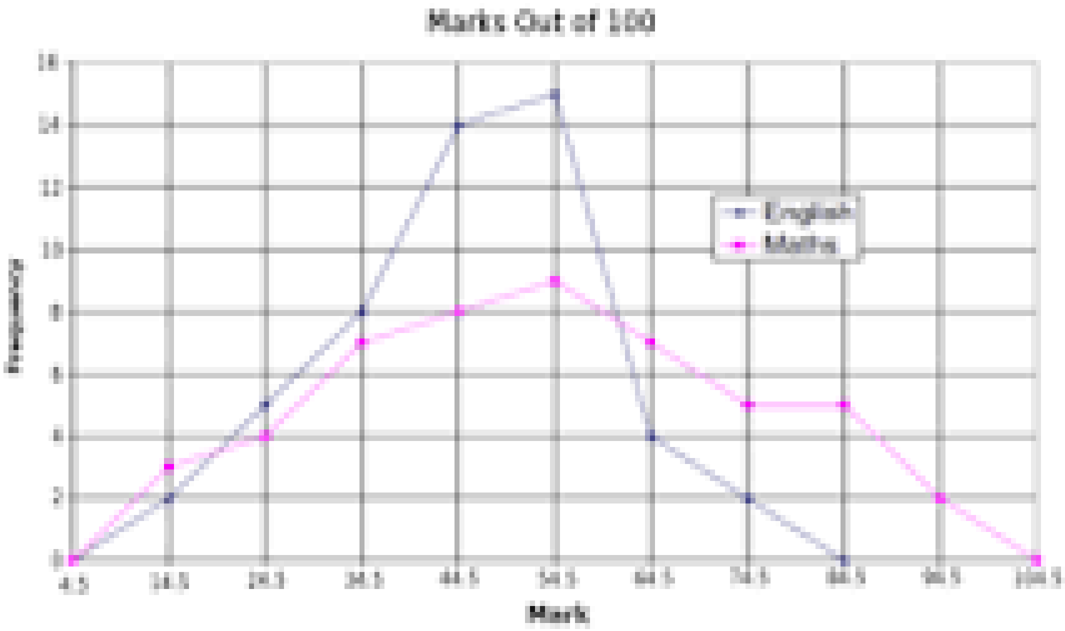
**Probability Histogram**

A Probability Histogram shows a pictorial representation of a discrete probability distribution. It consists of a rectangle centered on every value of x, and the area of each rectangle is proportional to the probability of the corresponding value. The probability histogram diagram is begun by selecting the classes. The probabilities of each outcome are the heights of the bars of the histogram.

**Frequency Polygon**

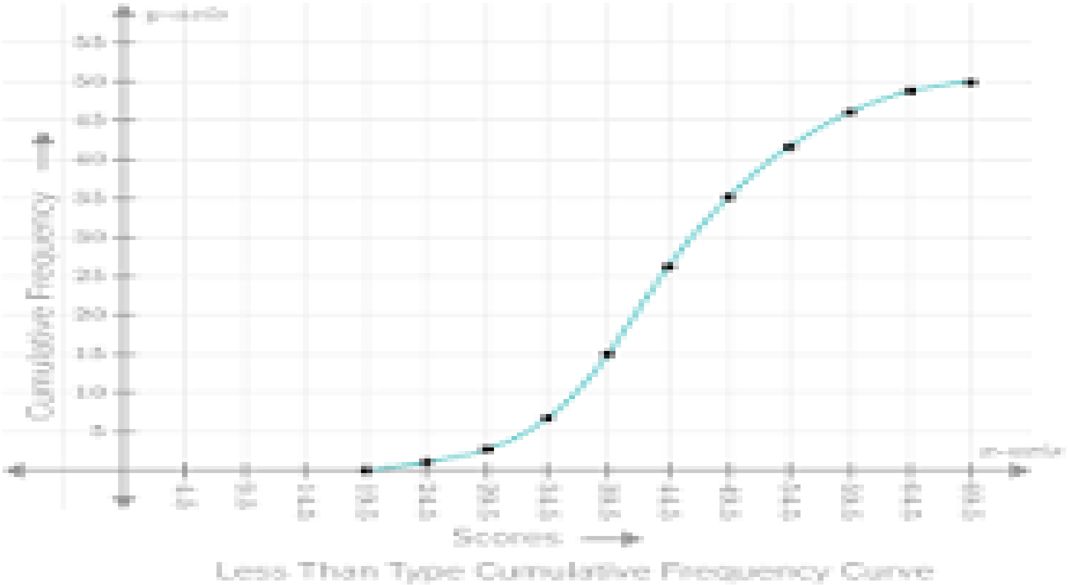
A frequency polygon is a line graph of class frequency plotted against class midpoint. It can be obtained by joining the midpoints of the tops of the rectangles in the histogram.

The percentage of a frequency polygon



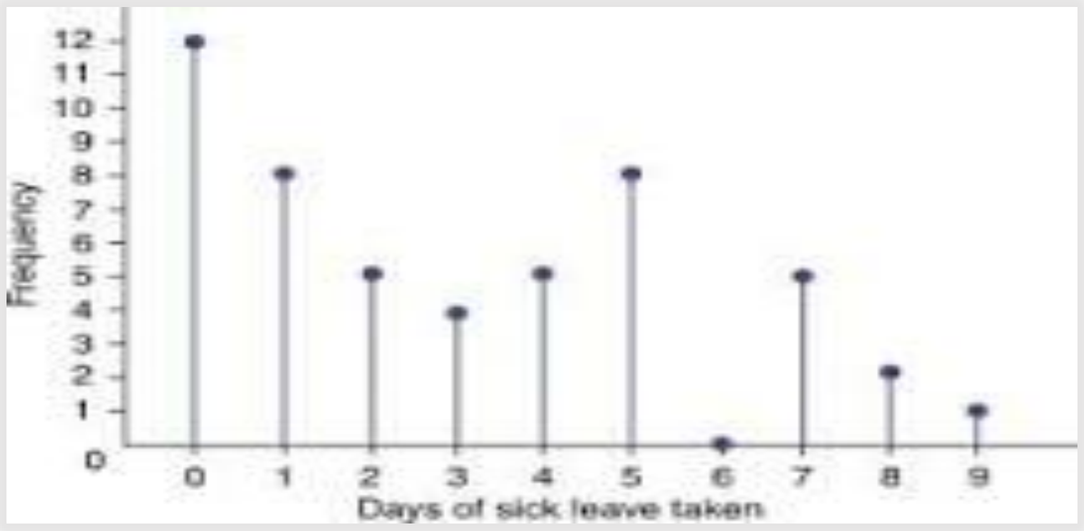
To construct a relative frequency polygon:Sum the number of points in each interval, divide the sum of each interval by the total number of data points, and multiply by 100. The result is the percentage of the total number of data points that is represented by each interval.

**Community frequency polygon**



The cumulative frequency polygon is essentially a line graph drawn on graph paper by plotting actual lower or upper limits of the class intervals on the -axis and the respective cumulative frequencies of these class intervals on the –axis.

**Frequency table**



A frequency table is simply a “t-chart” or two-column table which outlines the various possible outcomes and the associated frequencies observed in a sample.

**Ogive Curve**

The Ogive is defined as the frequency distribution graph of a series. The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative per cent frequencies on the vertical axis.

**Ogive Graph**

The graphs of the frequency distribution are frequency graphs that are used to exhibit the characteristics of discrete and continuous data. Such figures are more appealing to the eye than the tabulated data. It helps us to facilitate the comparative study of two or more frequency distributions. We can relate the shape and pattern of the two frequency distributions.

The two methods of Ogives are:

- Less than Ogive
- Greater than or more than Ogive

### Less than Ogive

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

### Greater than or More than Ogive

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class, second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

### Uses of Ogive Curve

Ogive Graph or the cumulative frequency graphs are used to find the median of the given set of data. If both, less than and greater than, cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which, both the curve intersects, corresponding to the x-axis, gives the median value.  Apart from finding the medians, Ogives are used in computing the percentiles of the data set values.

### Lorenz Curve

A **Lorenz curve**, developed by American economist Max Lorenz in 1905, is a graphical representation of income inequality or wealth inequality.

### Lorenz curve

A Lorenz curve is a graphical representation of income inequality or wealth inequality. The graph plots percentiles of the population on the horizontal axis according to income or wealth. Lorenz curve represents the distribution of income in an economy. It is represented by a straight line that depicts the perfect distribution of income. Lorenz curve is beneath that line which shows the estimated distribution of income.

### Merits of Lorenz curve

The Lorenz curve shows how income or wealth is distributed among a population. It plots the cumulative percentage of people on the x-axis and the cumulative percentage of income or wealth on the y-axis. The straight diagonal line represents perfect equality, where everyone has the same share of income or wealth.

### Demerits of Lorenz curve

It is not possible to determine which distribution has more inequality. In the lifetime of an individual, there will be variation in income and this variation is not taken into consideration when inequality in the Lorenz Curves is analyzed. These are the limitations of the Lorenz Curves.

### Interpretation and Report Writing

Interpretation refers to the task of drawing inferences from the collected facts after an analytical and/or experimental study. In fact, it is a search for broader meaning of research findings. The task of interpretation has two major aspects viz., (i) the effort to establish continuity in research through linking the results of a given study with those of another, and (ii) the establishment of some explanatory concepts. “In one sense, interpretation is concerned with relationships within the collected data, partially overlapping analysis.

### Technique of Interpretation

The task of interpretation is not an easy job, rather it requires a great skill and dexterity on the part of researcher. Interpretation is an art that one learns through practice and experience. The researcher may, at times, seek the guidance from experts for accomplishing the task of interpretation.

The technique of interpretation often involves the following steps:

(i) Researcher must give reasonable explanations of the relations which he has found and he must interpret the lines of relationship in terms of the underlying processes and must try to find out the thread of uniformity that lies under the surface layer of his diversified research findings.

(ii) Extraneous information, if collected during the study, must be considered while interpreting the final results of research study.

(iii) It is advisable, before embarking upon final interpretation, to consult someone having insight into the study and who is frank and honest and will not hesitate to point out omissions and errors in logical argumentation.

(iv) Researcher must accomplish the task of interpretation only after considering all relevant factors affecting the problem to avoid false generalization.

**Different Steps in Writing Report**

Research reports are the product of slow, painstaking, accurate inductive work. The usual steps involved in writing report are: (a) logical analysis of the subject-matter; (b) preparation of the final outline; (c) preparation of the rough draft; (d) rewriting and polishing; (c) preparation of the final bibliography; and (f) writing the final draft. Though all these steps are self explanatory, yet a brief mention of each one of these will be appropriate for better understanding.

**Logical analysis of the subject matter**: It is the first step which is primarily concerned with the development of a subject. There are two ways in which to develop a subject (a) logically and (b) chronologically. The logical development is made on the basis of mental connections and associations between the one thing and another by means of analysis..

**Preparation of the final outline**: It is the next step in writing the research report "Outlines are the framework upon which long written works are constructed. They are an aid to the logical organisation of the material and a reminder of the points to be stressed in the report."

Preparation of the rough draft: This follows the logical analysis of the subject and the preparation of the final outline. Such a step is of utmost importance for the researcher now sits to write down what he has done in the context of his research study.

**Rewriting and polishing of the rough draft**: This step happens to be most difficult part of all formal writing. Usually this step requires more time than the writing of the rough draft. The careful revision makes the difference between a mediocre and a good piece of writing. While rewriting and polishing, one should check the report for weaknesses in logical development or presentation. He should check the mechanics of writing—grammar, spelling and usage.

**Preparation of the final bibliography**: Next in order comes the task of the preparation of the final bibliography. The bibliography, which is generally appended to the research report, is a list of books

**Writing the final draft***:* This constitutes the last step. The final draft should be written in a concise and objective style and in simple language, avoiding vague expressions such as "it seems", "there may be", and the like ones. While writing the final draft, the researcher must avoid abstract terminology and technical jargon. A research report should not be dull, but must enthuse people and maintain interest and must show originality. It must be remembered that every report should be an attempt to solve some intellectual problem and must contribute to the solution of a problem and must add to the knowledge of both the researcher and the reader.

## Preparing a Research Proposal

Title and abstract

The title should be concise and descriptive, and the abstract should summarize the research question, objectives, methodology, and expected outcomes.

Introduction

The introduction should introduce the topic, provide background and context, and outline the problem statement and research questions. It should also demonstrate the relevance of the research and show that it is interesting, original, and important.

Literature review

The literature review should synthesize prior research and show how the proposed research is original and adds to current knowledge.

Research methodology

The research methodology should be clearly and logically written and organized. It should describe the proposed research methodology, data analysis tools, and other details.

Research design and type

Explain if the research is experimental, correlational, or descriptive. If conducting social science research, describe the demographic being studied.

Data collection

Explain how the subjects will be chosen and data will be collected from them.

Tools

Explain the tools that will be used to conduct the research, such as surveys, interviews, or videos.

Budget and time frame

Include information about the budget and time frame for the research.

References and citations

Include a full and accurate list of all sources used to write the proposal.

A project proposal is a detailed pitch for a project idea that explains why the project is important, how it will be done, who is involved, and how much it will cost. The goal is to convince people to support the project by showing that it is well thought out and worth investing in

# UNIT IV

# DATA ANALYSIS - I

## Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error.

## Probability Sampling Types

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

## Importance in sampling theory

The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these "important" values are emphasized by sampling more frequently, then the estimator variance can be reduced.

## Sampling Distribution

A sampling distribution isa probability distribution of a statistic that is obtained through repeated sampling of a specific population. It describes a range of possible outcomes for a statistic, such as the mean or mode of some variable, of a population.

## Types of Sampling Distributions

Here is a brief description of the types of sampling distributions:

- **Sampling Distribution of the Mean:** This method shows a normal distribution where the middle is the mean of the sampling distribution. As such, it represents the mean of the overall population. In order to get to this point, the researcher must figure out the mean of each sample group and map out the individual data.

- **Sampling Distribution of Proportion:** This method involves choosing a sample set from the overall population to get the proportion of the sample. The mean of the proportions ends up becoming the proportions of the larger group.
- **T-Distribution:** This type of sampling distribution is common in cases of small sample sizes. It may also be used when there is very little information about the entire population. T-distributions are used to make estimates about the mean and other statistical points.

**Parameter and Statistic**

A parameter is a number describing a whole population (e.g., population mean), while a statistic is a number describing a sample (e.g., sample mean).

**The difference between parameter and statistic**

The key difference between parameters and statistics is that parameters describe populations, while statistics describe samples. You can easily remember this distinction using the alliterations for population, parameter, and sample statistic.

| Points | Statistic | Parameter |
|---|---|---|
| 1 | Derived from sample data | Derived from population data |
| 2 | Used to estimate population characteristics | Represents population characteristics |
| 3 | Subject to sampling variability | Fixed value |
| 4 | Provides information about a sample | Provides information about a population |
| 5 | Varied values across different samples | Consistent value for the entire population |
| 6 | Estimation based on inference techniques | Known or can be determined with complete data |
| 7 | Used to draw conclusions about a population | Describes a population |
| 8 | Often denoted using Greek letters | Often denoted using English letters |
| 9 | Can change with different samples | Remains constant for a specific population |
| 10 | Used in hypothesis testing and confidence intervals | Used in defining populations and subgroups |

### Two Types of Errors in testing of Hypothesis

1. The hypothesis is true but our test rejects it (Type I error).
2. The hypothesis is false but our test accepts it (Type II error).
3. The hypothesis is true and our test accepts it (Correct Decision).
4. The hypothesis is false and our test rejects it (Correct Decision).

### Type I and Type II Errors

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

### Table of Type I and Type II Error

The relationship between truth or false of the null hypothesis and outcomes or result of the test is given in the tabular form:

| Error Types | When $H_0$ is True | When $H_0$ is False |
|---|---|---|
| **Don't Reject** | Correct Decision (True negative)<br>Probability = 1 – α | Type II Error (False negative)<br>Probability = β |
| **Reject** | Type II Error (False Positive)<br>Probability = α | Correct Decision (True Positive)<br>Probability = 1 – β |

### Type I and Type II Errors Example

Check out some real-life examples to understand the type-I and type-II error in the null hypothesis.

**Example 1**: Let us consider a null hypothesis – A man is not guilty of a crime.

Then in this case:

| Type I error (False Positive) | Type II error (False Negative) |
|---|---|
| He is condemned to crime, though he is not guilty or committed the crime. | He is condemned not guilty when the court actually does commit the crime by letting the guilty one go free. |

**Example 2:** Null hypothesis- A patient's signs after treatment A, are the same from a place box.

| Type I error (False Positive) | Type II error (False Negative) |
|---|---|
| Treatment A is more efficient than the placebo | Treatment A is more powerful than placebo even though it truly is more efficient. |

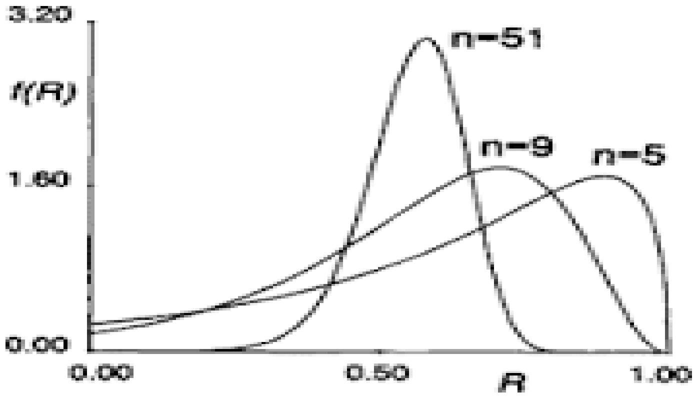
**Level of Significance**

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

**The 4 levels of significance**

The null hypothesis is a hypothesis that states that the data has no result, association between variables, or disparity between variables. There are four levels in statistics that are organized by level of complexity and precision. They are nominal, ordinal, interval, and ratio.

**5 -significance level**

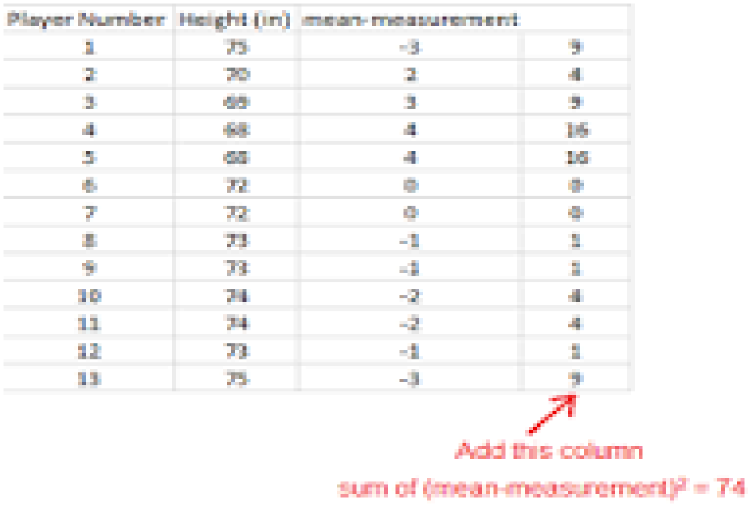


The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, α = 0.05, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting $H_0$ when it is true.

**STANDARD ERROR**

Standard error is the approximate standard deviation of a statistical sample population. The standard error describes the variation between the calculated mean of the population and one which is considered known, or accepted as accurate.

**An example of using standard error**



| Player Number | Height (in) | mean- measurement | |
|---|---|---|---|
| 1 | 75 | -3 | 9 |
| 2 | 70 | 2 | 4 |
| 3 | 69 | 3 | 9 |
| 4 | 68 | 4 | 16 |
| 5 | 68 | 4 | 16 |
| 6 | 72 | 0 | 0 |
| 7 | 72 | 0 | 0 |
| 8 | 73 | -1 | 1 |
| 9 | 73 | -1 | 1 |
| 10 | 74 | -2 | 4 |
| 11 | 74 | -2 | 4 |
| 12 | 73 | -1 | 1 |
| 13 | 75 | -3 | 9 |

Add this column
sum of (mean-measurement)$^2$ = 74

For example, you would construct a 95% confidence interval by adding and subtracting 1.96 times the standard error from the sample mean. Therefore, the 95% confidence interval for high school basketball player height would be 70.65 inches to 73.35 inches.

**Standard Error**

$$SE = \frac{\sigma}{\sqrt{n}} \quad \begin{array}{l}\longleftarrow \text{Standard deviation}\\ \longleftarrow \text{Number of samples}\end{array}$$

Standard error is calculated by dividing the standard deviation of the sample by the square root of the sample size. Calculate the mean of the total population. Calculate each measurement's deviation from the mean.

**The symbol for standard error**

The standard error of a statistic is usually designated by the Greek letter sigma (σ) with a subscript indicating the statistic. For instance, the standard error of the mean is indicated by the symbol: $\sigma_M$.

**Properties of Good Estimator**

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

**An example of estimate**

We need to estimate how much paint we'll need for the job. The cost of the project has been estimated at/as about 10 million dollars. He estimates that current oil reserves are 20 percent lower than they were a year ago. Damage from the hurricane is estimated (to be) in the billions of dollars.

**Testing of Hypothesis**

Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

There are three types of hypothesis tests: right-tailed, left-tailed, and two-tailed. When the null and alternative hypotheses are stated, it is observed that the null hypothesis is a neutral statement against which the alternative hypothesis is tested.

**Procedure of Testing Hypothesis**

The procedure of testing hypothesis is as follows:

**1. Set up a hypothesis**

The null hypothesis can be thought of as the opposite of the "guess" the researchers made: in this example, the biologist thinks the plant height will be different for the fertilizers. So the null would be that there will be no difference among the groups of plants. Specifically, in more statistical language the null for an ANOVA is that the means are the same.

**2. Set up a suitable significance level**

The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, $\alpha = 0.05$, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting $H_0$ when it is true.

**3. Setting a test criterion**

This involves selecting an appropriate probability distribution for the particular test, that is, a probability distribution which can properly be applied.

**4. Doing Computations**

A *computation* is any type of arithmetic or non-arithmetic *calculation* that is well-defined. Common examples of *computations* are mathematical equations.

**5. Making Decisions**

Finally as a fifth step, we may conclude statistical conclusions and take decisions. A statistical conclusions or statistical decision is a decision either to reject or to accept the null hypothesis.

**The steps of testing hypothesis**

**Table of contents**

- Step 1: State your null and alternate hypothesis.
- Step 2: Collect data.
- Step 3: Perform a statistical test.
- Step 4: Decide whether to reject or fail to reject your null hypothesis.
- Step 5: Present your findings.

**The advantages of hypothesis**

A hypothesis can help you to formulate a specific and testable research problem, and to design an appropriate method to collect and analyze data. A hypothesis can also help you to establish a clear direction and focus for your research, and to communicate your expectations and assumptions to your readers or audience.

**Hypothesis test to use**

A z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

# UNIT V

# DATA ANALYSIS – II

## Difference between Large and small samples

When the sample size is under 30, statisticians are supposed to use the Student T distribution instead. It has a much greater chance of being wrong. In statistical context, a sample is considered to be large if it is at least 30. On the other hand, a sample is considered small is it is less than 30.

The difference between a small sample size and a large sample size lies in the number of observations or data points included in each sample. Here's a comparison of the characteristics of small and large sample sizes:

Small Sample Size:

**1. Limited Representation**: A small sample size may not fully represent the population from which it is drawn. It may not capture the full range of variability and characteristics present in the population.

**2. Higher Sampling Error**: Small samples tend to have higher sampling error or variability. The observed data points may deviate more from the true population values, leading to less precise estimates or conclusions.

**3. Reduced Statistical Power**: Small samples may have lower statistical power, making it more challenging to detect significant effects or relationships. This can increase the likelihood of Type II errors (failing to detect true effects).

**4. Narrow Confidence Intervals**: With smaller sample sizes, the confidence intervals around estimates or statistical parameters tend to be wider, reflecting increased uncertainty or imprecision in the results.

**5. Limited Generalizability**: The findings or conclusions from a small sample may have limited generalizability to a larger population. The relationships or patterns observed in the small sample may not hold true for the broader population.

**Large Sample Size:**

**1. Improved Representation**: A large sample size is more likely to capture the characteristics and variability of the population. It provides a better representation of the overall population, leading to more reliable and generalizable results.

**2. Lower Sampling Error**: Large samples tend to have lower sampling error or variability. The observed data points are closer to the true population values, resulting in more precise estimates and reduced random fluctuations.

**3. Higher Statistical Power**: Large samples have higher statistical power, enabling a better chance of detecting significant effects or relationships. This reduces the risk of Type II errors.

**4. Narrow Confidence Intervals**: With larger sample sizes, the confidence intervals around estimates or statistical parameters tend to be narrower, indicating greater precision and reduced uncertainty in the results.

**5. Enhanced Generalizability**: Findings from a large sample are more likely to be generalizable to the broader population, increasing the confidence in applying the conclusions to a wider context.

In summary, larger sample sizes generally offer more accurate and reliable estimates, higher statistical power, and increased generalizability. However, the appropriate sample size depends on the research question, desired level of accuracy, effect size, variability, and statistical techniques employed. Statisticians employ power analysis and other methods to determine the sample size needed for specific research studies.

**Test of Significance for Large Samples**

Before moving to large samples, test of significance has to be seen in brief. Test of significance is performed after framing the hypothesis (tentative statements) at say 1%, 5%

and 10% level. The level of significance (denoted as α or alpha) represents the probability of error or chances of making wrong decisions.

**Statistical test is used for large sample size**

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

**Test for Two Means and Standard Deviations**

The 2-Sample Standard Deviation test compares the standard deviations of 2 samples, and the Standard Deviations test compares the standard deviations of more than 2 samples. In this paper, we refer to k-sample designs with k = 2 as 2- sample designs and k-sample designs with k > 2 as multiple-sample designs.

## Normal Distribution:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The binominal and the Poisson distribution described above are the most useful theoretical distributions for discrete variables, i.e., they relate to the occurrence of distinct events. In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, a continuous distribution in needed. The normal distribution, also called the normal probability distribution, happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. In fact normal distribution is the cornerstone of modern statistics.

### Importance of the Normal Distribution

The normal distribution has been long occupied a central place in the theory of statistics. Its importance will be clear from the following points.

1. The normal distribution has the remarkable property stated in the socalled central limit theorem. According to this theorem as the sample size n increases the distribution of mean, $\bar{X}$ of a random sample taken from practically any population approaches a normal distribution.
2. As n becomes large the normal distribution serves as a good approximation of many discrete distributions whenever the exact discrete probability is laborious to obtain or impossible to calculate accurately.
3. In theoretical statistic many problems can be solved only under the assumption of a normal population
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. The normal distribution is used extensively in statistical quality control in industry in setting up of control limits.

## Properties of Normal Distribution:

The following are the important properties.

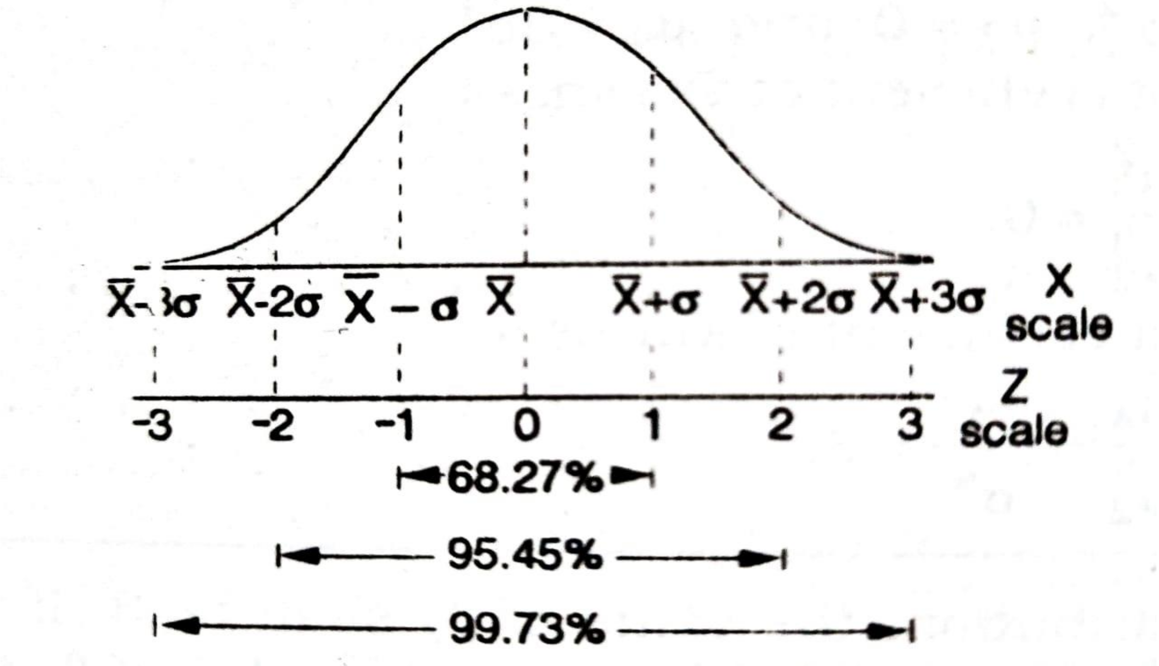
1. The normal curve is bell-shaped curve and symmetrical in its appearance. If the curves were folded along its vertical axis, the two halves would coincide.
2. The height of the normal curve is at its maximum at the mean. Hence, the mean and mode of the normal distribution coincide. Thus mean, median and mode are equal.
3. There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in either direction from the mean.

4. The curve approaches nearer and nearer to the base but it never touches it, i.e., the curve is asymptotic to the base on either side. Hence its range is unlimited or infinite in both directions.
5. Since there is only one maximum point, the normal curve is unimodal, i.e., it has only one mode.
6. The points of inflexion, I, e., the points where the change in curvature occurs are $\bar{X} \pm \sigma$.
7. As distinguished from Binomial and Poisson distribution where the variable is discrete, the variable distributed according to the normal curve is a continuous one.
8. The first and third quartiles are equidistant from the median.
9. The mean deviation is 4th or more precisely 0.7979 of the standard deviation.
10. The area under the normal curve distributed as follows:

$a) Mean \pm 1\sigma \text{ cov} ers\, 68.27\%\, area; 34.135\%\, area\, will\, lie\, on\, either\, side\, of\, the\, mean$

$b) Mean \pm 2\sigma \text{ cov} ers\, 95.45\%\, area$

$c) Mean \pm 3\sigma \text{ cov} ers\, 99.73\%\, area$



## Significance of the Normal Distribution

Normal distribution is mostly used for the following purposes.

1. To approximate of fit a distribution of measurement under certain conditions.
2. To approximate the binomial distribution and other discrete of continuous probability distribution under suitable condition.
3. To approximate the distribution of means and certain other quantities calculated from samples, especially large samples.

**The t-test for means and standard deviation**

The t-test is a test used for hypothesis testing in statistics. Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values. T-tests can be dependent or independent.

**Used for standard deviation**

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

**The t-test with means**

A t test is used to measure the difference between exactly two means. Its focus is on the same numeric data variable rather than counts or correlations between multiple variables.

**Proportion and Confidence of Fit**

Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers. The build a confidence interval for population proportion p, we use: ˆp–zα2·√ˆp(1−ˆp)n<p<ˆp+zα2·√ˆp(1−ˆp)n.

Therefore, the 99% confidence interval is 0.37 to 0.43. That is, we are 99% confident that the true proportion is in the range 0.37 to 0.43.

### Small sample Test

t, and F $\chi$ -tests are some commonly used small sample tests. Unit which are based on $\chi 2$ and F-distributions described in Unit 3 and Unit 4 of this course respectively. This unit is divided into eight sections.

### Best for small sample size

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

### Small sample size sampling

The size of the sample is small when compared to the size of the population. When the target population is less than approximately 5000, or if the sample size is a significant proportion of the population size, such as 20% or more, then the standard sampling and statistical analysis techniques need to be changed.

### Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Correlation refers to the statistical relationship between two entities. In other words, it's how two variables move in relation to one another. Correlation can be used for various data sets, as well. In some cases, you might have predicted how things will correlate, while in others, the relationship will be a surprise to you. It's important to understand that correlation does not mean the relationship is causal.

To understand how correlation works, it's important to understand the following terms:

**Positive Correlation**

A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.

**Negative Correlation**

A negative correlation is -1. This means the two variables moved in opposite directions.

**Zero or no Correlation**:

A correlation of zero means there is no relationship between the two variables. In other words, as one variable moves one way, the other moved in another unrelated direction.

**Types of correlation coefficients**

While correlation studies how two entities relate to one another, a correlation coefficient measures the strength of the relationship between the two variables. In statistics, there are three types of correlation coefficients. They are as follows:

**Pearson correlation:**

The Pearson correlation is the most commonly used measurement for a linear relationship between two variables. The stronger the correlation between these two datasets, the closer it'll be to +1 or -1.

**Spearman correlation:**

This type of correlation is used to determine the monotonic relationship or association between two datasets. Unlike the Pearson correlation coefficient, it's based on the ranked values for each dataset and uses skewed or ordinal variables rather than normally distributed ones.

**Kendall Correlation:**

This type of correlation measures the strength of dependence between two datasets.

**KARL PEARSON’S COEFFICIENT OF CORRELATION**

The value of the correlation is obtained by the below formula and the value always lie between $\pm 1$.

To calculate the Karl Pearson’s Coefficient of Correlation, the following formula is

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

$$x = (X - \overline{X}); y = (Y - \overline{Y})$$

$$\sigma_x = St\,\text{an}\,dard\ deviation\,of\ series\ X$$

$$\sigma_y = St\,\text{an}\,dard\ deviation\,of\ series\,Y$$

## t-test:

A t-test is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

**The three types of t-tests**

There are three forms of Student's t-test about which physicians, particularly physician-scientists, need to be aware: (1) one-sample t-test; (2) two-sample t-test; and (3) two-sample paired t-test.

**(1) one-sample t-test**

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

$$t = \frac{(\overline{X_1} - \mu)\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

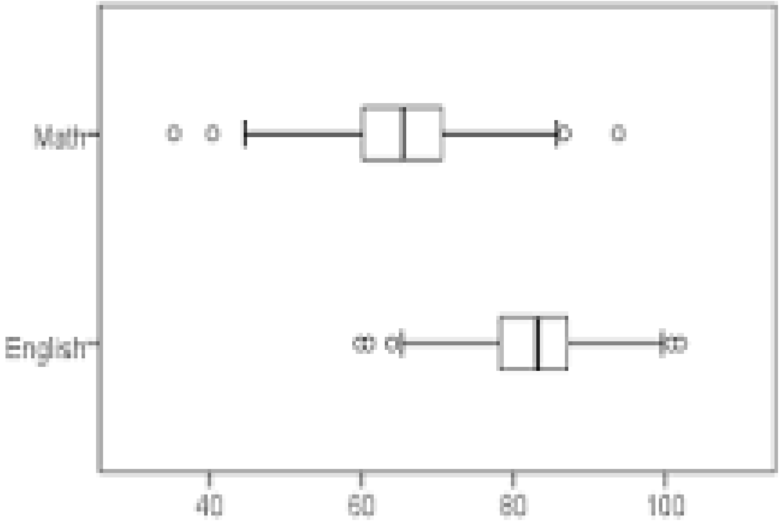
**(2) Two-sample t-test**

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\sum (X_1 - \overline{X}_1)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

**(3) two-sample paired t-test.**



The Paired Samples t Test compares the means of two measurements taken from the same individual, object, or related units. These "paired" measurements can represent things like: A measurement taken at two different times (e.g., pre-test and post-test score with an intervention administered between the two time points).

**Chi-square test:**

The $\chi^2$ test (chi-square) is one of the simplest and most widely used non-parametric test in statistical work. The symbol $\chi^2$ is the Greek letter Chi. The $\chi^2$ test was first used by Karl Pearson in the year 1900.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

For a Chi-square test, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

It is defined as $\chi^2 = \frac{\sum (O-E)^2}{E}$

Where O refers the observed frequencies and E refers to the expected frequencies. To calculate the expected frequencies

$$E = \frac{RT \times CT}{N}$$

E = Expected frequency

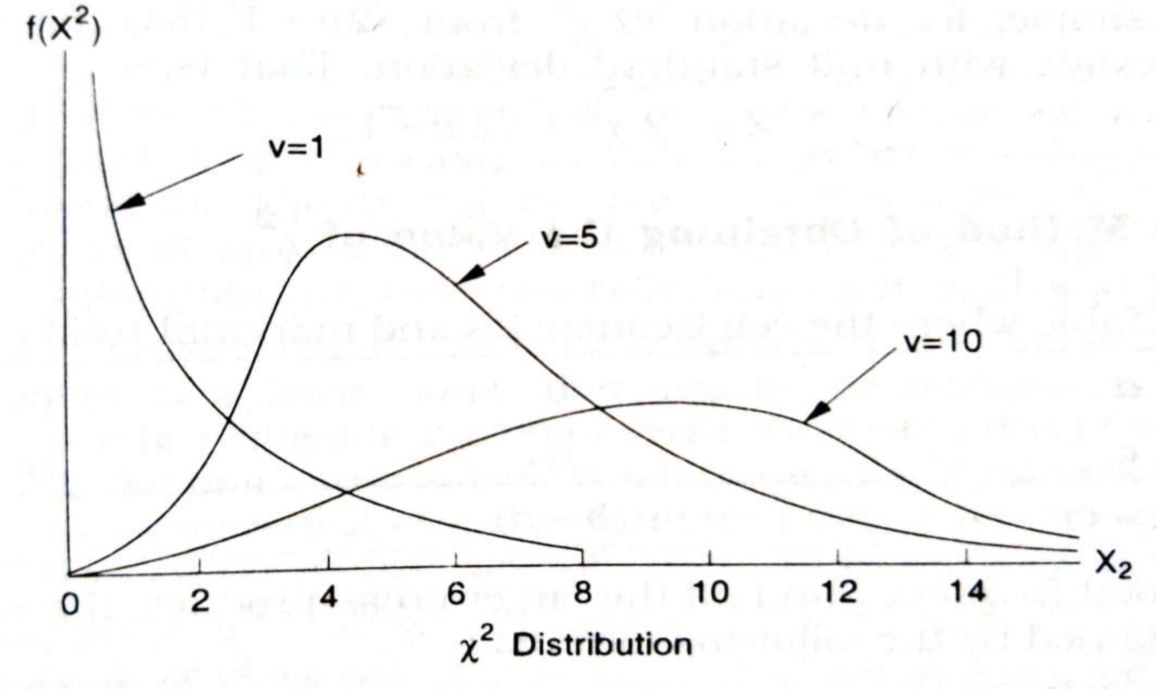
RT = The row for the row containing the cell

CT = the column total for the column containing the cell

N = the total number of observations.

**The Chi-Square distribution**

The following diagram gives the $\chi^2$ distribution for 1.54 and 10 degree of freedom.



**Uses of $\chi^2$ test**

1. $\chi^2$ test as a test of independence

2. $\chi^2$ test as a goodness of fit
3. $\chi^2$ test as a test of homogeneity

**Test and Goodness of Fit**

In the previous chapter various tests of significance such as t, F and Z were discussed. These tests were based on the assumption that the samples were drawn from normally distributed populations, or more accurately that the sample means were normally distributed. Since the testing procedure request assumption about the type of the population or parameters.

Though non-parametric theory developed as early as the middle of the nineteenth

Century, it was only after 1945 that non-parametric test came to be used widely. Originated in sociological and psychological research, non-parametric tests today are very popular in behavioural sciences. The following three reasons account for the increasing use of non-parametric tests in business research:

(1) These statistical tests are distribution-free (can be used with any shape of population distribution)

(2) They are usually computationally easier to handle and understand than parametric tests;

(3) They can be used with types of measurements that prohibit the use of parametric tests.

The increasing popularity of non-parametric tests should not lead the reader to form an impression that they are usually superior to the parametric tests. In fact, in a situation where parametric and non-parametric tests both apply, the former are more desirable than the latter.

**Association of Attributes**

Association of attributes is a statistical method that measures the relationship between two phenomena by studying the presence or absence of a particular attribute. It is used when the sizes of the phenomena are not directly measurable.

Here are some key concepts related to association of attributes:

Positive and negative association

Two attributes are positively associated when they are present or absent together. Negative association, also known as disassociation, occurs when attributes appear below expectation level.

Independence

Two attributes are independent when the proportion of one attribute among the other is the same as the proportion of the other attribute among the first.

Classification

When studying a single attribute, the population is divided into two classes based on the presence or absence of the attribute. This is called division by dichotomy.

Measures of association

The strength of the relationship between two attributes is measured using measures of association. Some measures of association include Yule's Coefficient, Coefficient of Colligation, Chi-Square coefficient, Karl Pearson's Coefficient, and Tschuprow's Coefficient.

Chi-square test

This test is commonly used to test associations between events, proportions, and goodness of fit to a theory.

***

**Compiled by**
Dr. G. Monikanda Prasad
Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli – 627 012