Limitations of Statistics

Although statistics is very widely used in all spheres of human activity, it has its own limitations. The following are some of the limitations.

1. **Statistics does not study qualitative phenomenon :** Statistics are numerical statements of facts. It can be applied only to such problems that can be measured quantitatively. Statistics cannot be used directly for the study of qualitative characteristics such as honesty, beauty, intelligence, culture etc. However, it may be possible to analyze such characteristics indirectly. For example we may study the intelligence of students on the basis of marks secured by them in an examination.
2. **Statistics does not study individual measurements :** A single or isolated figure cannot be regarded as statistics unless it is part of the aggregate of fact relating to a field of enquiry. Statistical methods do not give any recognition to an object, person or an event in isolation. The average income of a group of persons might have remained the same over two periods, yet some persons in the group might have become poorer than what they were before, statistical methods ignore such individual cases.
3. **Statistical laws are true only on average :** Statistics, as a science is not as accurate as many other sciences. Statistical laws are not universally true like laws of physics etc. They are true only on an average. Statistics deals with such phenomena that are affected by a multiplicity of causes and it is difficult to study the effects of each of these factors separately. Due to this limitation, the conclusions arrived at are not perfectly accurate.
4. **Statistics can be miscued :** Statistics are liable to be miscued. Any person can misuse statistics and draw wrong conclusions. Statistical methods are dangerous tools in the hands of the non-expects. Biggest limitation of statistics is that it deals with figures which are innocent and do not bear on their face the label of their quality. They can be easily distorted and manipulated by dishonest or unskilled users for selfish motives. Statistics neither proves nor disproves anything. It is merely a tool which, If rightly used, may prove extremely useful but, if miscued, may lead to fallacious conclusions. In the words of King "Statistics are like clay of which you can make a god or devil as you please." According to him "Science of statistics is the useful servant, but only of great value to those who understand its proper use. An example of wrong interpretation is given below.

   "In India the percentage of death among sick persons is higher in hospitals than at home" was stated by a person who collected information of deaths in hospitals and at home. This may lead to the conclusion that more persons die in hospitals than at home due to lack of proper treatment and care. But the investigator failed to take into consideration the fact that in India only seriously ill persons are hospitalized.
5. **Statistics do not reveal the entire story :** Statistical analysis may focus on only certain aspects of the study and may not bring to light the entire story. To illustrate, the opposition in the U.S.A to outsourcing of jobs to India is often justified citing statistics on job losses. However, if the demographic profile of the USA is studied, one finds that America is actually heading for an era of shortage of personnel. A good statistician must recognize the need to consider all facts and the complete story, rather than rely on a few tools of statistical analysis.
6. **No 'Cause' and 'Effect' relationship :** Statistics does not necessarily bring out the 'cause' and 'effect' relationship between various parameters. It only reveals the association amongst such parameters. The investigator needs to apply judgment to determine the relationship between the parameters. "To illustrate, if the 'sales' and

'advertisement expenditure' of an organization is studied, it does not necessarily tell us that there is a 'cause' and 'effect' relationship between the two. In fact, the technique of regression can be applied to find out what would be the 'advertisement expenditure' for a given figure of sales. This may not be practical, as sales does not directly cause impact on advertisement.

7. **Cannot be applied to any situation :** Statistics are collected for a defined purpose. The analysis resulting out of such statistics cannot be indiscriminately applied to any situation. Indiscriminate use of statistics to any situation can lead to false conclusions. This is particularly true when 'sample data' is used.
8. **Statistics are Contextual :** In order to understand a statistical study in its full detail, it is necessary that we also study the context of the study. A Change in definition of a variable, defective sample, inappropriate tools can all result in a totally different conclusion.

   Hence, one needs to be careful while drawing inferences on the basis of information collected by statistical investigator. If statistical conclusions are drawn form incomplete, inaccurate data, result will also be wrong and misleading.

Distrust of Statistics

By distrust of statistics, we mean, lack of faith or confidence in statistical methods and statements. Inexperienced, irresponsible and dishonest people have used statistical data and methods to fulfill their selfish motives. This has led to wrong interpretation of data. It is often commented by people that "An ounce of Truth will produce tons of statistics." Or "Statistics are lies of first order" or that "there are three types of lies-lies, damned lies and Statistics, wicked in the order of their naming". A Paris banker remarked "Statistics are like mini skirts. They cover up the essentials, but give you ideas". Some of the reasons for the distrust of statistics are as follows.

1. **Quality of Data :** "Figures are convincing. Therefore, people are easily led to believe them. It is said, "It is said, "If figures say so, it cannot be otherwise" or "Figures do not lie". Figures are innocent and easily believable. It is human psychology when facts supported by figures are given, people will believe them to the true. But figures do not have any label of quality on their face. For example, the executives of a marketing agency that was required to conduct a survey of users of product can fill in the questionnaires themselves and the agency can come up with erroneous conclusions as the base data going into the research is of poor quality.
2. **Incomplete Data :** To establish certain results or conclusions, some people make use of incomplete data or inaccurate figures. The truth is distorted and wrong conclusions are drawn. To illustrate, an education institution may claim that is generating 100% pass result since inception. However, the institution may be only 2 year old with hardly 10 to 15 students. Without knowing all the details, users can arrive at faulty conclusions.
3. **Manipulated Data :** Though accurate, the figures might be molded or manipulated by dishonest persons to conceal the truth and to present a false figure. When people realize that even statistical statements are not correct, their faith in the science of statistics is shaken and they start condemning statistics in the strongest possible language. In such cases the fault lies, not with statistics, but with persons who make use of it. If wrong figures are used, they are sure to give wrong conclusions. It is the duty of the persons who use statistics to see the figures used by them are free from bias and have been properly collected and scientifically analyzed.
4. **Statistics is only a tool :** It is said that statistics can prove anything. Some people charge that it does not prove a particular thing in a particular manner. But it must be

noted clearly that statistics does not prove anything. It is only a tool in the hands of a statistician. Different types of conclusions can be drawn form the same set of figures. A lay man has to be very cautious. If the figures are incomplete, inaccurate or collected by persons who have bias, then the conclusions arrived at are bound to be wrong. We can conclude in the words of walls and Roberts. "He who accepts statistics indiscriminately will often be duped unnecessarily, but he who distrusts statistics indiscriminately will often be ignorant unnecessarily."

Statistical Investigation

We have understood the importance and wide spread use of Statistics. However, we need to have quantitative data that has to be analyzed from the point of view of a defend objective. Statistical investigation or statistical enquiry is a process where relevant quantitative data is collected for the purpose of analysis to arrive at a conclusion. Data can be collected by way of a 'survey' or and 'experiment'. Data is collected from experiments in the case of physical sciences. In case of social sciences, data is largely collected through conduct of statistical surveys.

Stages of Statistical Investigation

A statistical investigation in order to be useful, must pass through the following stages.

1. **Observation :** A Statistical enquiry is often triggered with an 'Observation'. For example, a businessman dealing with hundreds of products 'observes' that there is a decline in sales of a particular product (e.g xx soap). This will prompt him to 'observe' the trend of that product with greater interest. He may observe a change in consumer behavior. For example, some consumers would look at packaging, while some may look for aroma, with some others asking for 'health' oriented brands. However, he may not be able to arrive at any conclusion. Thus, observation is a preliminary assessment of the problem. It requires an understanding of the nature and size of the problem. It will need collection of facts and their analysis. There is a great amount of 'subjectivity' at this stage.
2. **Framing of Hypothesis :** Hypothesis is the conclusion that is arrived at on the basis of observation, using deductive logic. It may or may not be true, and hence needs to be tested. However, the hypothesis provided direction to the statistical enquiry. Continuing the example, the businessman may find that more and more people are making the choice based on aroma. He probably arrives at this conclusion by observing that most soap buyers invariably smell the soap. This is only a hypothesis. In reality, consumers may be smelling the soap, but deciding on which soap to buy based on it price.
3. **Investigation :** Based on the observation and hypothesis, data needs to be collected to check whether the hypothesis is true. While the hypothesis is the basis for conducting the investigation, all related data is also collected. Continuing with our example, the investigation will not restrict to preference of consumers for aroma. It will also collect data on price, Packaging, usage preference, and such other factors. Statistical methods are used to analyze the data collected.
4. **Prediction :** Predictions are the outcome of the investigation. They refer to anticipation about the future, as derived from the results of the investigation. For example, in the case of our businessman, we may be able to deduce from the data collected that sales would pick up if they are sold as package (such as in Buy 3, get 1 free). This prediction is also not a certainty. It also requires confirmation and verification.
5. **Verification :** Verification is the process of putting the prediction to test. In order to complete the statistical enquiry, our businessman will actually need to introduce the

offer and await customer reaction. However, in case of large organizations or large markets, such experiments are conducted only in 'test' markets. Further action will depend on the result of the experiment.

Planning a Statistical Survey

Any person planning a survey must bear the following

(i) **Objective :** Any survey being taken up must have a clear objective. All the activities that are required to be carried out are directly or indirectly dependent on the objective of the survey. For example, the impact of offering a product as a package on total sales can be an objective of a particular survey. Defining the objective is particularly important if the conduct of the survey is being outsourced to a marketing research firm or any such third party.

(ii) **Scope :** The scope of the survey defines the aspects that need to be covered to accomplish the objective. The scope can be in terms of Geography (Hyderabad market), Product, competition analysis, etc.

(iii) **Definition of terms :** The terms that are very commonly used but interpreted differently by different persons need to be properly defined. For example, if we are studying increase in 'sales, is it increase in number of units sold, or increase in sales revenues. There should not be any scope fro confusion in the mind of the person conducting the survey.

(iv) **Dummy Report :** When the scope of the survey cannot be clearly defined before the start of the survey, it is preferable to prepare a 'dummy' report on the basis of current understanding. Such report will indicate the additional study that is required and provide clarity on scope.

(v) **Laying down of Hypothesis :** A Hypothesis can be framed based on preliminary observations. Such hypothesis can also be framed as part of statement of the 'objective' Laying down hypothesis will provide direction in planning and execution of the survey.

Collection of Data

Data are collection of any number of related observations. A collections of data is called a data set, and a single observation is called a data point. Data constitute the foundation of statistical analysis and interpretation. Hence, collection of data is a very critical step in statistical analysis. Data can be collected either through internal records, or from Primary and Secondary sources.

Internal Data

It refers to data available from records kept by an organization on a routine basis. Organizations need to maintain a lot of records on account of statutory requirement or for information of management. Salts, all major expenses quality of production, number of employees etc, all constitute Internal data. This data is relatively easy to collect but it may not be available in one department. Hence, it requires a bit of compilation. The extensive use of Computers has given rise to development of robust Management Information Systems (MIS) that generates all internal data without any difficulty. Thus, collecting data from internal data costs less and is relatively easy. However, it is often incomplete and needs to be supplemented with primary or secondary data.

Primary Data

It refers to the data collected for the first time and is original in character. Primary data is the results of surveys conducted by government and also by individuals, institutions and research bodies. Primary data is essentially raw and statistical methods will have to be applied on such data for the purpose of analysis and interpretation.

Secondary Data

Data, which is not originally collected but is obtained from published or unpublished sources, is secondary data. They are collected and processed by some agency and made use of by some other agency for their statistical work. However, secondary data would have already been treated statistically by the agency, which actually collected it..

For example, a company manufacturing UPS systems wants to plan for an expansion. For that, it needs to know what is the size of the market for UPS systems and what it should do to further improve its product. For this, it decides to meet its customers. It can generate the list of customers from its own records. This is an example of Internal data. To know about product improvement, it will have to collect information on levels of customer satisfaction. This is an example of primary data. To estimate demand potential, it approaches a research agency to know the size of the computer market. This is an example of secondary data.

Difference between Primary and Secondary Data

The difference between primary and secondary data is one of degrees only. Data, which is primary in the hands of one, becomes secondary in the hands of another.

Primary Data	Secondary Data
1. Primary Data is collected by the person Conducting the statistical enquiry	1. Secondary data is data collected by person agencies other than the one conducting the enquiry.
2. It is 'raw' and original	2. It is processed and hence, not Original
3. It is most relevant to the study being conducted	3. It may or may not be directly relevant to the study
4. It involves huge costs, time and Effort of the investigation.	4. It is relatively less expensive and takes less time and effort
5. There is no need for extra precautions in using the data	5. It should be used with great care.

Advantages of Primary Data

1. Primary data is very reliable as it is being collected first hand.
2. Primary data is most suitable for the study as it is being collected for the purpose of the study.
3. Since data is being directly collected, the scope for mis-interpretation or loss of data is minimal.
4. There is greater degree of participation from the persons from whom data is being collected.

Limitations of Primary Data

1. Primary data is vulnerable to manipulation. The person collecting the data can drop some of the response from the total responses collected to tilt the results in a particular direction.
2. It is a very expensive proposition.
3. It involves a lot of time and effort on part of the investigator.
4. It is raw and needs to be processed by applying various statistical techniques.

Advantages of Secondary Data

1. Secondary data is processed data. Various statistical techniques would have already been applied on raw data and hence, it is possible to arrive at conclusions based on secondary data.
2. The reliability of secondary data is high if the source of such data is a reputed agency or organization.

3. Secondary data saves huge efforts and time that is required for collecting primary data.
4. Sourcing of secondary data is relatively inexpensive. Most data is freely available or can be obtained by paying a small fee.

Limitations of Secondary Data

1. Secondary data may not be directly relevant to the study being conducted.
2. Secondary data is as reliable as the reputation of the agency providing it.
3. It is processed data and hence, would have already lost some of the forced of the original data collected.
4. There is possibility that the secondary data might have become outdated, as it is being collected in the present situation.
5. Selective adoption of secondary data could result in misleading conclusions

Choice between Primary and Secondary data: The choice between primary and secondary data depends on the following considerations:

1. Nature an scope of the enquiry
2. Availability of finance
3. Availability of time
4. Degree of accuracy desired and
5. The collecting agency-whether government, an institution or an individual

Methods of Collection of Data

There is a lot difference in the method of collection of primary and secondary data. In the case of primary data, the entire scheme of the plan starting with definitions of various terms used, units to be employed, type of enquiry to be conducted, extent of the accuracy aimed etc, is to be formulated, whereas the collection of secondary data is in the form of mere compilation of the existing data.

Methods of collecting Primary data : The following are the methods of collection of Primary Data

1. Direct personal interviews
2. Indirect oral investigations
3. Information received through local agencies
4. Mailed questionnaire method
5. Schedules sent through enumerators

Direct personal interviews

Under this method, the investigator collects information personally from the persons from whom it is to be obtained. In other words, the investigator has to go to the field personally, meet people, ask questions pertaining to the survey and collect the required information. In such cases, it is necessary that the investigator should have keen sense of observation, should be polite, courteous and tactful. He should acquaint himself with local conditions. The questions should be asked in simple and easy language so that he is able to obtain precise and correct answer.

Merits

(a) The information obtained is more reliable and accurate While collecting information, the investigator can check information given by respondents by cross-examining them
(b) People will be more willing to give information when they are approached
(c) Information on very sensitive matters can be collected by tactfully mixing such questions with other questions and twist the questions depending on the reaction of the investigator.
(d) The investigator can adopt the language of communication according to status and educational background of the informant.
(e) The investigator may be able to obtain supplementary information, which may help in the interpretation of the data collected.

Limitations

(a) This type of investigation is restrictive in nature. It is suited for intensive studies and not for extensive enquiries. This method is not suitable if the field of investigation is too wide in terms of number of persons to be interviewed or the area to be covered.

(b) This method is very costly, time consuming and requires much manpower. The informants can be approached only at their convenience. If the information is to be collected from working class, then information can be collected only in the evenings and at weekends. As a result of this, the investigation gets spread over a long period.

(c) It is Subjective in nature. Success of the investigation depends upon tact, intelligence, skill and courage of the investigator.

(d) The interviewers have to be thoroughly trained and supervised. Moreover, the personal bias and prejudice of the investigator may affect the findings of the enquiry.

Indirect oral investigation

This method is followed, when the information to be obtained is of a complex nature, or the persons concerned are unwilling to give the required information. For example, if the information to be obtained is on certain social evils such as drinking, gambling, smoking etc, the persons concerned will be reluctant to furnish the required information. Such information can be obtained by interviewing person who are directly or indirectly concerned with the subject matter of the enquiry. In such cases, enumerators are appointed for such purposes. A small list of questions pertaining to the investigation is prepared and these questions are put to the persons (known as witness) and these questions are put to the persons (known as witness) and their answers are recorded. Enquiry commissions or committees appointed by Government generally adopt this methods to get facts relating to the enquiry.

This is a very popular method, but correctness of information depends upon a number of factors. The success of the method depends upon the skill, intelligence and efficiency of the enumerators. They should be properly trained. Personal bias or Prejudice of the enumerators, should not be allowed to effect result of the enquiry. The accuracy of the information collected also depends upon the nature of the witnesses. It should be seen that witness are unbiased. The findings of the enquiry should not be based on information supplied by single person. A number of persons should be interviewed to know the facts.

Merits of Indirect Oral Investigation

1. Since informants /respondents are personally contacted all the advantages of Direct personal interview, namely reliable and accurate information, willingness of people to share information, use of tact and intelligence to extract correct information that is sensitive and choice of appropriate language for communication, are applicable to this method.
2. This method is relatively less expensive and takes lesser time
3. A wide area can be covered in this method.
4. The prejudices of direct respondents can be avoided
5. Expert views of specialists can be obtained to formulate and conduct the enquiry move effectively.

Demerits of Indirect Oral investigation

1. Direct supervision of investigator and personal touch of the respondents is lacking.
2. Accuracy of data collected and the inferences drawn are relatively less reliable.
3. Prejudices of witnesses motivated by selfish considerations may impact the study and create a bias in the results.

Information received through local agencies

Under this method, the investigator appoints local agents (commonly known as correspondents) in different places to collect information. These correspondents collect

information and transmit it to the central office where the data is processed. Newspaper agencies generally adopt this method. Correspondents supply information regarding important developments in their areas in the filed of sports, accidents, riots, strikes, political developments etc., to the office. This method is also adopted by the government in case of crop estimates. This method is cheap and appropriate for extensive investigation. However, the reports may now always be correct due to personal bias or prejudice of the correspondents.

Merits of Information through local Agencies

1. This method is relatively less expensive and less time consuming
2. Large areas are to be covered
3. This method is suitable when data needs to be collected on a continual basis.

Limitations

1. The reliability of data is low
2. Collection of data is not consistent, as each person is likely to have a different style.
3. There is greater chance of under-estimation

Mailed Questionnaire Method

Under this method a questionnaire is prepared and sent to the informants by post. A questionnaire consists of number of questions relating to the field of enquiry. Space is provided for answers to be filled by the informants (called respondents). This is sent with a covering letter explaining in details the aims and objective of collecting information. The respondents are requested to extend their full co-operation by furnishing correct replies within a reasonable time. Respondents are taken into confidence by assuring them that the information supplied by them will be kept confidential. To ensure quick and better response, the return postage expenses are usually borne by the investigator. This method is generally used by the research scholars, non-government officials, individuals and even by government. This method is less expensive and information can be collected form a wide area. Success of the method depends upon cooperation of informants. Generally, it is found that informants adopt and attitude of indifference towards such enquiries. There is always some uncertainty about their response. Cooperation on part of informants may be difficult to presume. The information supplied by the informants may not be correct and it is very difficult to verity its accuracy. Moreover, this method can be adopted only when informants are literate people.

Merits of Questionnaire

1. This method can be employed to cover vast areas simultaneously
2. It is relatively inexpensive and fast
3. The bias of the persons collecting data is reduced to the minimum
4. This method suits studies when same data is required to be collected frequently.
5. It ensures consistency across respondents, geographic and multiple studies.
6. Data collected is relatively more accurate
7. Much thinking is done in the drafting of questionnaire. Thus, there is complete clarity on what the investigator needs. This results in quality of data being better and more relevant
8. The respondents can be assured of confidentiality by making the personal details section (such as Name, Address Telephone number etc) optional.

Limitations

1. This method cannot be used for collecting data from people who cannot read and write.
2. There is no flexibility. There is no scope for the investigator to use tact, twist questions or cross examine the responses.
3. The hit ratio is very poor as questionnaires are impersonal. Usually, respondents are indifferent to mailed questionnaires and do not respond at all. This results in the whole effort going waste

4. There is a possibility that a respondent may not understand the question properly. Hence, the response may be faulty.
5. There is no way of verifying the accuracy/authenticity of data provided. People may willfully provide wrong information.
6. Respondents resist filling questionnaires in their own hand writing.

Drafting a Questionnaire

Questionnaire is the major medium of communication between Investigator and respondents. It should be drafted with utmost care and caution so that all relevant and essential information for enquiry is collected without any ambiguity or difficulty. Drafting a good questionnaire is a skilled job requiring care, efficiency and experience. No hard and fast rules can be laid down for framing a questionnaire. However, the following points may be kept in mind while framing a questionnaire.

1. **Size :** The number of questions must be restricted to the minimum possible. More the number of questions, lesser the likelihood of it being answered by the respondents.
2. **Clarity :** Questions should be clear and unambiguous. Simple language must be used. Words that can have multiple meanings or interpretations should be explained.
3. **Sequence :** The questions should be arranged in a logical sequence. This will help the respondent to understand the context and answer appropriately.
4. **Easy to Respond :** It should be easy for the respondent to answer the questions. It is preferable to provided multiple choices as answers, of which the respondent will generally choose the most appropriate answer (unless multiple ticks are allowed). Open ended questions should not require lengthy or elaborate answers.
5. **Confidentiality :** The questionnaire should make all questions that reveal the identity of the respondent (such as Name, Telephone Number, address, etc) optional. The respondents should also be assured that information provided by them would be kept confidential.
6. **Questions to be avoided :** Questions that involve prestige or status, leading questions (Questions that lead the respondent to an answer e.g Is Hero Honda your favorite motor cycle brand), questions of sensitive or personal nature, questions whose answers require detailed calculations, subjective & irrelevant questions need to be avoided.
7. **Look and Feel :** Good quality paper and printer must be used. The layout and presentation should be made interesting.
8. **Post Survey Usage :** Mode of Tabulation and Analysis of data collected should also be kept in mind at the time of design of questionnaire.
9. **Cheeks and Accuracy :** The questionnaire must contain internal cheeks to ensure accuracy of data being collected. For example 'age of marriage' and age of the first child' are two questions that can help the investigator to know if data provided is accurate or otherwise.

Steps Apart from the above factors that need to be considered in designing the questionnaire, the following additional steps should be taken to ensure best results from the method.

1. **Covering Letter :** A covering letter informing the respondents about the objective of the questionnaire and a request for their co-operation should be sent along with the questionnaire. The letter may also assure the respondents about maintaining confidentiality.
2. **pre-paid Reply Card :** A self-addressed, appropriately stamped envelope should also be sent along with the questionnaire to reduce the inconvenience of the respondent.

3. **Incentives :** Respondents may be incentivized to respond to the questionnaire by announcing gifts and prizes for people responding to the questionnaire is full
4. **Training :** In the event of questionnaires being sent with enumerators (Schedules) the name of the enumerator should be stated in the covering letter. All enumerators should be properly trained on what each question means and how to get the required information from the respondents.
5. **Pre-test :** It is preferable to pre-test the questionnaire on a small sample before a complete launch so that suitable changes can be made before it is sent to all respondents.
6. **Supervision and Scrutiny :** The work of the investigators must be properly supervised. Test checks may be considered to ensure that the investigators are not manipulating data. The responses should be properly scrutinized for detecting material inconsistencies.

Schedules sent through enumerators

Under this method, the enumerators go to the informants along with the questionnaire, get replies to the questions contained in the schedules and fill them in their own hand-writing. The essential difference between questionnaire method and schedule method is that questionnaire is sent to the informants by post and filled up by them (informant), while schedule is carried by the enumerators personally to the informants and is filled up by enumerators. The enumerators explain aims and objectives of the investigation and remove the difficulties which informants may have in understanding the questions. This method is very useful in extensive enquiries and fairly dependable results can be expected. However, this is very expensive. It is used by Government, big business houses and reputed research institutions. Population census all over the world is conducted by this method

Difference between Questionnaire and Schedule

The Schedule is also a questionnaire. However, we can note the following differences.

Questionnaire	Schedule
1. Questionnaire is mailed to respondents	1. Schedules are taken to respondent by enumerators.
2. Questionnaires can be used only with educated respondents.	2. Schedules can be used with uneducated respondents
3. Responses are filled in by respondents	3. Responses are filled in by enumerators
4. Questionnaires do not have adequate flexibility	4. Enumerators can use tack and intelligence to elicit the required information information.
5. Suitable if large areas are to simultaneously covered	5. Best suited for survey of small, focused areas
6. The percentage of people responding is low	6. The percentage of people responding is high
7. There is minimal influence on the respondent	7. Enumerators can influence the respondent
8. Confidentiality can be ensured.	8. Confidentiality is not promised
9. It is a cost effective method	9. It is relatively expensive
10 The most important factor is the design of the questionnaire	10. The most important factor is the personality of the enumerator.

Merits of Schedules

1. Since schedule is also a questionnaire, it has almost all the advantages of the mailed questionnaire method such as reduced bias, frequent collection of data, accuracy of data, consistency, greater clarity and relevance.

2. Enumerators can explain the objective of the study and clear the confusion in the minds of the respondents.
3. The personal touch of the enumerator ensures that the response is much greater.
4. It allows the enumerator to use tact and judgment in cross examining the responses
5. This method can be used even if the respondents are illiterate.

Limitations

1. It is an expensive and time consuming method.
2. Respondents may not answer all questions appropriately as confidentiality is not ensured.
3. Using the method across wide areas will result in inconsistencies, as more than one enumerator is involved
4. The success of the method depends on the capabilities of the enumerators.

Sources of Secondary Data

Secondary data are those that have been already collected and analysis by some agency. This may be published or unpublished.

Published Sources

- Official publications of International bodies like UNO or its subsidiaries, foreign Governments, etc.
- Official Publications of Central and State Governments
- Semi Official publications of various local bodies such as Municipal Corporations and District Boards.
- Private Publications such as Publications of Trade and Professional bodies like Institute of Chartered Accountants of India, Financial and Economic Journals, Annual reports of Joint Stock companies, Publications brought out by Research Institutes etc.

Unpublished Data

Some statistical information, though collected, may not be published. This includes the wok of research bodies, trade associations etc. Such data can be made use of if it is Reliable, Suitable and Adequate for the purpose of investigation.

Precautions to be taken while using secondary data

Secondary data should be used after subjecting it to a thorough and careful scrutiny. One should make sure that the data is reliable, accurate, adequate and suitable for the purpose for which it is sought to use. A qualified person or agency should have collected data. It should be as recent as possible. Such a scrutiny will help the user in avoiding the limitations and misuse of Statistics.

Statistical Units

A Statistical unit is a well-defined and identifiable object or group of objects with which the measurements or counts in any statistical enquiry can be undertaken. It is the unit of measurement in Statistics. For example, in an exercise aimed at finding out the extent to which a disease is spreading, number of persons affected by the disease will need to be counted. Thus, in this exercise, 'person' is the statistical unit. In most studies, the units are physical units of measurements. Examples can be kilometers (of, road traveled) hours (spent by children on sports), tons (of commodities transported, etc), However, in many studies, particularly in case of studies of socio-economic nature, we do not have precise definition of the units of measurement. For example, if a study of number of families having access to a telephone is to be conducted, the statistical unit in the case is 'family'. However, family is not a physical unit. It is an 'arbitrary' unit and needs to be defined properly so that there is no inconsistency at the time of collecting data.

Requisites of a Statistical Unit :

In order to qualify as statistical unit, it should be ensured that

1. **It is well-defined :** There should not be any scope for interpreting it in multiple ways. In case of arbitrary units (such as 'family') it should be tightly defined. In the above example, the investigator may define as a set or people living together in a household. Thus 2. brothers staying with their respective wives and children in one house will all be part of one family.
2. **It should be stable :** It should be stable over long periods of time and across places. This will ensure that data collected at different times and at different places is comparable. For example, in studying the consumption patterns of a commodity by the people, it is preferable to study the quantity consumed, rather than value, as value is relatively less stable on account of inflation.
3. **It should be appropriate to the enquiry :** The unit selected should be appropriate to the enquiry. For example, if retail spending patterns are being studies, we need to compare retail prices rather than wholesale prices.
4. **It should be uniform :** The unit adopted should be uniform throughout the investigation. For example, if land owned by people is being studied, we are likely to get the numbers in square feet, square yards acres, grunts, beeches etc which will lead to confusion. Hence, either all the observations must be in one unit or a reliable conversion scale must be adopted to convert all data into comparable units.
5. **It should be distinct :** The unit should be such that all the elements of the study should belong to one and only one statistical unit.

Types of Statistical Units

The statistical units may be classified into Units of Collection and Units of Analysis

1. **Units of Collection :** Units of collection are further sub divided into

(i) **Units of Enumeration** : It is the basic unit on which observations are to be made. It can be person, family, industry, etc. It must be clearly defined and the definition must be clearly explained to the enumerators so that there is no inconsistency in data collection.

(ii) **Units of Recording ;** These are the basic units in terms of which the data is collected or recorded. They are the units of quantification. Height, weight, meters, price etc are examples of units of recording.

Units of Recording are of the following types

(a) **Simple Units :** Units that represent only on e condition without any qualification are simple units. They have a single determining characteristic. Examples of such units are meters, hours, number (of person) etc.

(b) **Composite or Compound Unit :** A simple unit with some qualifying words is a 'composite' unit. A simple unit with only one qualifying word is called a 'compound' unit. Examples of compound units are 'skilled' labor, 'monthly' wages, etc. If two or more qualifying units are added to a simple unit, it becomes a 'complex' unit. Measures of Productivity (output per man hour, kilometers per liter of diesel, etc) are examples of 'complex' unit. Both 'compound' and 'complex' unit are together called 'composite' units.

(c) **Hypothetical Units :** The are non-existent units but are used for comparison of 'subjective' issues. Jessant Singh, former Finance Minister of India, coined the term 'Gross Contentment Index', which cannot be objectively measured.

Units of recording are also accompanied by units of estimation: Units of estimation are essentially units of recording that are rounded off. For example, sales revenue of the top 500 companies can be studied in 'Rupees in Cores'. This is a unit of estimation where 'Rupees' is the unit of recording.

Units of Analysis and Interpretation

Units of analysis are those units in the form of which statistical data are ultimately analyzed and interpreted. Rates, Ratios, Percentages, etc are examples of Units of analysis provide relative figures that are independent of units of recording.

Methods of Collecting of Data

Statistical data may be collected in any one of the following ways:

1. **Census Method :** In this system, information is collected from the entire 'population'. The term 'population' means the entire universe of units that fall under the scope of study. For example, if we are studying the heights of students of a school, all the students of the school together constitute the population. In census method, we will need to measure the heights of each and very student of the school. The census of population of human beings in India, collected every 10 years, is an example of data collection by 'census' method.

Advantages

1. The data collected is accurate and most authentic. There is no statistical error.
2. It reduces biases, estimations probabilities and resultant uncertainties to the minimum
3. Data is likely to be more consistent.

Disadvantages

1. It is prohibitively expensive and time consuming.
2. Data may become out-dated by the time it is collected in entirety and analyzed.
3. It requires qualified enumerators in large numbers, which is a challenge
4. The authenticate of data is a function of the sincerity of enumerators.

Suitability

The census method is suitable in the following circumstances.

1. It is very critical to obtain accurate date. For example, number of persons infected by SARS virus.
2 The universe or population is small
3. Individual response is vital. For example, feedback, on service provided to guest by a 5 star hotel.
4. Adequate resources in terms of people and finance are available.

Sampling Method

A 'sample' is a part of the 'Population'. It is subset of the entire set of units that fall within the scope of study. For example, to study the brand preference of consumers in aerated drinks (Coca Cola vs. Pepsi) we don't have to ask each and every person who consumes aerated drinks about their choice. Only a small number of persons are covered, who form the 'sample'. The number of units (or' elements') in the sample is called sample size.

The method of obtaining data or conducting a statistical investigation by studying a sample, rather than the entire population, is 'sampling method'. The most important aspect of this method is that the sample should be 'representative' of the population.

Advantages

1. It is very cost-effective in comparison to census method
2. It has lower turnaround time (i.e. takes less time) and will provided the latest updates, rather than outdated information.
3. Accuracy can be ensured by appropriate selection of sample and by conducting various statistical tests.
4. If offers great flexibility in organizing the statistical investigation if it requires wide coverage.

Disadvantages

1. Choice of sample is of paramount importance. Wrong sample chosen could lead to erroneous results.
2. There is lot of scope for statistical error to creep in
3. There is lot of scope for individual bias and manipulation of data. For example, the study of TRP ratings resulted in a controversy (between Star plus, Sony and Zee Television channels)

Suitability

1. It is suitable and hence almost universally applied in market research and other business studies.
2. It is the only method that can be used when conducting the experiment leads to consumption of the test unit (example, checking if grapes are sweet, before buying them).
3. It is used when the universe (population) is infinite or very large.
4. It is also used where statistical error can be tolerated and factored in
5. It is used where the data to be studied is largely homogenous.

Methods of Sampling

There are various methods of sampling, which can be broadly divided into Random Sampling and Non-Random Sampling methods. Irrespective of the method being followed. It is of paramount importance that the sample is representative of the population.

Random Sampling Methods

A random sample is one which is selected in such a way that every item in the population has equal chance of getting included in the sample, 'Random' does not mean 'haphazard'. For example, if a study is being conducted to find out whether the general public has liked a particular movie, a person stands outside the movie theatre and keeps asking the opinion of the people coming out of the theatre at random. The various random sampling methods are :

Simple Random Sampling

Under this method, the entire 'population' is taken as one single composite unit for the purpose of selecting a sample. No attempt is made to study the 'characteristic' of the sample and make sure that the sample is 'representative' of all characters of population. Such sample can be drawn by

(a) **Lottery :** All the 'elements' or' constituents' of a population are numbered on identical slips of paper and put in a drum. After rotating the drum for quite some time required number (which is the decided sample size) of slips are picked up by blindfold selection. All the constituents represented by the selected slips form the random sample. This is the simplest and hence most popular method.

(b) **Random Number Table :** We have readymade 'random number tables, that are so constructed that each of the number 0, 1,2,3,4,5,6,7,8 and 9 appear with approximately the same frequency and independently of each other. In order to choose a sample, all the elements of the population are allocated numbers (0 to 9 or 00 to 99 or 000 to 999, etc depending on sample size). We then select, at random, any page of the random number table and then any row, column or diagonal of the table and pick up the numbers given in that row, column or diagonal. The elements represented by the numbers form the sample.

This method is scientific, economical and representative. However, complete list of all elements of population may not be available. Also, numbering them may become very cumbersome. However, this method is well suited where the size of the population is not very large.

(ii) **Stratified Random Sampling :** If the 'population' consists of diverse segments that can be clearly distinguished, such population is first classified into groups each representing a divers segment, and then simple random techniques are used within each group, to come up with a total sample that is more representative of the diverse population. These diverse segments are called a 'strata' (Stratum', if singular) and this method is called 'stratified random sampling. The classification of population into various 'strata' can be done on the basis of its Geographic, sociological or Economic characteristics. The size of sample from each, strata is normally proportional to the total size of that strata to the total population size. However, it can also be of disproportional size.

This method enhances the 'representative' character of the sample. Hence, it is likely to result in more accurate results, particularly where the distribution of population is skewed. However, in many cases, it is very difficult to classify the population into different strata.

(iii) **Systematic Sampling :** In case of systematic sampling, the first sample unit is selected at random. Thereafter, depending on sample size, the rest of the units are selected automatically by using a systematic process. For example, in a colony of 1000 houses (numbered from 0001 to 1000). We need to study 50 houses to know the impact of Television of children, we will select the first house at random (say, H.No.679). Thereafter we keep selecting every 20th house (population size /sample size = 1000 /50 =20). Thus, the sample will be H.No.s 679, 699, 719, ...999, 019, 039,.....) '20' is called the 'Sample Interval.'

This method retains the advantages of simple random sampling. It reduces the time and effort in constituting the sample. However, there is scope for 'bias' or 'preference' of the investigator to creep in and he can design the 'system' in such a manner that a particular group may have a relatively greater probability of getting selected.

(iv) **Multi-stage Random Sampling :** Under this method, the final sample is obtained after a drill down exercise, with random selection at each stage. To illustrate, if we wish to study the reaction of consumers to the nation-wide launch of a new Flat TV model by LG, we can drill down to the sample as follows. Select 4 states at random. From each state, select 3 cities /towns and within each of the 12 cities select people who have used or are aware of the new Television model. The choice of states, then cities and also the final 'persons' constituting the sample is done 'at random'. This method is called Multistage random sampling.

This method is very useful, when wide areas are to be comprehensively covered in a very short time. It saves time, effort and costs. However, the quality of representation may not be very high, particularly if there is heterogeneity within the population.

Non-Random Sampling Methods

Non-Random sampling

(ii) **Cluster Sampling :** In this method, the entire population is divided into some recognizable sub-groups called 'clusters' A random sample of the clusters is drawn and then all the units belonging to the clusters are considered as part of the sample. For example, if a study is to be conducted to estimate the reach of 'Eenadu' newspaper in the city of Hyderabad, the city of Hyderabad is subdivided into, day 100 Blocks, and by random sampling, 10 blocks are chosen. All the households falling into the chosen 10 blocks are covered under the sample. Each of these block is a 'cluster'. This method is usually employed when some traits of the population are being studied.

(iii) **Convenience Sampling :** Under this method, a sample is obtained by selecting such units that are convenient to locate or contract. An example of convenience sampling could be a list obtained from telephone directory. This method neither has the advantage of statistical tools/ techniques nor the intuitive or judgmental capabilities or an experienced investigation. This method is not advisable as it is not likely to yield accurate results. It can be used only to reinforce a known fact.

(iv) **Sequential Sampling :** Sequential sampling is not a separate method of drawing a sample. It is more of a process. A sample is drawn and the statistical investigation is conducted. If the result are satisfactory, nothing else is done. However, if study of the sample is inconclusive, another sample is drawn. This process may go on for any number of iterations until the investigator can arrive at a conclusion. This process is applied in Statistical Quality Control (SQC).

(v) **Quota Sampling** : It is a combination of stratified Sampling and Judgment sampling. In this method, the population is divided into 'quotas'. Each quota consists of units with specified characteristics. Each investigator or enumerator is told about the number of units he/she needs to select from the quota. This is normally done using judgment. This should not be confused with cluster sampling. In cluster sampling, only few out of all clusters are picked and all units belonging to selected cluster are selected. In Quota sampling all quotas are selected but sampling is applied within each quota.

There is no one best method of picking up a sample. The choice of method depends on the purpose of the investigation. It is possible that an investigator may decide to use multiple methods to meet the requirements of the study being undertaken.

Sampling Errors or Statistical Errors

In statistics, we tale the help of samples and arrive at some conclusions with respect to population. The actual value of population may or may not be identical to the estimate arrived at based a sampling, Statistical 'error' is the difference between the 'true' value and 'estimated value' of any subject of study.

Reasons for Statistical Errors

Statistical errors may crop up on account of any one or more of the following reasons.

1. **Errors of origin :** These could be (a) bias of enumerators / investigators (ii) faulty definition of statistical units (iii) Imperfection of measuring instruments and (iv) Inherent instability or erratic tendency of collected data.
2. **Errors of Inadequacy :** These errors are caused due to incomplete information or inadequacy of size.
3. **Errors of Manipulation :** These errors are committed by the investigation team at the time of counting, measurement or approximation.

Type of Errors : Errors may be classified into Biased and Unbiased errors.

Biased Errors :

These errors are caused on account of a bias or prejudice of enumerators or of measuring instruments. Such errors are dangerous and cannot be tolerated, as they are likely to impact the purpose of the study.

Unbiased Errors

There errors arise on account of 'chance'. They are an offshoot of various statistical techniques. They are not very serious and do not affect the results significantly.

Classification of Data

The largest amount of data and the greatest amount of detail may not convey the most useful information for decision making. An important aspect of Statistics is to organize and present data so as to convey critical information quickly.

Classification means grouping of a whole into different groups or classes. However, each of these groups should have a common characteristic. For example, all living things can be classified into plants and animals. Animals can be classified into Amphibians, Reptiles, and Mammals etc.

According to Horace Secrets, "Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different but related parts".

Objectives of Classification

1. To condense the mass of data into a form that is easily understandable.
2. To enable the user to get a proper grasp of the significance of the information contained in the data.
3. To sort data in such a manner that irrelevant details are ignored and relevant details only are considered.
4. To present the data in tabular form and enable the user to apply statistical tools and techniques on the same
5. To enable interpretation, analysis and generalization of data
6. To facilitate comparative study of variables.
7. To establish relationships between variable on study of various characteristics of data
8. To bring out similarities and disparities in data collected.
9. To provide an executive summary of the data collected at a glance.

Rules of Classification

There are no hard and fast rules of classification but the following principles must be observed.

1. Classification must be **clear and unambiguous**. It should not lead to multiple interpretations and increase confusion in the minds of the user.
2. Classification must be clear **exhaustive**. Each and every item in the data must belong to a particular defined class.
3. Classification should be **mutually exclusive**. Every item in the data should belong to one and only one class. For example, if data collected is being classified into educated, uneducated and rich, a particular respondent could be educated and rich, hence it is not a proper classification.
4. The Classification should be **suitable** and relevant to the purpose of study. To illustrate, data collected to know the preference for some particular brands of soft drinks can be or no use if it is classified into people belonging to various professions.
5. The basis of classification should be **stable.** For example, if sales are classified by geographical areas, then selling expenses should also be classified by geographical areas.
6. The data classified into each class should be **homogenous.** If a particular class contains dis-similar components, such class can be sub-divided further so that each class has a common characteristic.

7. The classification of data should be **consistent**. If data is being classified on the basis of geography into North, East, West and South, we should ensure that all the areas, particularly in central region, are consistently being taken as part of the same class.
8. Classification should be **flexible**. It should allow new ideas to be incorporated in future. It should be such that data can be re-classified at a later point in time as per changed requirements.
9. Classification should allow for **comparison** between classes.
10. Classification should **aid statistical enquiry.** It should be presentable in various forms and allow the application of various statistical tools on the data.

Characteristics of Classification

In classification, data collected is distributed into different groups. This is done on the basis of certain common characteristics such as age, sex, income religion etc. Such groups should ideally be objective (income less than ₹ 50,000 p.a, ₹ 50,000 p.a more than ₹ 1,50,000 p.a) rather than subjective (poor, middle class, rich). The classification should bring out the similarity within the class and the diversity amongst different classes distinctly.

Basis of Classification

Data is classified depending on the objective of the study. Normally, data is classified on the following criteria.

1. **Geographical :** The criteria will highlight the location difference in case of the subject of study. For example, companies fighting for leadership in markets (e.g coca cola and Pepsi) are interested in knowing the sales of their products by Geography. While Coca Cola may be the leader in Andhra Pradesh, Pepsi may be the top brand in Delhi.
2. **Chronological :** Data is classified on the basis of differences in time. For example, the number of mobile telephone users from 1996 to 2001 will give valuable information on hour this service has expanded.
3. **Qualitative :** Data is classified as per certain qualitative factors, which cannot be quantified. Data is classified into two classes based on presence or absence of such attributes. Examples of such attributed could be Gender (Male / Female), Marital Status (Married / Single), Education (Literate, Illiterate), occupation (Business/ Service) etc.
4. **Quantitative :** Data is classified on the basis of some features, which are capable of measurement. The quantitative phenomenon under study is known as 'variables' and hence this classification is also known as classification by variables.

   For example, number of people who prefer to have their own business rather than take up or continue employment can be classified by age as under.

Age	% Of people preferring Employment	% Of people preferring Business
18-25	82	18
25-35	33	67
35-45	60	40
45-55	72	28
Above 55	88	12

Variable or Variety

A characteristic, which can be expressed numerically, is called a 'variety' or 'variable'. It refers to the characteristics that vary in amount or magnitude in a frequency distribution. The quantitative phenomenon like marks in a test, heights and weights of students, wages of workers of a factory, runs scored by each player or team in the 2003 Cricket World Cup etc are all examples of a vitiate.

Discrete Variable

'Discrete' variable that cannot take all values. In the above list of examples, the runs scored by a player are always a whole number. If cannot be fraction. Similarly, number of children in a family cannot be anything other than a whole number. Thus, variables that cannot take all possible values within a specified range are termed as 'discrete' variables. They are characterized by jumps and gaps between one value and the next.

Continuous Variable

Variables that can take all possible values in a given range are termed as continuous variables. These variables include fractional as well as integral values. Height and weight of children income of various households, percentage marks or percentile obtained by students, etc are all examples of continuous variable.

Frequency : The number of times each variety occurs in known as frequency

Frequency Distribution

A classification showing different values of a variety and the corresponding is known as 'Frequency Distribution'

Types of series

Statistical information can be classified on the basis of values that variety can take. It can be classified under three heads namely individual observations, Discrete Series and Continuous series.

Individual Observations

Individual Observations represent a series where items are listed simply after observation. They are not arranged into groups. Runs scored by top eight players of a team are an example of individual Observations. It will be presented as under:

Player	A	B	C	D	E	F	G	H
Score	15	28	81	43	67	114	16	21

Marks awarded to students of a class, wages earned by employees in factory etc are other examples of Individual Observations. Individual Observations being simple and independent, frequency distribution table cannot be prepared.

Discrete Series

Discrete Series represent items arranged in such a manner that each unit of data is separate and complete. Each unit of data is capable of exact measurement. A feature of discrete series is that there are intermittent gaps between two values. There is no continuity. Discrete service variable cannot be expressed in fractions. It has to be necessarily a whole number. Examples of Discrete series would include data on number of families having a defined number of children, number of students who scored a defined set of marks etc. It is presented as under :

No. ofChildren	0	1	2	3	4	5
No. of Family	8	17	35	24	12	3

In case of discrete series the order in which the series is written does not make any difference. Data can be in ascending or descending order.

Preparation of Frequency Distribution Table for Discrete Service

In a discrete series, there is a possibility of the same variety occurring a number of times. Hence, a frequency distribution table can be prepared. By preparing a Frequency distribution table, we are trying to count the number of times a particular variety has appeared. It is done through the following steps.

Let us understand the process with the help of an illustration.

Illustration 1. (Preparation of Discrete Frequency Table)

The following data shows the previous teaching experience of lecturers in a college. Prepare a Frequency distribution table in Discrete form.

0,1,2,3,10,4,6,3,5,5,5,3,7,3,2,8,0,9,2,10,8,4,2

Solution

Step 1 : List all possible values of the variable (year of experience) in order. The possible values are 0,1,2,3,10,4,6,5,7,8,9 and 10. It is arranged from the lowest to highest in column 1.

Step 2 : Look at each value provided in the data and put a tally across that value in the column 2. For example, the first value is '0'. Therefore, in the second column, against '0', we draw a bar. The next value is 2. We draw a bar against 2. We proceed this way till we finish putting a bar for all items. The last value is 2, which is appearing for the 5th time. Hence, instead of another vertical bar, we draw a diagonal bar across the 4 vertical bars and each such block implies 5 counts. Lastly, we count the number of bars and put the final tally in the last column.

Years of experience	Tally Bars	Frequency
0	III	3
1	I	1
2	IIII	5
3	IIII	4
4	II	2
5	III	3
6	I	1
7	I	1
8	II	2
9	I	1
10	II	2

In case of discrete series, the order in which the service is written does not make any difference. Data can be written in ascending or descending order.

Continuous Series

When data can take any value without any gap, the series becomes a Continuous Series. For example, the maximum day temperature recorded at various places in the country can take any value such as 39.8 degree or 41.3 degrees etc. Such data is studied by placing items into certain well-defined limits. For example, the prices at which shares of various companies get traded can be presented as under:

Price (₹)	0-00	100-200	200-300	300-400	400-500	500-600
No.	75	28	33	18	15	6

Before we attempt to prepare a frequency distribution table for continuous series, we need to understand certain terms.

Class : Each stated interval, into which the continuous variables get classified, is a class.

Class Limits

The class limits are the lowest and the highest values that can be included in the class. For example, in the class 10-19, 10 is the lowest value and 19 is the highest value. These limits are called lower and upper limits.

Class intervals

The difference between the upper and lower limit of a class is known as class interval. For example, in the class 10-20 the class interval is ten.

Mid-Point or Mid-Value

It is the value lying half-way between lower and upper class limits of a class. It is ascertained as follows :

$$\frac{\text{Upper limit of the class + Lower limit of the class}}{2}$$

There are two ways in which data are classified on the basis of class interval namely (i) 'Exclusive' Method and (ii) 'Inclusive' Method.

Exclusive' Method

In Exclusive method, the upper limit of the class is excluded from that class and is included in the next class. In such series, the upper limit of one class is the lower limit of the other class. For example.

Marks	No. of Students
0-10	5
10-20	12
20-30	24

'Inclusive' Method

Under the inclusive method of classification, the upper limit of the class in included in that class itself. For example.

Marks	No. of Students
0-9	10
10-19	14
20-29	23

The student securing 9 marks will be included in 0-9 class interval.

To ensure continuity, it is better to adopt 'exclusive' method of classification. However, where 'inclusive' method has been adopted, it is necessary to make an adjustment to determine the correct class-interval and to have continuity. This adjustment is done by the formula.

$$\text{Correction factor} = \frac{\text{Lower limit of Succeeding class – upper limit of preceding class}}{2}$$

The value so obtained should be subtracted from lower limits and added to upper limits of all classes. For example.

Wages	No. of Workers
10-19	10
20-29	18
30-39	35

The class interval is adjusted as follows : (20 -19)/2 = 0.5

Upper Class Boundary = Upper Class Limit +d
Lower Class Boundary = Lower class Limit –d where d = correction factor

Open-End Classes

Open-end classes are those in which lower limit of the first class and upper limit of the last class are not known

Income	No. of Persons
Less than 100	20
100-200	15
200-300	20
Above 300	9

Preparation of Continuous Frequency Distribution Table

If any instructions are given in the problem, such instructions must be followed. In the absence of instructions, the following guidelines can be followed.

1. See the smallest and largest values in the distribution and find the difference between them.
2. Decide the number of Classed in the range or 6 to 15 classes. The number of Classes can be arrived at by dividing the range with the desired class Interval, subject to the limits of 6 to 15

Surge's Formula

In order to determine the number of classes, Prof H.A Surges has suggested formula, which is described below.

$k=1+3.322 \log_{10} N$ where

k= number of classes (rounded off to nearest whole number)

N= Total frequency

Thus, if the total frequency (a total number of observation) is 100, then number of classes = $1+3.322 \log_{10} 100 = 1+3.3222 \times 2 (\log_{10} 10)$

$= 1+3.3222 \times 2 (\log_{10} 10 = 1+(3.322) \times 2 = 1+6.644 = 7.644 = 8$

This formula is useful in determining the number of classes, but fails in case the total frequency is too high.

3. In the absence of instructions, adopt Exclusive method
4. Class Interval should be fixed such that each class has a convenient mid-value
5. Prepare Frequency Table by putting class Intervals in Column 1, Tally Bars in Column 2 and Frequency in Column 3.

Illustration 2 : From a frequency distribution by taking a suitable class-interval for the following data giving the ages of 52 employees in a government agency.

67,34,36,48,49,31,61,34,43,45,38,32,28,61,29,47,36,50,46,30,46,32,30,33,45,49, 48, 41,53,36,47,47,30,46,50,28,35,35,38,46,43,34,36,62,69,50,28,44,43,60,38

(B.A. Madurai)

Note

The smallest value is 28 and the largest 69. Difference between the two values is (69-28) =41. If we take a class interval of 5, nine classes will be formed.

Solution 2

Frequency Distribution of the Age of Employee

Age (Year)	Tally Bars	Frequency
25-30	IIII	4
30-35	~~IIII IIII~~	10
35-40	~~IIII IIII~~	10
40-45	~~IIII~~	5

45-50	~~IIII~~ ~~IIII~~ III	13
50-55	III	4
55-60	-	-
60-65	IIII	4
65-70	II	2
Total		52

Illustration 3 : The following is an array of 65 marks obtained by students in a certain examination :

26	45	27	50	45	32	36	41	31	41	48	27	46
47	31	34	42	45	31	28	27	49	48	47	32	33
35	37	47	28	46	26	46	31	35	33	42	31	41
45	42	44	41	36	37	39	51	54	53	38	55	39
52	38	54	36	37	38	56	59	61	65	64	72	64

Draw up a frequency distribution table classified on the basis of marks with Class-Intervals of 5.

Solution :Frequency distribution of marks obtained by students

Class-Interval of marks	Tally Marks	Frequency
25-29	~~IIII~~ 11	7
30-34	~~IIII~~ ~~IIII~~	10
35-39	~~IIII~~ ~~IIII~~ 111	13
40-44	~~IIII~~ 111	8
45-49	~~IIII~~ ~~IIII~~ 111	13
50-54	~~IIII~~ 11	6
55-59	111	3
60-64	111	3
70-74	1	1
Total	-	65

Illustration 4 : The following data pertains to number of units produced by 50 workers in a factory. Using Surge's formula, classify the data into a frequency table.

162	258	237	231	228	159	242	189	168	189
243	196	287	222	105	138	128	62	53	91
192	230	246	183	242	215	171	220	153	208
152	295	239	220	116	166	102	178	137	211
206	212	275	271	290	128	59	180	296	218

Solution

Number of Classes (as per Surge's formula) = $1+3.322 \log_{10} 50$

= 1+(3.322*1.699) = 6.64 rounded off to 7.

Size of Class Interval = Highest Value – Lowest Value / Number of Classes

= (296-53) /7 = 34.7, rounded off to 35.

Thus, the Class Intervals would be 53-88, 88-123, 123-158, ...

Class-Interval of units Produced	Tally marks	Frequency
53-88	III	3
88-123	IIII	4
123-158	IIII I	6
158-193	IIII IIII I	11
193-228	IIII IIII I	11
228-263	IIII IIII	9
263-298	IIII I	6

Simple and Cumulative Frequency Series

Statistical series can be either imple or cumulative. In simple, frequency against each class interval or value is shown separately and Individually.

Illustrations 2 and 3 are examples of a simple frequency series. In cumulative series, the frequencies are progressively totaled and aggregates are shown. They can be shown by 'Less than' method or 'more than' method. 'Less than' method indicates to us the total number of observations below a particular value. Similarly, 'more than' method indicates to us the total number of observations more than a particular value. In case of 'less than' method we start with upper limits of the classes and go on adding the frequencies. In case of 'more than' method. We start with lower limit of the classers. Total Frequency is written against the first number, then frequency of the first class is deducted and the balance is written against the next class, then the process is repeated for the remaining classed.

Illustration 5 : Convert the following frequency distribution into cumulative frequency distribution by (i) "less than" method and (ii) 'more than' method.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	6	9	10	15	30	18	12

Solution 4 :Cumulative Frequency Table :

(i) "Less than" method (ii) 'More Than' Method

Marks less than	**No. of students**	**Marks more than**	**No.of students**
0	0	0	100
10	6	10	94
20	15	20	85
30	25	30	75
40	40	40	60
50	70	50	30
60	88	60	12
70	100	70	0

Explanation

The given frequency distribution table tells us the number of students falling within a particular range of marks. In preparing the 'less than' cumulative frequency table, we would like to know how many students scored below 10 marks, how many score below 20 marks etc. Students who scored below 20 marks will not only include students scoring below 10 marks, but also students scoring between 10 and 20 marks. Hence, the number of observations would be 6+9 i.e. 15. In case of 'more than' table, the number of students scoring more than 0, which is the lowest marks, will be equal to total number of observation i.e 100. The students who have scored between 0 to 10 will be excluded in calculating the no. of students scoring more than 10 marks. Hence, total observations will be 100-6=94. The rest of the problem can be similarly solved.

Conversion of Cumulative Frequency Distribution into simple Frequency

Distribution

A cumulative frequency distribution table can be converted into simple frequency distribution table by splitting the total number of observations. 'less than' 'more than' a particular value in appropriate class intervals. For example, if 'less than' a particular value in appropriate class intervals. For example, if 'less than' frequency table is given, students scoring marks below 100 will include students scoring between 90 to 100 and students

scoring below 90. Thus number of observations between 90 and 100 can be ascertained by deducting number of students scoring below 90 from total observations below 100. In case of conversion of 'more than' type of cumulative frequency table, observations falling in the class interval of 0 to 10 marks can be calculation by deducting total observations more than 10, from total observations more than '0'.

Illustration 6 : Convert the following cumulative frequency distribution into simple frequency distribution.

Marks above	0	10	20	30	40	50	60	70	80	90	100
No. of Students	80	77	72	65	55	43	28	16	10	8	0

Solution

Simple Frequency Distribution

Marks		**Frequency**
0-10	(80-77)	3
10-20	(77-72)	3
20-30	(72-65)	7
30-40	(65-55)	10
40-50	(55-43)	12
50-60	(43-28)	15
60-70	(28-16)	12
70-80	(16-10)	6
80-90	(10-8)	2
90-100	(8-0)	8
Total		80

Illustration 7 : Convert the following cumulative frequency into simple frequency table :

Income (₹) Below	500	600	700	800	900	1000
No. of Students	8	20	40	70	87	100

Solution :

Conversion of Cumulative frequency into simple frequency table.

Income (₹)		**No. of Employees**
400-500		8
500-600	(20-8)	12
600-700	(40-20)	20
700-800	(70-40)	30
800-900	(87-70)	17
900-1000	(100-87)	13
		100

Bivariate Frequency Distribution

Bivariate frequency distribution that classifies the given data with respect to two variable simultaneously. Let us understand this with the help of an example. Consider a group of 50 couples. The ages of all the 50 men and 50 women were collected. If we make a simple frequency distribution by age, we may get frequency distribution telling us how many people in the group of 100 are aged 18,19,20 ..... up to 27. We can take it one level lower and have two frequency distributions. The women were in the age-group of 18 to 22. The men were in the age group 22 to 27/ The first listing number women aged 18,19,20,21 and 22. The Second listing number of men aged 22,23,24,25,26 and 27.

However, if we consider each couple as a unit and wish to find out how many women aged 18 years husbands with ages 22,23,24,25,26,27 we are talking about a bivariate frequency distribution. The result of this exercise would be as under:

	Age of Husband					
Age of Wife	**22**	**23**	**24**	**25**	**26**	**27**
18	3	2	1	2	1	0
19	1	6	3	1	1	0
20	1	3	4	2	2	1
21	1	0	1	2	0	1
22	0	1	2	3	3	2

While the above example is of a discrete bivariate distribution, the same can also be prepared in case of continuous variables.

Tabulation

A Table is a systematic arrangement of statistical data in columns and rows. Rows are horizontal arrangements, whereas columns are vertical. Tabulation is systematic presentation of data in a form suitable for analysis and interpretation.

Classification vs. Tabulation

Classification and Tabulation are part of the process of data compilation. After data is collected, it is first classified and then tabulation. The following are the main points of difference between classification and tabulation.

1. Classification is the pre-requisite for tabulation. Tabulation begins where classification ends.
2. In Tabulation, the emphasis is on presentation of data. In classification, the emphasis is on highlighting the similarities and differences in data.
3. In case of classification, data is arranged according to attributes and variables. In case of tabulation, data is arranged in columns and rows.

Objectives of Tabulation

1. **Simplify Data :** The main objective of tabulation is to simplify a mass of information into its simplest form, making it easy to understand.
2. **Provide Clarity :** Tabulation provides clarity on data collected, Many of the questions that are related to the objective of the study can be easily answered through tabulation
3. **Facilitate Comparison :** Tabulation facilitates easy comparison. The two competing options that need to be evaluated can be placed side by side and compared.
4. **Trend Analysis :** Tabulation helps to catch the trend by more observation. Tabulated data is well placed to perform statistical calculation of trend and other features of data.
5. **Economy of Space and Time :** Presentation of data in tabular form saves space without any compromise on quality or completeness of data. Not only presentation, but usage of data also results in savings of time for the users of such data.
6. **Enables Easy Reference :** Tabulated information with appropriate title and number are easy to remember and convenient to refer at future date. Most Research papers will have a separate. Table of contents listing down the various table in which useful statistical data is presented and their page numbers.
7. **Detections of Errors and Omissions :** Particularly in case of numeric data, tabulation serves as a control. If any of the totals of rows or columns are not as they should be, it gets immediately known.

8. **Facilitate Statistical Processing :** It is only after tabulation that data becomes fit for statistical processing we can apply various statistical methods and techniques such as correlation, calculation of various measures of central tendency and dispersion etc, only after data is correctly tabulated.
9. **Clarity on Characteristics of data :** A concise tabular form clearly reveals the characteristics of data and highlights its significant characteristics.

Requisites of a Table (Rules for Preparing a Table)

1. The table should serve the purpose for which it is being prepared. By looking at the table, the user should be able to understand what the person preparing the table wishes to convey. Including data, though not relevant, just because it is available will defeat the purpose with which a table is prepared.
2. The table should be prepared in a systematic and logically organized manner.
3. The table should be clear in terms of its various captions and stubs. They should be clearly defined so that there is no ambiguity and hence, no possibility of over laps.
4. The table should be compact. It should ideally fit the size of the paper or screen (in case of video presentations). The column width should be adjusted or the table broken down into smaller tables so that at any point in time, it is not a strain on the user to go through the table.
5. Unless the need for accuracy is very high, the data can be rounded off or a higher unit of measurement can be used. This will ensure that the user is not bogged down by too many details and gets the 'broad' or 'big' picture.
6. The best practices with respect to all parts of the table stating unit of measurement, use of symbols for footnotes, (e.g, numbering of columns, etc), should be adopted.
7. The arrangement of various stubs and columns should follow a systematic method. It can be chronological (year, wise). Geographical, Alphabetical in times of size, or any other sequence. Within the sequence, ascending or descending order should be used consistently.
8. The presentation of the table should be appealing to the eye. Rows and Columns should be appropriately spaced. So that there is no cluttered appearance.
9. The table should be able to highlight the relationship between various items. Data that needs to be compared should ideally be placed side by side. If that is not possible, it should be placed at least within noticeable distance so that the two parts can be looked at the same time.
10. Data that needs to be emphasized can be highlighted by use of different color, font, font size or by simply circling it or putting it in a box. However, it is not advisable to use too much of styling.
11. The notations used should not be confusing. For example, N.A could mean 'Not Applicable' or 'Not Available' such notations should be avoided. It is also advised to avoid ditto marks, dashes and other symbols such as '0' that can be misunderstood for something else.
12. Abbreviations should be avoided. However, standard abbreviations and symbols can be used. For example, in measuring height, feet and inches are depicted by 'and' and it is universally understood. Similarly, 'Govt.' in place of 'Government' is also an example where abbreviation can be used.
13. If standard classifications are available, it is preferable to use them instead of creating a different set of classifications for a particular table.
14. It is always advisable to have an extra column for 'Remarks'. Text that can help the user in understanding the corresponding data can be written here.

15. The most important requisite is that the data being presented is 'error-free' Each element being presented should be verified at least twice to ensure its accuracy. The entire effort of tabulation will go waste. If the data presented is erroneous, even if all other requisites are satisfied.

Parts of a Table

1. **Table Number :** Indicates the serial number of the table. Table Numbers help in easy identification and reference of the table in future. Hence, they should be stated in clear and visible terms. It is preferred to write the Table Number at the top of the table, usually in center.
2. **Table Title :** Indicates the subject matter of the table. The title should be as unambiguous as possible. It should be clearly worded and indicate the nature of data contained in the table. The title should ideally indicate then 'what, where and How of the classified data and hint at the period of which it relates. Titles should not be too long, However, if the title is becoming very lengthy, it is advised to have a 'catch-line' with the rest of title stated in the next line. The table title should be very prominent in bold, and at the center, just below the Table Number.
3. **Captions :** refers to the headings of the columns. Captions usually have a heading and also sub-heads, depending on the type of data being tabulated. A caption should be brief, concise and self-explanatory. Care should be taken that each of the column names does not spill into other column areas. It is preferable to number the captions and the sub-heads for case of reference, particularly when calculations are involved.
4. **Sub :** refers to the headings of the rows. As with captions, rows can also have further sub-divisions. However, stubs do not need to be very brief. Often, the choice of arrangement is done in such a way that the more descriptive items are taken as stubs. As in case of captions, it is desirable to number the stubs as well.
5. **Body :** refers to the numerical information entered into the table. The entire statistical data that is to be presented constitutes the 'body'. Tabulation is an effort to present this 'body' in a more meaningful manner.
6. **Head Note :** It is a brief explanatory statement with respect to some or all parts of the table to facilitate better understanding of the nature of information provided in the table.
7. **Foot Note :** Footnotes provide further explanation with regard to the contents of the table. They may include exceptions to data, classifications etc.
8. **Source Data :** In case the table is not generated from primary data collected by the statistical investigator, the source from which data is collected is given at the bottom of the table. The source is stated below the footnotes. Stating the source adds credibility to the data. Usually, we state the 'original source' of data along with other indicators such as page number, table number, etc so that it is easy for anybody to refer to it at any time.

Types of Tabulation

1. **Simple and Complex :** This classification of tables is done on the basis of number of characteristics studied. A simple table studies only a single feature. It is also known as one-way table complex tables present two or more characteristics.

(a) **One-way Table :** It presents only one characteristics and helps in answering one or more independent questions with regard to that characteristic. The frequency distribution table prepared in Illustration 2 is an example of simple or one-way table.

(b) **Two-way Table (Double Tabulation) :** It contains sub division of a total and is able to answer two mutually dependent questions. Following is an example of a two-way table.

Table No. xx

Employee distribution according to Age and Sex of Employees

Age (year)	Male	Female	Total
25-35	4	10	14
35-45	10	5	15
45-55	11	6	17
55-6 5	3	1	4
65-75	2	0	2

(c) **Three-way table :** It sub-divides the total into three distinct categories. It is capable of answering three mutually dependent questions. For example male and female can be further classified into local and non-local employees. This type of tabulation is known as treble tabulation.

(d) **Manifold Tabulation :** If more than three characteristics are simultaneously shown, it is known as manifold tabulation. As the number of characteristics increases, there is more confusion and there may be loss of clarity. It is preferable to prepare a separate table.

General Purpose and Special Purpose Table

General-purpose tables or Reference tables, provide information for general use. They usually contain detailed information and are not constructed for specific discussions. Special purpose tables, or summary tables, or analytical tables provide information for a particular discussion and highlight relationship between different figures. Special purpose table are ordinarily taken from the general-purpose tables and emphasize the relationship, which the user wishes to stress. They are prepared with a stress on analysis. It should be designed such that the reader may easily refer to the table for comparison. Analysis concerning the particular discussion.

Original Table and Derived Table

Original tables contain data that is collected from an original source. They contain primary data and hence are also called as 'Primary' tables. They are also known as 'classification' tables. Data in such tables is 'raw' data that is not processed in any manner (such as rounding off, calculation of percentage, etc). Derived tables present data in a form that is desired from the original numbers. A table containing averages, percentages, ratios, etc is a derived table. These tables can present both primary and secondary data, but such data is already processed in some form.

Illustration 1 : Prepare a blank table to give as much information as possible of the summary results of the distribution of population according to sex and three religions at four age groups in North and South India. ***(B.A. Kerala Adapted)***

Classification vs. Tabulation

Classification and Tabulation are part of the process of data compilation. After data is collected, it is first classified and then tabulated. Thus, Classification is the pre-requisite for tabulation. Tabulation begins where classification ends. In Tabulation, the emphasis is on presentation of data. In classification, the emphasis is on highlighting the similarities and differences in data. Lastly, in case of classification, data is arranged according to attributes and variable. In case of tabulation, data is arranged in columns and rows.

Solution to Illustration 7

Table No.

Distribution of population according to sex, religion, region and age

Region	Religion	Below 15 years			15 – 25 years			25 – 35 years			35 years & above			Total		
		M	F	Total	M	F	Total	M	F	Total	M	F	Total	M	F	Total
North India	Hindu															
	Muslim															
	Christian															
South India	Hindu															
	Muslim															
	Christian															
	Total															

Statistical System in India

Statistics has been in use in India ever since the medieval period. However, there has been no formal organization responsible for collection and publication of statistics even during the British Raj. On the recommendations of the Bowley-Robertson Committee, the first attempt in this direction was made by establishing the office of the Economic Adviser to the Government of India in 1938. However a serious attempt was made to collect regular and reliable statistics only after independence. The preparation and successful implementation of the Five year plans underlined the need for reliable and adequate statistical information. Consequently, a number of organizations have been set up by the Central and State Governments.

The statistical system in India closely corresponds to the federal structure of our constitution. According to our constitution, subjects like Defense, Railways Currency and Foreign Exchange, foreign trade, income tax, customs and excise are in the Union list. The Central Government is responsible for the collection of statistics relating to these subjects. Similarly subjects like public health, agriculture, livestock, irrigation, forests an fisheries come under the State list. The State Governments are responsible for the collection of statistics relating to these subjects. Subjects like economic and social planning trade unions, social insurance, labor welfare, relief and rehabilitation come under the Concurrent list. The responsibility for collection of statistics relating to these subjects is with both the State and Central Governments. However, this demarcation is not very rigid. The Central Government acts as an apex organization by issuing directives and framing acts and rules. It acts as co-organization by issuing directives and framing acts and rules. It acts as a co-ordinating agency and publishes the statistics collected by the various State Governments on an all India basis.

EXERCISE –I

1. In a survey of 30 families in a village, the number of children per family was recorded and following data obtained.

6	5	2	8	3	1	7	8	7	0	2	4	6	7	2
5	2	3	4	2	3	5	5	3	4	3	4	4	6	3

Represent the data in the form of a discrete frequency distribution.

2. The class of 2004 of Osmania University had 30 students who scored the following letter grades.

A	B	D	C	C	B	D	C	B	C	B	C	B	A	B
C	B	A	D	D	C	B	B	D	D	A	B	C	C	B

Represent the data in the form of a discrete frequency distribution.

3. Following are the marks secured by 60 students. Arrange the data in a frequency table by 'Exclusive Method'.

56	71	18	31	56	40	81	64	59	49
42	12	67	7	39	76	48	7	80	84
34	58	61	28	45	0	63	61	11	81
12	24	75	64	68	43	58	79	36	83
09									