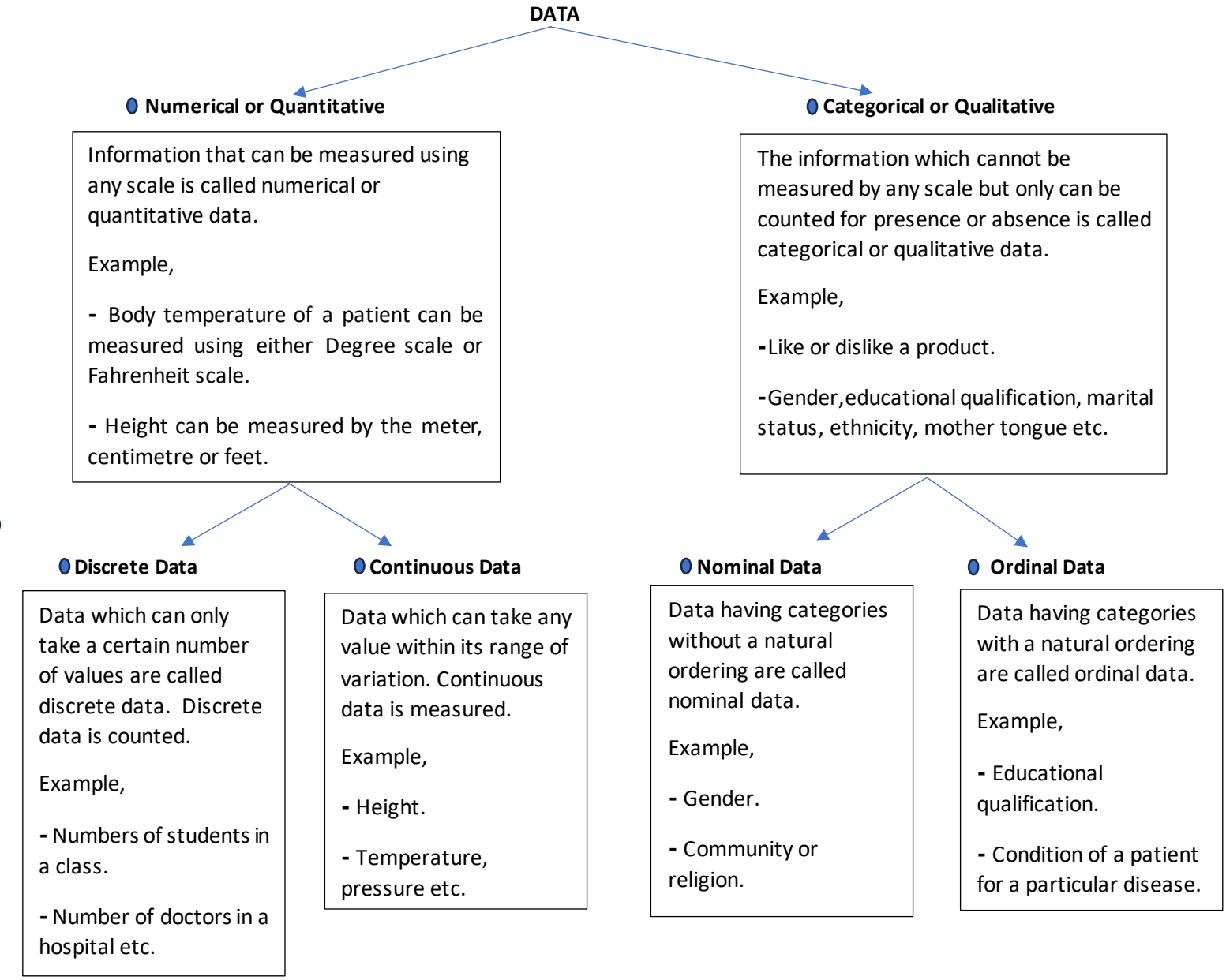
Statistical Data:

Information collected with a predetermined objective in mind, either in terms of numbers or attributes marked by "uncertainty and variability" are called statistical data. Uncertainty & variability are two major characteristics of Statistical Data. Not all quantitative data is statistical data. Example of statistical data – Suppose we study the 'Heights of students in a particular college'. Here we can't predict the height of an individual with certainty & there will be variation in heights of students. Counter Example: Multiplication table in a tabular form is a quantitative data, but since there is no uncertainty & variability involved in the data so it's not a Statistical Data.

Types of Statistical Data:

- Depending upon the **nature** of information, data can be broadly classified into two types:
- Numerical or Quantitative Data
- Categorical or Qualitative Data



Depending upon how information is **counted or described**, we have two more classification on data:

- Frequency Data
- Non-frequency Data

DATA

Frequency

When we are interested in knowing how frequently each of the different values of a variable occurs in a data set it is said be frequency data.

Example,

- Students' test scores out of 100,

Test Score Range	No. of Students
60-69	7
70-79	10
80-89	7
90-100	6

- Favourite fruits among a group of people,

Fruit	No. of people
Apple	12
Banana	15
Orange	8
Grape	5

Non-frequency

When we are not concerned about the frequency of the values of a variable in a data set it is said be non-frequency data.

Example,

- Raw test score of 5 students out of 100,

Student	Test Score
Saurav	96
Sangeeta	72
Sujeet	85
Sakuntala	68

- Favourite fruits for 5 individual people,

Person	Favourite Fruit
Saurav	Apple
Sruti	Banana
Sakuntala	Orange
Sujeet	Grape

Cross-sectional Data

When data is collected by observing many subjects (such as individuals, firms, countries, or regions) at the same or approximately the same point in time, or without regard to differences in time is known as cross-sectional data.

Example,

- Production of rice in different states of India

States	Production of Rice (in tons)
Bihar	10
West Bengal	16
Assam	12
Orissa	14

Time series/historical/chronological/longitudinal Data

When data are arranged according to the order of time, the data is known as time series data or historical data or chronological data. Here the values of one or more variables are given for different points or periods of time. In such a case, we are interested in the relationship between the time and the variable.

Example,

- Production of rice in Assam

Years	Production (in tons)
1991	10
1992	10.5
1993	11
1994	12

Spatial/Geographical Data

When data is collected in accordance with geographical region it is called spatial or geographical data. Place is the dominating factor here which can be countries, states, districts etc.

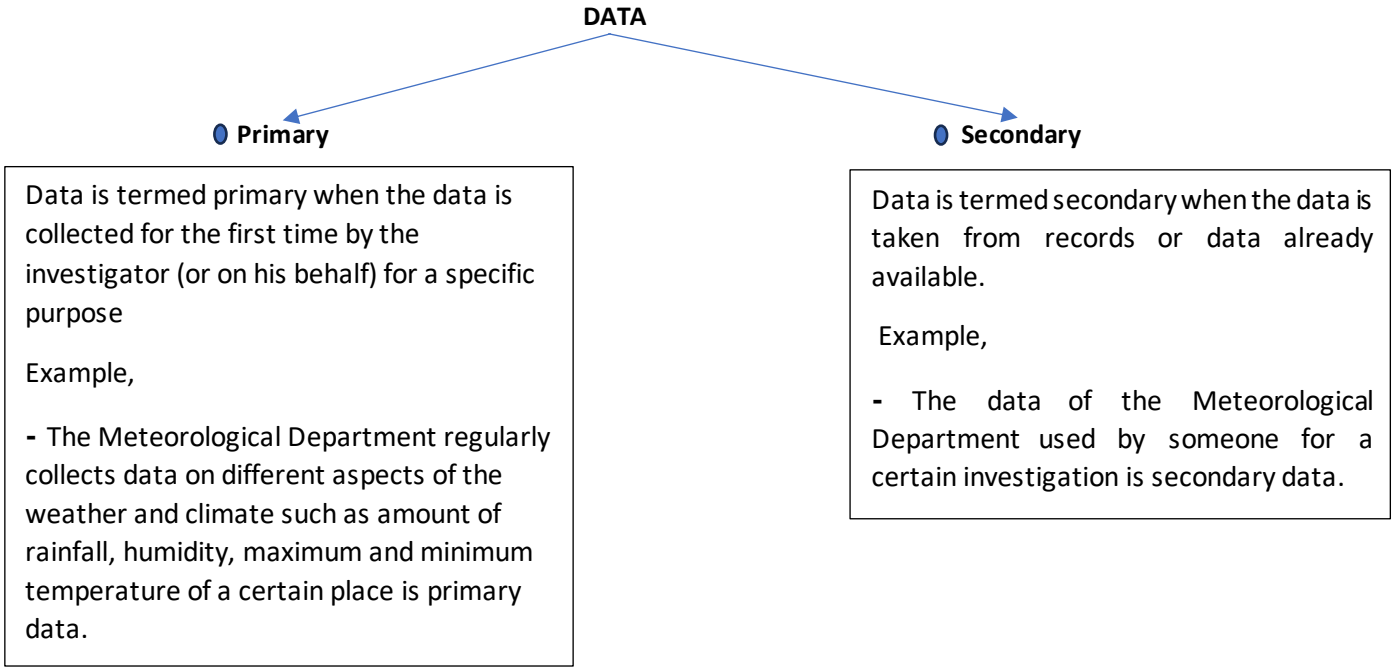
Example,

- Population of 5 districts of Assam according to 1991 census,

Districts	Population ('000)
Dibrugarh	750
Golaghat	450
Jorhat	700
Tinsukia	600

Depending upon the way how information is **collected**, we have two more classification on data:

- Primary Data
- Secondary Data



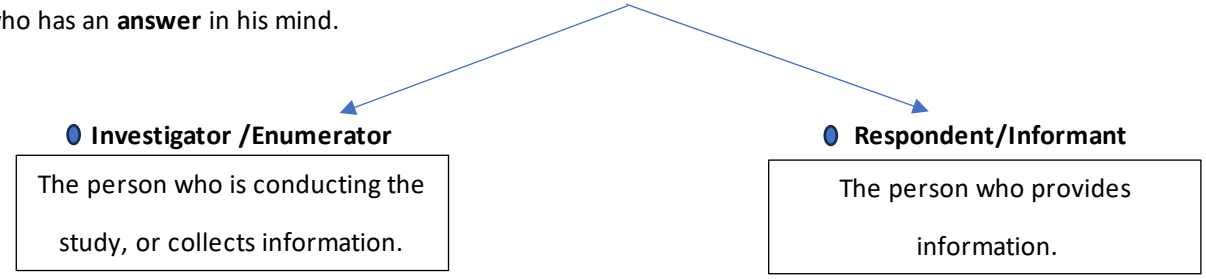
Distinguish between Primary data and Secondary data:

(1) Primary data are those which are to be collected for the first time by the investigator (or on his behalf) and the therefore, it is original in nature, whereas Secondary data are those which do not originate from the investigator (or from the field of enquiry) but which are obtained from someone else's records.

(2) Primary data may be used with greater confidence because the investigator will himself decide upon the coverage of the data, whereas secondary data is not so reliable. The secondary data may contain mistakes due to errors in transcription made when figures were copied.

Types of people involved in data collection:

The process of collecting data clearly gets divided between **two types of people** - one who has a **question** in his mind and one who has an **answer** in his mind.



Population and sample:

Population

Universe of concern or aggregate of the entire group of items, individuals, objects or subjects under study.

Based on numbers

Finite

If a population consists of a countable number of elements.

Example,

- Consider a class of 50 students. To study their average height, the population size is limited to 50. You can list and count all the students, making it a finite population.

Infinite

If a population consists of an uncountable or exceedingly large number of elements, so large that it is practically impossible to count or list all the elements.

Example,

-Population of sand particles in a river beach.

Based on subjects

Real

If a population consists of elements that actually exist or have existed at a specific point in time and are observable.

Example,

- Population of books in a bookstall.

Hypothetical

If a population consists of potential observations or outcomes that could occur under certain conditions but do not exist as a concrete group of elements at the time of analysis.

Example,

- Population of outcomes while rolling a die can be considered as hypothetical population as it doesn't exist physically as a fixed group. Instead, it represents a conceptual set of all possible outcomes of the rolling process.

Sample

Few items selected from the universe to know about the characteristics of the universe.

Example,

- Consider a class of 100 students. To study their average height, if we select 30 students and by measuring those selected 30 students infer about the average height of 100 students then those 30 students constitute a sample.

Methods of data collection:

There are two principal methods of data collection. Through a **census**, through a **sample survey**. Census implies complete enumeration of each and every element of the source. Data obtained by taking relevant measurement or observation of each and every element of the source constitute census data. When only some selected elements of the source (selected according to some valid procedure) are taken and measurement or observations of these selected elements are recorded, the data is said to be collected through a sample enquiry and is said to be sample data. For example, if we want to find out the average weight of all the 90 persons of a locality in a census enquiry, we will measure the actual weights of all the 90 persons and the total will be divided by 90 which will give us the average weight. Whereas, in a sample enquiry, we will take a sample of suitable size (say 20 or 25 persons) and divide the aggregate by the respective number to obtain the estimates for the population average.

Sample survey vs census:

The advantages of Sample Survey method over the Census method of enquiry are as follows:

(1) It requires less time, labour and money because it is based on a part of the population.
(2) Greater scope: As there is a possibility to collect more information in a sample enquiry than in a complete count.
(3) Greater Accuracy: It is possible to engage better trained personnel for collection of data in the case of a sample enquiry than in a complete count. Processing of data is also much easier with sample data. All these factors lead to greater accuracy in data collected.
(4) If the test is destructive, sample survey is the only way e.g. testing crackers and explosives.
(5) If the population is hypothetical in nature, sample survey is the scientific way.

Disadvantages of sampling:

(1) If the population is small, sampling is unnecessary.
(2) If information is required for each and every one then sample survey cannot be done.
(3) Sample survey requires trained personnels and sophisticated tools, but this might not be possible many times.
(4) If proper care is not taken then sample survey might result in error.

Tools of data collection:

Statistical data are frequently obtained by a process in which the desired information is obtained from the source, either by having an enumerator visit to the informant, ask the necessary questions and enter the replies on a schedule, or by sending to the informant a list of questions (sometimes called a questionnaire) which he may answer at his convenience.

Questionnaire: The term 'questionnaire' means a list of certain systematically arranged questions relating to the subject of enquiry. It is necessary that questionnaire is designed with due care so that necessary data may be easily collected.

Schedule: In the schedule one finds a list of questions, on which information will be collected, the exact forms of the questions to be put to the informants are not given to the respondent and task of questioning, explaining the desired information is left to the investigator.

Questionnaire vs Schedule:

(1) Questionnaire is a list of questions which are answered by the respondent himself in his own handwriting while schedule is the method of getting answers to the questions in a form which are filled by the interviewer in a face-to-face situation.
(2) Questionnaire cannot be applied when the informants are illiterate whereas schedule can be applied even when the informants are illiterate.

Framing of Questionnaire or Schedule:

Great care is to be taken in drafting a questionnaire or schedule, as this is the medium through which information is collected. There are a few general points which should be borne in mind:

(i) The questions put should be clear, concise and unambiguous.

(ii) Delicate questions are to be put with greater care, often indirect questions should be put to get answers to some pertinent point. It is sometimes desirable to avoid very delicate questions.

(iii) The size of the questionnaire/schedule should be small. It saves time, both for the enumerator and the respondent. A large questionnaire is likely to exhaust the patience of the respondent.

(iv) There should be a natural, logical order in which questions are put.

(v) It should be noted that the information collected through questions should be such that it is usable.

Sample questionnaire:

Name : _____

Age (in years) : _____

Sex :

Male ☐ Female ☐ Transgender ☐

Educational qualifications :

Matric ☐ Intermediate ☐ Graduation ☐

Postgraduation Any other technical qualification

Marital Status :

Married ☐ Unmarried ☐ Divorced ☐ Widow ☐

Nature of Job :

Permanent ☐ Contractual ☐

Category :

Skilled ☐ Semiskilled ☐ Unskilled ☐

Designation : _____

Department : _____

Monthly income (in Rs) : _____

Methods of collection of primary data:

Direct Personal Investigation: In this method, the enumerator himself personally goes to the person/source and collects the necessary information. This method is suitable when the field of enquiry is small and a high degree of accuracy is desired. But the method is costly and time consuming. Also, the personal bias of the enumerator may enter into the data.

Indirect Oral Investigation: This method is used when informants are reluctant to supply information or when it is very difficult to contact them directly. Under such situation indirect evidence of third parties who are in touch with the facts

desired, is recorded. This method is economic and time saving but the information obtained from the third parties may not be reliable at times.

Investigation through local agencies: Here there is no formal collection of any data but the local correspondents residing in different areas collect the information and report the same to the authority. This method is adopted by newspapers and periodicals. This method is very cheap and a wide area can be covered but the reliability of data may be a matter of doubt.

Mailed Questionnaire Method: In this method the schedules of questions known as questionnaires are mailed to the informants with the request of quick response after duly filled in. This method is the least expensive method where the bias of the investigator is completely ruled out. However, this method cannot be applied when the informants are illiterate and success of the method depends upon the co-operation of the respondents.

Schedules sent through enumerators: In this method the enumerators go to the respondents with the schedule and record their replies. Population census is done by this technique. This method can be applied even when the informants are illiterate. By this method maximum possible results can be obtained. This method is quite expensive and time consuming. However, the accuracy largely depends upon the skill and efficiency of the investigator.

Types of Sampling:

The technique of selecting a sample is of fundamental importance in sampling theory and usually depends upon the nature of the investigation. The sampling procedures which are commonly used may be broadly classified under the following heads:

- *Probability*: Probability sampling is the scientific method of selecting samples accordingly to some laws of chance in which each unit in the population has some definite pre-assigned probability of being selected in the sample. The following are the different types of probability sampling:

    1. Simple Random Sampling: In this method, every individual or item in the population has an equal chance of being selected. For example, if we pick names from a box without looking, that is simple random sampling.

    2. Stratified Sampling: Here, the population is divided into smaller groups called strata based on some characteristics (like gender or income), and samples are taken from each group. This ensures fair representation of all groups. For example, suppose a school has 600 students where 300 are boys and 300 are girls. If we want to study students' study habits, we first divide them into two groups (boys and girls) and then randomly select 30 boys and 30 girls. This ensures both groups are fairly represented.

    3. Systematic Sampling: In this method, we select every k^{th} item from a list after choosing the first one at random. For example, if we choose every 5^{th} student from a roll list, it is systematic sampling.

    4. Probability Proportional to Size (PPS) Sampling: In this type, bigger groups or units have a higher chance of being selected. For example, if we are selecting villages, larger villages get a higher chance than smaller ones.

    5. Cluster Sampling: The population is divided into clusters or groups (like schools or villages), and some clusters are randomly chosen. Then all individuals from those selected clusters are studied. For example, let us imagine a district has 50 schools, and we want to survey students' eating habits. Instead of visiting all 50 schools, we randomly select 5 schools (these are our clusters) and we survey all the students in those 5 schools.

- *Non-Probability or Judgement or Purposive sampling*: In this method the investigator selects those units of the population in the sample which he/she thinks to be the true representative of the population, keeping the definite purpose in view. Here the chances of the selection of some units of the population in the sample is very high while the chances of some other units are very low. This method is highly subjective since the selection of the unit entirely depends on the judgement of the investigator.

For example, if a teacher is asked to select five students from his class for a particular job, he will select those five students which he feels fit for that job.
Judgement sampling suffers from nepotism and favouritism. It suffers from prejudices of the investigator. The most serious objection is that it is impossible to compute the degree of precision of the estimates from the sample values.

- *Mixed Sampling:* If the samples are selected partly according to some laws of chance and partly according to a fixed rule, they are called mixed samples and the method of selecting such samples is known as mixed sampling. The merits of this sampling are the mixture of the merits of both sampling. Selection of units is more reliable in this method because that is the representation of various stages of population.