# Unit -3

Sampling-

When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a **sample.**

A sample is a smaller set of data that a researcher chooses or selects from a larger population using a pre-defined selection method. These elements are known as sample points, sampling units, or observations.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

- **Probability sampling** is a method of deriving a sample where the objects are selected from a population-based on probability theory. This method includes everyone in the population, and everyone has an equal chance of being selected. Hence, there is no bias whatsoever in this type of sample.

Probability sampling can be further classified into four distinct types of samples. They are:

**1.** simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: You want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number

to every employee in the company database from 1 to 1000 and use a random number generator to select 100 numbers.

**2.** systematic sampling

Systematic sampling is like simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.
If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

**3.** stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example: The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

**4.** cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

**Non-Probability Sampling** This type of sampling is used for preliminary research where the primary objective is to derive a hypothesis about the topic in research. Here each member does not have an equal chance of being a part of the sample population, and those parameters are known only post-selection to the sample.

**1. convenience sampling**

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias.

Example: You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

**2.voluntry response sampling**

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g., by responding to a public online survey).

Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others, leading to self-selection bias.

Example: You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

**3. purposive samling**

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments.

Example: You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

**4. Snowball sampling**

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias.

Example: You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

**5. quota sampling**

Quota sampling relies on the non-random selection of a predetermined number or proportion of units. This is called a quota.

You first divide the population into mutually exclusive subgroups (called strata) and then recruit sample units until you reach your quota. These units share specific characteristics, determined by you prior to forming your strata. The aim of quota sampling is to control what or who makes up your sample.

Example: You want to gauge consumer interest in a new produce delivery service in Boston, focused on dietary preferences. You divide the population into meat eaters, vegetarians, and vegans, drawing a sample of 1000 people. Since the company wants to cater to all consumers, you set a quota of 200 people for each dietary group. In this way, all dietary preferences are equally represented in your research, and you can easily compare these groups. You continue recruiting until you reach the quota of 200 participants for each subgroup.

## DATA COLLECTION-

Data collection is the process of gathering, measuring, and analysing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. The data collection methods allow a person to conclude an answer to the relevant question. Most of the organizations use data collection methods to make assumptions about future probabilities and trends. Once the data is collected, it is necessary to undergo the data organisation process.

The data collection methods are divided into two categories-

1) Primary Data Collection Methods
2) Secondary Data Collection Methods

Primary data collection methods-Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys, or observations. The primary data collection method is further classified into two types such as

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Let us discuss the different methods performed to collect the data under these two data collection methods.

## Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

## Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data such as-

## Observation Method

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

## Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

## Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple
- Should follow a logical sequence.
- Provide adequate space for answers.
- Avoid technical terms.
- Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent.

## Schedules

This method is like the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

## Organisation of data-

It refers to arrangement of figures in such a form that comparison becomes easy, and conclusion can be drawn. Organization of data means classification, tabulation, graphical presentation and diagrammatic presentation of data.

**Classification-** It is the process of arranging things in groups and classes according to their resemblances and affinities.

1. **Objective of classification**
   a] To simplify complex data.
   b] To facilitate understanding.
   c] To facilitate comparison.
   d] To make analysis and interpretation easy.

e] To arrange and put the data according to their common characteristics.

**Characteristic of Good Classification**

1. Comprehensiveness
2. Clarity
3. Homogeneity
4. Suitability
5. Stability
6. Elastic

2. Concept of variable - A characteristic or a phenomenon which is capable of being measured and changes its value overtime is called variable.

3. A) Discrete Variable - Those variables that increase in jumps or in complete numbers. (No fraction is possible) Eg. Number of students in a class, Number of cars in a show room etc. (1,2, 10, or 15 etc.)
   B) Continuous Variables that assume a range of values or increase not in jumps but continuously or in fractions are called continuous variables. E.g., Height of the boys – 5'1'', 5'3'' and so on, Marks in any range 0-10, 10-15, 15-20
4. Statistical series systematic arrangement of statistical data Raw data: Data collected in original or crude form.
   Series: Arranging of raw data in different classes according to a given order or sequence is called series.

**Data analysis**

Data analysis is the process of cleaning, changing, and processing raw data and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graph.

Why Data Analysis is Important?

- Better Customer Targeting: You don't want to waste your business's precious time, resources, and money putting together advertising campaigns targeted at demographic groups that have little to no interest in the goods and services you offer. Data analysis helps you see where you should be focusing your advertising and marketing efforts.
- You Will Know Your Target Customers Better: Data analysis tracks how well your products and campaigns are performing within your target demographic. Through data analysis, your business can get a better idea of your target audience's spending habits, disposable income, and most likely areas of interest. This data helps businesses set prices, determine the length of ad campaigns, and even help project the number of goods needed.
- Reduce Operational Costs: Data analysis shows you which areas in your business need more resources and money, and which areas are not producing and thus should be scaled back or eliminated outright.
- Better Problem-Solving Methods: Informed decisions are more likely to be successful decisions. Data provides businesses with information. You can see where this progression is leading. Data analysis helps businesses make the right choices and avoid costly pitfalls.
- You Get More Accurate Data: If you want to make informed decisions, you need data, but there's more to it. The data in question must be accurate. Data analysis helps businesses acquire relevant, accurate information, suitable for developing future marketing strategies, business plans, and realigning the company's vision or mission.

**Types of Data Analysis-**

Descriptive Analysis-

Descriptive analysis is used to summarize and describe the main features of a dataset. It involves calculating measures of central tendency and dispersion to describe the data. The descriptive analysis provides a comprehensive overview of the data and insights into its properties and structure.

## Inferential analysis-

The inferential analysis is used statistical models and testing to make inferences about the population parameters, such as the mean or proportion. This analysis involves using models and hypothesis testing to make predictions and draw conclusions about the population.

## Predictive analysis-

Predictive analysis is used to predict future events or outcomes based on historical data and other relevant information. It involves using statistical models and machine learning algorithms to identify patterns in the data and make predictions about future outcomes.

## Prescriptive analysis-

The prescriptive analysis is a decision-making analysis that uses mathematical modelling, optimization algorithms, and other data-driven techniques to identify the action for a given problem or situation. It combines mathematical models, data, and business constraints to find the best move or decision.

## Data mapping-

Data mapping is the process of connecting a data field from one source to a data field in another source. This reduces the potential for errors, helps standardize your data and makes it easier to understand your data. Data mapping helps us visualize and connect data fields

much like maps can help us visualize the best way to get from point A to point B. And just like taking the wrong turn can mean trouble when you're travelling, data mapping errors can negatively impact your mission-critical data management initiatives.

Data mapping provides a visual representation of data movement and transformation. It is often the first step in the process of executing end-to-end data integration. Data integration brings together data from one or more sources into a single destination in real time. You need data mapping to understand your data integration path and process. Given the complexity and volume of data in today's enterprise, data mapping has become more critical than ever, and it requires intelligent, automated solutions for success.

With a backdrop of exploding data volume and variety in the modern enterprise, it's important to decrease the potential for data errors while increasing the ability to deliver actionable data insights. The data visualization process integrates multiple data sources into data models so you can simplify and combine dispersed data sources.

Data systems each store data in their own way. To analyse and understand your data, use data mapping to standardize data across your enterprise. Data mapping helps ensure that complex data management processes — like data migration, data integration and master data management — yield quality data insights. To automate business processes, you need to integrate data from one application to another. Data mapping bridges the gap by synching data from one format to another.

Business analytics benefit from data mapping. Combining data sets from different sources gives you a holistic view and context for your data. Data mapping can identify subject records across all your data sources. Then, it matches and links records across sources and systems to create a 360-degree view of each individual data subject.

Understanding data at such a granular level enables you to achieve deeper insights that can enhance your organization's decision-making capabilities, giving you a competitive edge.

Data mapping provides the ability to link all data about an individual's attributes. This helps you establish a single source of truth. Data mapping enables the smooth flow of data through different systems, applications, and services. It is a critical element of any data privacy framework. Given today's changing privacy regulations, automated, reliable data mapping helps you address crucial data access and compliance requirements. Data mapping provides visibility into end-to-end data lineage. It also supports data governance and makes it easier to apply use consent and other rights.

## Parametric and Non-Parametric-

### Parametric -

Mohanty and Misra (2016, page 675) defined parametric statistical tests as "a test whose model specifies certain conditions about the parameters of the population from which the research sample was drawn".

Salkind (2014, page 468) described parametric statistics as "statistics used for the inference from a sample to a population that assume the variances of each group are similar and that the sample is large enough to represent the population". The definitions highlight the homogeneity of the variance as well as also focuses on the sample being representative of the population.

Statistical techniques like Pearson's Product Moment Correlation, t test and ANOVA can be termed as parametric statistical techniques.

### Non-Parametric-

Non-parametric statistical techniques are distribution free test. In other words, these tests are not based on data that is normally

distributed or any such assumption (as is the case with parametric statistical techniques). Non-parametric statistics can also be described as tests that do not involve testing of hypotheses related to population parameters (King and Minium,2013. The model of the non-parametric statistical tests does not state any conditions about the population parameters from which the sample for research has been taken.

Assumptions of Parametric Statistics

The following are the assumptions of parametric statistics:

1. The population from which the research sample is taken is to be normally distributed: A normal distribution is a continuous probability distribution for a variable. It is also known as Gaussian/ Gauss or LAPlace – Gauss distribution. The normal distribution is determined by two parameters, mean and variance. When the normal distribution is represented in form of a graph, it is known as normal probability distribution curve or simply normal curve. A normal curve is a bell-shaped curve, bilaterally symmetrical and is continuous frequency distribution curve. Such a curve is formed because of plotting frequencies of scores of a continuous variable in a large sample. The curve is known as normal probability distribution curve because its y ordinates provide relative frequencies or the probabilities instead of the observed frequencies. A continuous random variable can be said to be normally distributed if the histogram of its relative frequency has shape of a normal curve.

2. The variables are measured in terms of interval or ratio scale: As you have studied in BPCC104, there are four scales of measurement nominal, ordinal, interval, and ratio. In case of parametric statistics, the variables are measured in terms of the interval and ratio scale. Interval scales are similar to the ordinal scale as the categories can

be ranked and are exclusive as well, but the degree of difference between two participants is same. Ratio scale has all the properties of all the scales, nominal, ordinal and interval. Further, it also has an absolute zero, that indicates presence or absence of certain property or characteristics.

3. The observations need to be independent of each other: If any case is included or deleted from the sample, this should not have an effect on the results of the research. Thus, is a certain case is selected from the population and the same is included in the sample should not have an influence the chance of inclusion of any other case and a score that is given to certain case should not be biased by a score assigned to another case (Mohanty and Misra, 2016). This point mainly focuses on the random selection of sample from the population and an unbiased assignment of scores to the observations.

4. The population from which the research sample is taken needs to display a variance that is homogeneous: It is important that population from which the research sample is taken displays variance that is homogeneous or display variance that is same or in certain cases have a ratio of variance that is known. To understand the term variance, it is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean. The variance (s2 ) or mean square (MS) is the arithmetic mean of the squared deviations of individual scores from their means. In other words, it is the mean of the squared deviation of scores. Variance is expressed as V = SD2

Assumptions of Non-parametric Statistics

It is often possible that, the assumptions and conditions as stated under parametric statistics are not met. In such a situation, non-parametric statistics can be used.

The following are the assumptions of non-parametric statistics:

1. There is no assumptions of normality of the data.

2. The variables are measured in terms of nominal or ordinal scales: We discussed under parametric statistics that the variables are measured in terms of interval and ratio scales. In non-parametric statistics, the variables are measured in terms of nominal and ordinal scales. Nominal scale can be used to measure variables that are qualitative as well as exclusive in nature. Whereas, ordinal scale involves ranks, that is, the data can be assigned ranks based on whether they are less or more, low or high, bad or good and so on and the data is ranked in terms of its magnitude.

3. The observations need not be independent of each other: Under nonparametric statistics, as such there is no assumption regarding the independence of relationship between the data.

4. The population from which the research sample is taken from need not display homogeneous variance: The sample can be heterogeneous, and the variance could be heterogeneous or no assumption is made with regard to the variance.

Data interpretation-

Data interpretation is the process of reviewing data and drawing meaningful conclusions using a variety of analytical approaches. Data interpretation aids researchers in categorizing, manipulating, and summarising data in

order to make sound business decisions. The end goal for a data interpretation project is to develop a good marketing strategy or to expand its client user base.

## Importance of Data Interpretation-

### Informed decision making

To act and adopt new processes, the management board must evaluate the data. This underlines the need for well-evaluated data and a well-organized data collecting method. A choice is only as good as the information that went into making it. Industry leaders that make data-driven decisions have the opportunity to differentiate themselves apart from the competition. Only when a problem is recognized and a goal has been established will the most decisive steps be taken. Identification, thesis formulation, data collecting, and data communication should all be part of the data analysis process.

### Identifying trends and anticipating demands

Users may employ data analysis to gain useful insights that they can use to foresee trends. It would be based on the expectations of the customers. When industry trends are detected, they may be used to benefit the whole industry.

For Example, people are more concerned about their health post covid, hence people are more likely to buy an insurance policy. The fundamental datasets for data analysis, data cycle of collection, evaluation, decision-making and monitoring should be followed by all next-gen companies.

### Cost efficiency

Many business experts do not consider data interpretation to be an expenditure, despite the fact that many organisations invest money in

it. Instead, this investment will assist you in lowering expenses and increasing the efficiency of your company.

## Methods of Data Interpretation-

Data interpretation method is a way to analyze and help people make sense of numerical data which has been collected, analyzed and presented. When data is collected, it normally stays in a raw form which may be difficult for the normal person to comprehend and that is why analysts always try to break down the information gathered so that others can make sense of it.

For instance, when Founders present their pitches to his or her potential investors, they do that by interpreting the data such as market size, growth rate and so on for better understanding. There are 2 principal methods by which data interpretation can be done:

1. Qualitative methods
2. Quantitative methods

## Qualitative Data Interpretation Method

Qualitative data interpretation method is used to analyze qualitative data which is often termed as categorical data. This approach uses texts, rather than numbers or patterns to represent data. Qualitative data requires first to be coded into numbers before it can be analyzed. As the texts are usually cumbersome and take more time. Coding done by the analyst is also documented so that it can be reused by others and also examined further.

There are 2 main types of qualitative data, such as nominal and ordinal data. These two data types are both performed using the same method, but ordinal data interpretation is easier than that of nominal data.

In most of the cases, ordinal data is usually labeled with numbers throughout the process of data collection, and so many times coding

may not be required. This is different from nominal data which still requires to be coded for proper interpretation.

Types of Data Interpretation

The various types of Data Interpretation are given below:

1. Tabular DI
2. Pie Charts
3. Bar Graph
4. Line Graph
5. Caselet DI

**1. Tabular DI:** In Tabular DI, data is provided in horizontal rows and vertical columns called tabular form. A table is one of the simplest and most convenient tools used for summarizing data and presenting it in a meaningful way. In a table, data is arranged systematically in columns and rows. While reading a table, the following parts need to be given careful observation.

- **Title of the Table:** It gives the description of the content of the table and precisely describes the kind of data, measurements and the period for which it occurred.
- **Column Heading:** This defines the information contained in the various columns with specifications of the unit of measurement in some cases.
- **Head Note:** In general, the unit of measurement is specified in the head note.
- **Footnote:** These are used to point out any exceptions in arriving at the data.

**2. Pie Charts:** It is a circular chart divided into various sectors. The sectors of the circle are constructed in such a way that the area of each sector is proportional to the corresponding values of information provided. In pie charts, the total quantity is distributed over a total angle of 360° or 100%.

Pie graphs have the shape of a pie and each slice of the pie represents the portion of the entire pie allocated to each category. Here the data could be presented and converted into 360 degrees or in percentages or in fractions. Many times, Statisticians may use exact figures against these sectors inside or outside as the case may be. Pie charts can be classified into two main types such as Exploded Pie Chart and Doughnut Pie Charts.

**3. Bar Graph:** In Bar Graph, data is represented as horizontal or vertical bars. One of the parameters is given on the x-axis and other on y-axis. Here we need to understand the given information and thereafter answer the given questions. A bar graph or a bar chart presents the grouped data with the help of rectangular bars. These bars are either horizontal or vertical and their lengths are proportional to the value that they represent.

There are 2 axes in the graph in which one represents particular categories being compared and the other axis shows a discrete value. Those bar graphs in which clustered groups of more than one bar are presented are known as grouped bar graphs, And, bar graphs in which bars are divided into sub-parts to show cumulative effect are known as cumulative bar graphs or stacked bar graphs.

**4. Line Graph:** A line graph shows the quantitative information or a relationship between two changing quantities with a line or curve. We are required to understand the given information and thereafter answer the given questions. A line graph or a line chart is a geographical representation of the change in two variables over a period of time. A line graph is created by connecting various data points.

Each data point is obtained as a result of plotting a point when we are given the value of two variables such as one independent variable and one dependent variable. Line graphs are a small but important part of data interpretation. In line graph questions, candidates are provided with certain data in the form of a line graph. The data may be related

to various categories such as the following, Average income and expenses, Comparing pie charts, population or demographics study, demand and supply, funds, distribution and utilization etc.

**5. Caselet DI:** In Caselet DI, a long paragraph is provided and with that as the basis, some set of questions are asked. We need to understand the given information and then answer the given questions.