CSE 2027-Fundamental of Data Analysis

Module: 3: Data Collection, Processing and Analysis

Collection of Primary Data(Observation Method, Interview Method, Collection of Data through Questionnaires ,Collection of Data through Schedule) Difference between Questionnaires and Schedules, Some Other Methods of Data Collection, Collection of Secondary Data ,Difference between Survey and Experiment Processing Operations, correlation.

> *80 percent of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20 percent to actually perform analysis...*
>
> IBM Data Analytics

Definition

- Data collection is a term used to describe a process of preparing and collecting data. Systematic gathering of data for a particular purpose from various sources, that has been systematically observed, recorded, organized.
- Data are the basic inputs to any decision making process in business

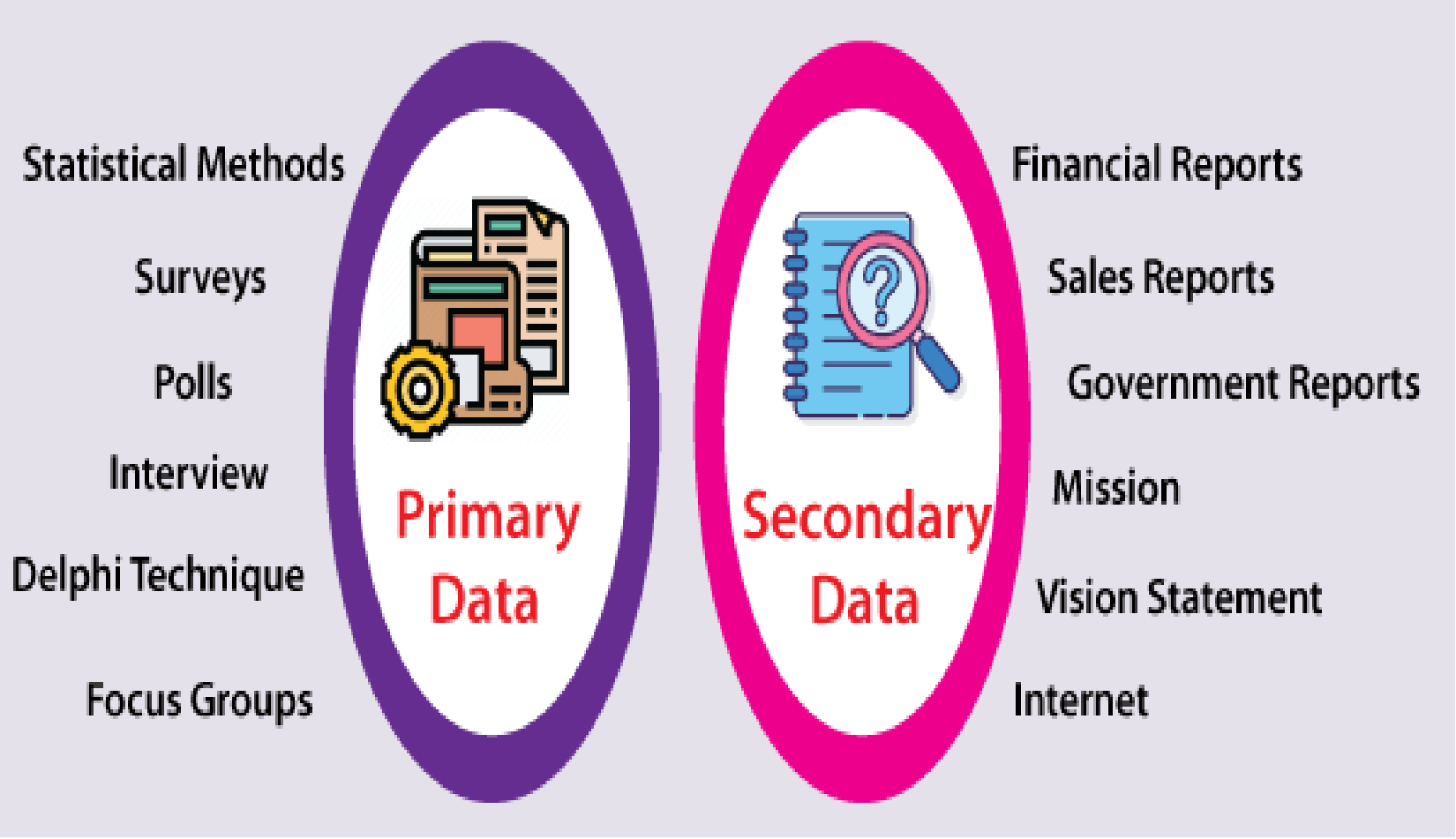
PURPOSE OF DATA COLLECTION

- To obtain information
- To record
- To take decisions about important issues
- To pass information on to others

CLASSIFICATION OF DATA

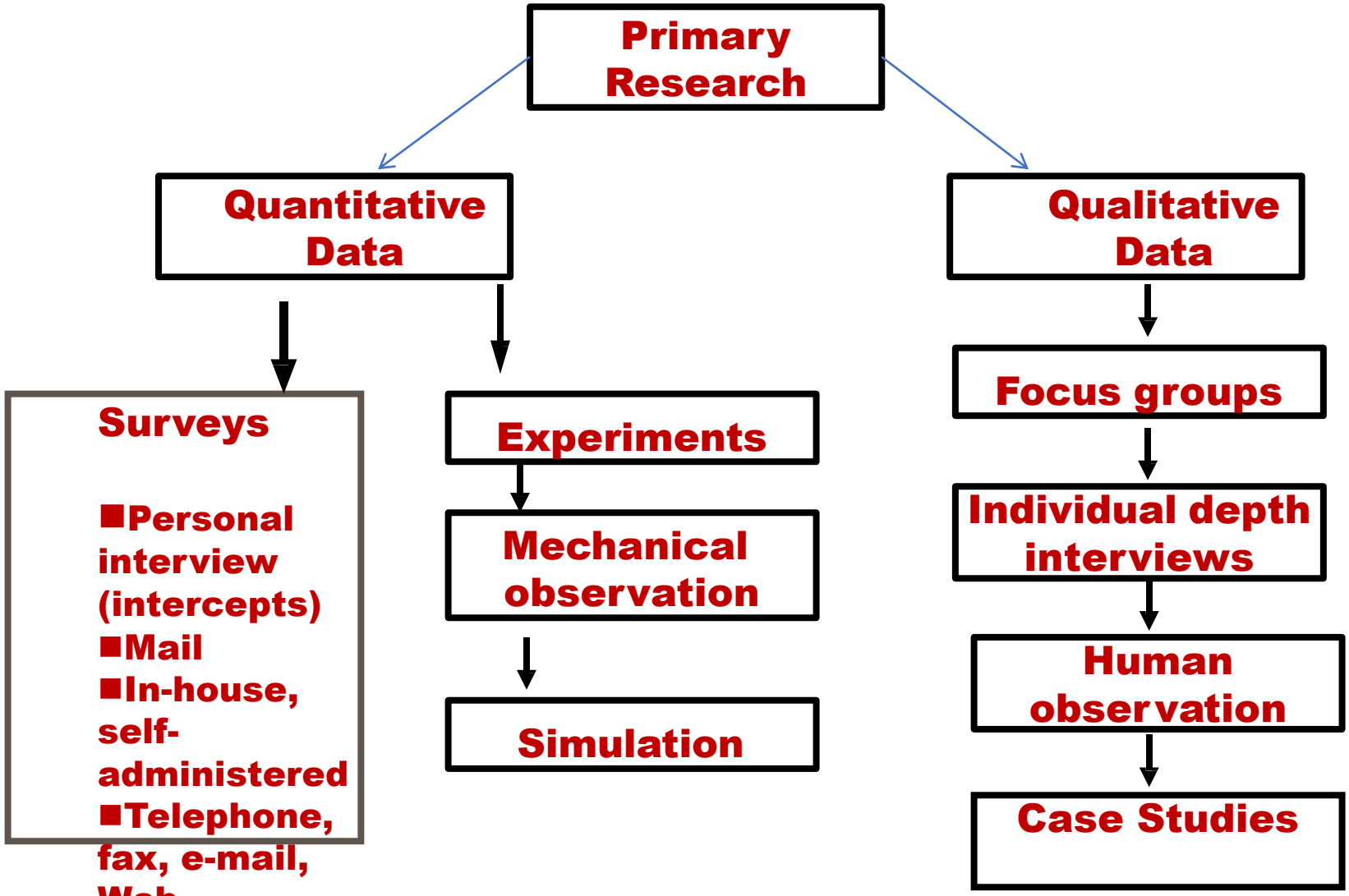
- Primary Data
- Secondary Data

Primary and Secondary Data



PRIMARY DATA

- The data which are collected from the field under the control and supervision of an investigator
- Primary data means original data that has been collected specially for the purpose in mind
- This type of data are generally afresh and collected for the first time
- It is useful for current studies as well as for future studies
- For example**:** your own questionnaire.


Primary Research
Quantitative Data
Qualitative Data
Surveys
■Personal interview (intercepts)
■Mail
■In-house, self-administered
■Telephone, fax, e-mail, Web
Experiments
Mechanical observation
Simulation
Focus groups
Individual depth interviews
Human observation
Case Studies

Primary Research Methods & Techniques

Quantitative and Qualitative Information:

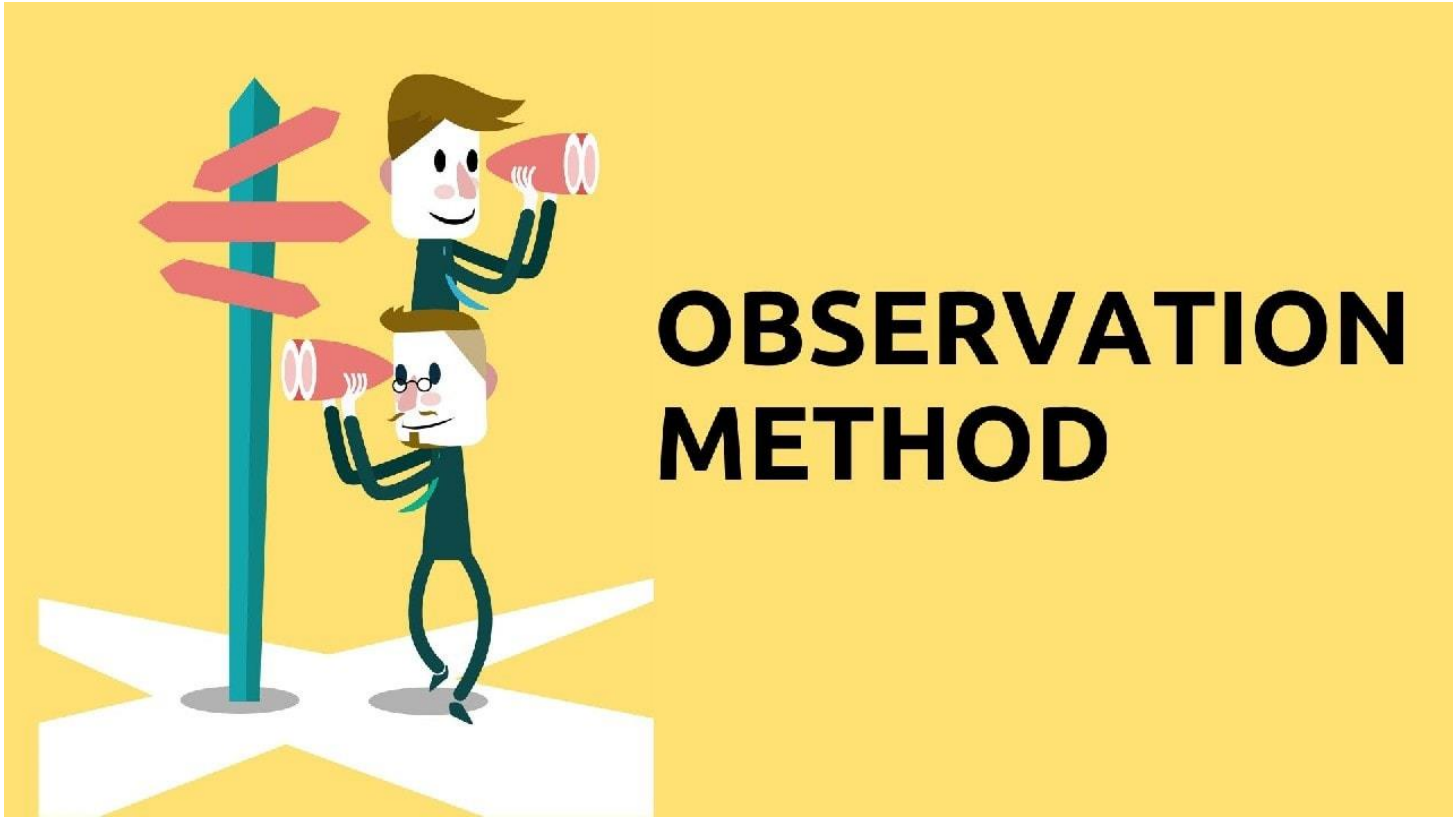
Quantitative – based on numbers – 56% of 18 year olds go for a trip at least four times a year - doesn't tell you why, when, how.

Qualitative – more detail – tells you why, when and how!

Methods of Primary Data Collection

- OBSERVATION METHOD
- INTERVIEW METHOD
- QUESTIONNAIRE
- SCHEDULE

Observation Method



Observation Method

- It is commonly used in studies relating to behavioral science.
- Under this method observation becomes scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose and is systematically planned and subjected to checks and controls.

Observation Method

(a) Structured (descriptive) and unstructured (exploratory) observation-

- When a observation is characterized by careful definition of units to be observed, style of observer, conditions of or observation and selection of pertinent data of observation it is a structured observation.
- When there characteristics are not thought of in advance or not present. it is a unstructured observation.

Observation Method

(b) Participant, Non-participant and disguised observation-

When the observer observes by making himself more or less, the member of the group he is observing, it is participant observation but when the observer observes by detaching himself from the group under observation it is non participant observation.

If the observer observes in such manner that his presence is unknown to the people he is observing it is disguised observation.

Observation Method

(c)Controlled(laboratory) and uncontrolled(exploratory)

- If the observation takes place in the natural setting it is a uncontrolled observation but when observation takes place according to some pre-arranged plans, involving experimental procedure it is a controlled observation.

Observation Method

- **Advantages-**
  - Subjective bias is eliminated.
  - Data is not affected by past behavior or future intentions.
  - Natural behavior of the group can be recorded.
- **Limitations-**
  - Expensive methodology.
  - Information provided is limited.
  - Unforeseen factors may interfere with the observational task

Interview Method

Interview Method

- This method of collecting data involves presentation of oral verbal stimuli and deeply in terms of oral- verbal responses. It can be achieved by two ways:-

(A) **Personal interview-**

- It requires a person known as interviewer to ask questions generally in a face to face contact to the other person. It can be –
- **Direct personal investigation-** The interviewer has to collect the information personally from the services concerned.
- **Indirect oral examination-** The interviewer has to cross examine other persons who are suppose to have a knowledge about the problem.
- **Structured interviews-** Interviews involving the use of pre-determined questions and of highly standard techniques of recording
- **Unstructured interviews-** It does not follow a system of pre-determined questions and is characterized by flexibility of approach to questioning.

Interview Method

- **Focused interview-** It is meant to focus attention on the given experience of the respondent and its effect. The interviewer may ask questions in any manner or sequence with the aim to explore reasons and motives of the respondent.
- **Clinical interviews-** It is concerned with broad underlying feeling and motives or individuals life experience which are used as method to collect information under this method at the interviewer direction.
- **Non directive interview-** The interviewer`s function is to encourage the respondent to talk about the given topic with a bare minimum of direct questioning.

(B) **Telephonic interviews-** It requires the interviewer to collect information by contacting respondents on telephone and asking questions or opinions orally.

Advantages and Disadvantages

- **Advantages-**
- More information and in depth can be obtained.
- Samples can be controlled.
- There is greater flexibility under this method
- Personal information can as well be obtained.
- Mis-interpretation can be avoided by unstructured interview.
- **Limitations**
- It is an expensive method.
- More time consuming.
- Possibility of imaginary info and less frank responses.
- High skilled interviewer is required

Questionnaire

Questionnaire

- In this method a questionnaire is sent (mailed) to the concerned respondents who are expected to read, understand and reply on their own and return the questionnaire. It consists of a number of questions printed or typed in a definite order on a form or set of forms.
- It is advisable to conduct a 'pilot study' which is the rehearsal of the main survey by experts for testing the questionnaire for weaknesses of the questions and techniques used.

Essential of a good questionnaire-

- It should be short and simple.
- Questions should processed in a logical sequence.
- Technical terms expression must be avoided.
- Control questions to check the reliability of the respondent must be present.
- Adequate space for answers must be provided.
- Brief directions with regard to filling up of questionnaire must be provided.
- The physical appearances-quality of paper, color etc must be good to attract the attention of the respondent

Advantages and Limitations

- **Advantages**
  - Free from bias of interviewer.
  - Respondents have adequate time to give answers
  - Respondents are easily and conveniently approachable
  - Large samples can be used to be more reliable.
- **LIMITATIONS**
  - Low rate of return of duly filled questionnaire.
  - Control over questions is lost once it is sent.
  - It is inflexible once it is sent.
  - Possibility of ambiguous omission of replies.
  - Time taking and slow process.

SCHEDULE

Schedule

- The schedule is a proforma which contains a list of questions filled by the research workers or enumerators, specially appointed for the purpose of data collection.
- Enumerators go to the informants with the schedule, and ask them the questions from the set, in the sequence and record the replies in the space provided. There are certain situations, where the schedule is distributed to the respondents, and the enumerators assist them in answering the questions.
- Enumerators play a major role in the collection of data, through schedules.

Schedule

- They explain the aims and objects of the research to the respondents and interpret the questions to them when required.
- This method is little expensive as the selection, appointment and training of the enumerators require a huge amount.
- It is used in case of extensive enquiries conducted by the government agencies, big organizations. Most common example of data collection through schedule is population census.

Enumerator

- An enumerator is a trained person who collects information and performs all the field work related with the collection of data. They help the respondents in filling the questionnaire in case of illiterate population.

Suitability of Enumerators :

- Wide Area: Enumerators are used when the region to be covered is wide and respondents are dispersed.
- No Additional Inquiry: No additional inquiry is required to perform this type of method.
- Control: The administration of enumerators is modest and effectively accessible.
- Training: Enumerators are trained persons in order to collect data. They guide respondents in filling the questionnaire.

Merits of Enumerators:

- Uneducated Persons: As the enumerator, himself tops the questionnaire, he/she can be utilized in those cases where the target population is not proficient.
- Assistance: Respondents can address perplexing and troublesome inquiries with the assistance of the enumerators.
- Less Chance of Bias: It leaves little scope for the poll to be biased.
- Reliability: The data collected is more reliable and correct.
- Responsiveness: There is less chance of non-response as the enumerator visits personally to the people.

Demerits of Enumerators :

- Costly: It is a costly method and requires a lot of money.
- Time Consuming: It is a very time-consuming process.
- Skilled Personnel: The outcome of this technique relies upon the accessibility of prepared and skilled enumerators.
- Personal Bias: The inclination of enumerators could impact the result of data.
- Affordability: It can only be afforded by big organisations.

Advantages & Disadvantages of Primary Data

- **Advantages**
  - Targeted Issues are addressed
  - Data interpretation is better
  - Efficient Spending for Information
  - Decency of Data
  - Proprietary Issues
  - Addresses Specific Research Issues
  - Greater Control

Advantages & Disadvantages of Primary Data

- **Disadvantages**
  - High Cost
  - Time Consuming
  - Inaccurate Feed-backs
  - More number of resources is required

SECONDARY DATA

- Data gathered and recorded by someone else prior to and for a purpose other than the current project
- Secondary data is data that has been collected for another purpose.
- It involves less cost, time and effort
- Secondary data is data that is being reused. Usually in a different context.
- For example: data from a book.

Sources

INTERNAL SOURCES

- Internal sources of secondary data are usually for marketing application-

✓ Sales Records
✓ Marketing Activity
✓ Cost Information
✓ Distributor reports and feedback
✓ Customer feedback

Sources

EXTERNAL SOURCES

- External sources of secondary data are usually for Financial application-

- ✓ Journals
- ✓ Books
- ✓ Magazines
- ✓ Newspaper
- ✓ Libraries
- ✓ The Internet

Advantages & Disadvantages of Secondary Data

- **Advantages**
  - Ease of Access
  - Low Cost to Acquire
  - Clarification of Research Question
  - May Answer Research Question

Advantages & Disadvantages of Secondary Data

- **Disadvantages**
  - Quality of Research
  - Not Specific to Researcher's Needs
  - Incomplete Information
  - Not Timely

Comparison Questionnaire and Schedule

BASIS FOR COMPARISON	QUESTIONNAIRE	SCHEDULE
Meaning	Questionnaire refers to a technique of data collection which consist of a series of written questions along with alternative answers.	Schedule is a formalized set of questions, statements and spaces for answers, provided to the enumerators who ask questions to the respondents and note down the answers.
Filled by	Respondents	Enumerators
Response Rate	Low	High
Coverage	Large	Comparatively small
Cost	Economical	Expensive
Respondent's identity	Not known	Known
Success relies on	Quality of the questionnaire	Honesty and competence of the enumerator.
Usage	Only when the people are literate and cooperative.	Used on both literate and illiterate people.

EXPERIMENTS AND SURVEYS

- Surveys and Experiments are two important statistical techniques used in research and data collection. When the research type is experimental, **experiments** are considered as a major source of primary data. On the other end, **surveys** are performed when the research is descriptive in nature.

Definition of Survey

- By the term survey, we mean a method of securing information relating to the variable under study from all or a specified number of respondents of the universe.
- It may be a sample survey or a census survey. This method relies on the questioning of the informants on a specific subject. Survey follows structured form of data collection, in which a formal questionnaire is prepared, and the questions are asked in a predefined order.

Definition of Experiment

The term experiment means a systematic and logical scientific procedure in which one or more independent variables under test are manipulated, and any change on one or more dependent variable is measured while controlling for the effect of the extraneous variable.

Here extraneous variable is an independent variable which is not associated with the objective of study but may affect the response of test units.

Experiment

- In an experiment, the investigator attempts to observe the outcome of the experiment conducted by him intentionally, to test the hypothesis or to discover something or to demonstrate a known fact.
- An experiment aims at drawing conclusions concerning the factor on the study group and making inferences from sample to larger population of interest

Comparison Between Survey and Experiment

BASIS FOR COMPARISON	SURVEY	EXPERIMENT
Meaning	Survey refers to a technique of gathering information regarding a variable under study, from the respondents of the population.	Experiment implies a scientific procedure wherein the factor under study is isolated to test hypothesis.
Used in	Descriptive Research	Experimental Research
Samples	Large	Relatively small
Suitable for	Social and Behavioral sciences	Physical and natural sciences
Example of	Field research	Laboratory research
Data collection	Observation, interview, questionnaire, case study etc.	Through several readings of experiment.

Correlation

Correlation is the degree of inter-relatedness among two or more variables.

Correlation analysis is a process to find out the degree of relationship between two or more variables by applying various statistical tools and techniques.

According to Conner

"if two or more quantities vary in sympathy, so that movement in one tend to be accompanied by corresponding movements in the other , then they said to be correlated."

Types of Correlation

- **Positive Correlation –**
  - When the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- **Negative Correlation –**
  - When the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- **No Correlation –**
  - When there is no linear dependence or no relation between the two variables.

Correlation Coefficient, r

The correlation coefficient, r, is a measure that describes the extent of the statistical relationship between two ratio level variables.

The correlation coefficient is scaled so that it is always between -1 and +1.

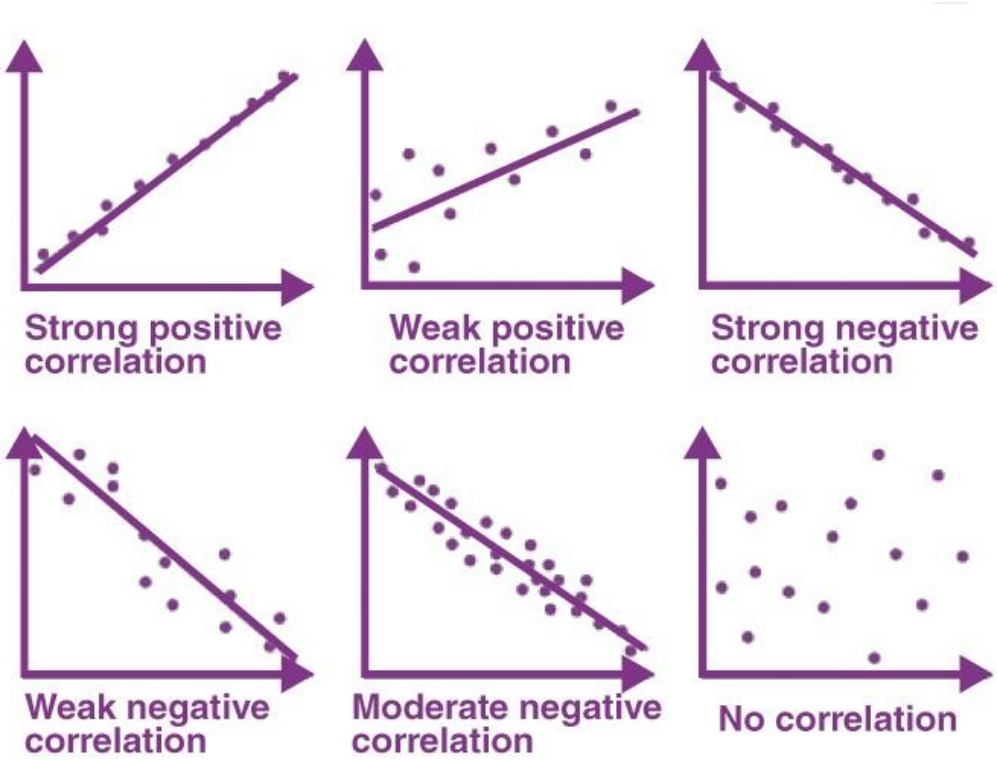
When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

Scatter Diagram

A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other.

The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Later, when the regression model is used, one of the variables is defined as an independent variable, and the other is defined as a dependent variable.



The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables .


Thank you!