Unit 1 : Data Management

1.2 Various Sources of Data

Data can be generated from two types of sources namely

1. Primary data
2. Secondary data

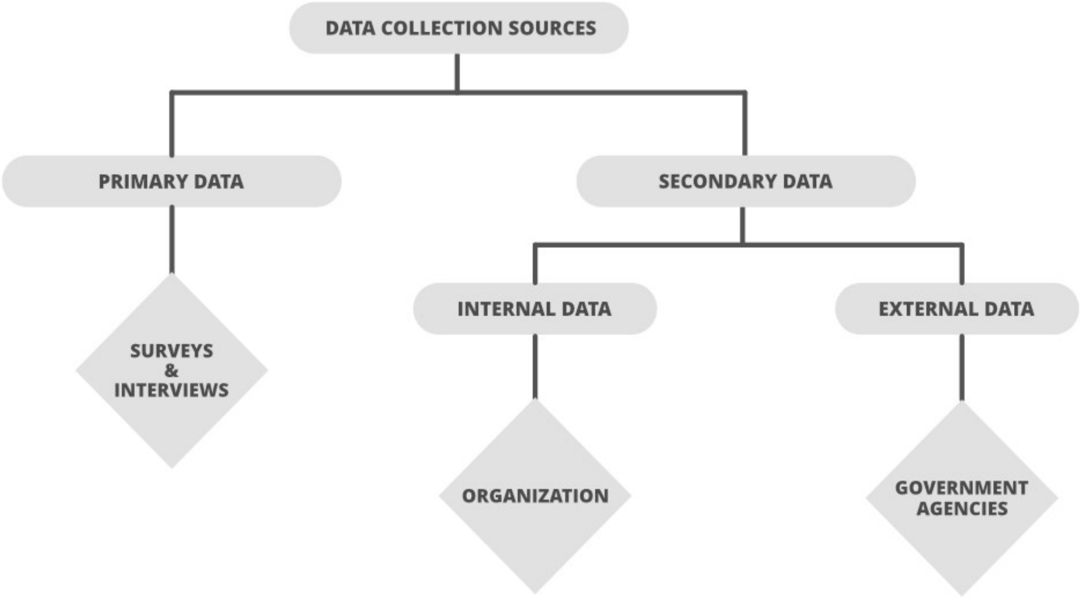


figure 1.3

Understand various primary sources of the Data

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing. Few methods of collecting primary data:

- Interview Method
- Observation Method
- Survey Method
- ExperimentalMethod

Interview method:

The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

Observation method:

An observation is a data collection method, by which you gather knowledge of the researched phenomenon through making observations of the phenomena, as and when it occurs. The main aim is to focus on observations of human behavior, the use of the phenomenon and human interactions related to the phenomenon. We can also make observations on verbal and nonverbal expressions. In making and documenting observations, we need to clearly differentiate our own observations from the observations provided to us by other people. The range of data storage genre found in Archives and Collections, is suitablefor documenting observations e.g. audio, visual, textual and digital including sub-genresof note taking, audio recording and video recording.

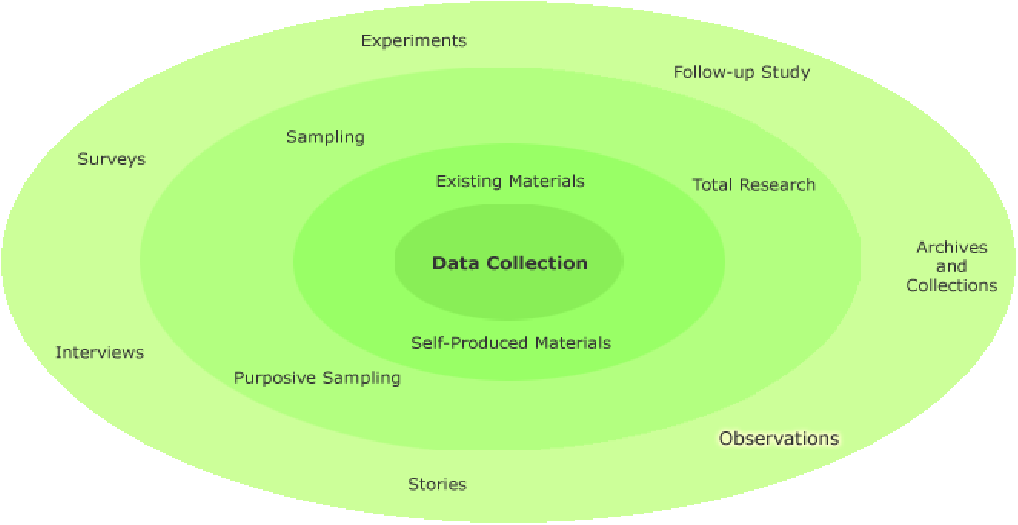


Figure 1.4

There exist various observation practices, and our role as an observer may vary according to the research approach. We make observations from either the outsider or insider point of view in relation to the researched phenomenon and the observation technique can be structured or unstructured. The degree of the outsider or insider points of view can be seen as a movable point in a continuum between the extremes of outsider and insider. If you decideto take the insider point of view, you will be a participant observer *in situ* and actively participate in the observed situation or community. The activity of a Participant observer *in situ* is called field work. This observation technique has traditionally belonged to the data collection methods of ethnology and anthropology. If you decide to take the outsider point of view, you try to try to distance yourself from your own cultural ties and observe the researched community as an outsider observer. These details are seen in figure 1.4.

Survey method:

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

Experimental method:

The experimental method is the process of collecting data through performing experiments, research, and investigation. This method is commonly used in scientific research. The most frequently used experiment methods are CRD, RBD, LSD, FD.

CRD - Completely Randomized Design

A completely randomized design (CRD) is one where the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. For the CRD, any difference among experimental units receiving the same treatment is considered as experimental error. Hence, CRD is appropriate only for experiments with homogeneous experimental units, such as laboratory experiments, where environmental effects are relatively easy to control. For field experiments, where there is generally large variation among experimental plots in such environmental factors as soil, the CRD is rarely used. CRD is mainly used in agricultural field.

Step 1. Determine the total number of experimental plots (*n*) as the product of the number of treatments (*t*) and the number of replications (*r*); that is, *n* = *rt*. For our example, *n* = 5 x 4 = 20. Here, one pot with a single plant in it may be called a plot. In case the number of replications is not the same for all the treatments, the total number of experimental pots is to be obtained as the sum of the replications for each treatment. *i.e.*,

$$n= \sum_{i=1}^{t} {}_{i}r$$

where r_i, is the number of times the *i*th treatment replicated

Step 2. Assign a plot number to each experimental plot in any convenient manner; forexample, consecutively from 1 to *n*.

Step 3. Assign the treatments to the experimental plots randomly using a table of random numbers.

Example 1: Assume that a farmer wishes to perform the experiment to determine which of his 3 fertilizers to use on 2800 tress. Assuming that farmer has a farm divided in to 3 terraces, where those 2800 trees can be divided in the below format

Lower Terrace	1200
Middle Terrace	1000
Upper Terrace	600

Design a CRD for this experiment

Solution Scenario 1

First we divide the 2800 trees in to random assignment of almost 3 equal partsRandom Assignment1: 933 trees
Random Assignment2: 933 treesRandom Assignment3: 934 trees
So for example random assignment1 we can assign fertilizer1, random assignment2 we canassign fertilizer2, random assignment3 we can assign fertilizer3.

Scenario 2

2800 trees is divided into terrace as shown below

Total no of trees	**Terrace**	**Random assignment**	**Fertilizer usage**
2800 Trees	Upper Terrace (600 trees)	200	fertilizer1
		200	fertilizer2
		200	fertilizer3
	Middle Terrace (1200 trees)	400	fertilizer1
		400	fertilizer2
		400	fertilizer3
	Lower Terrace (1000 trees)	333	fertilizer1
		333	fertilizer2
		334	fertilizer3

Thus the farmer will be able analyze and compare various fertilizer performance on different terrace.

RBD - Randomized Block Design

A **randomized block design**, the experimenter divides subjects into subgroups called **blocks**, such that the **variability within blocks is less than the variability between blocks**. Then, subjects within each block are randomly assigned to treatment conditions. Compared to a completely randomized design, this design reduces variability within treatment conditionsand potential confounding, producing a better estimate of treatment effects.

The table below shows a randomized block design for a hypothetical medical experiment.

Gender	**Treatment**	
	Placebo	**Vaccine**
Male	250	250
Female	250	250

Subjects are assigned to blocks, based on gender. Then, within each block, subjects are randomly assigned to treatments (either a placebo or a cold vaccine). For this design, 250 men get the placebo, 250 men get the vaccine, 250 women get the placebo, and 250 women get the vaccine.

It is known that men and women are physiologically different and react differently to medication. This design ensures that each treatment condition has an equal proportion of men and women. As a result, differences between treatment conditions cannot be attributed togender. This randomized block design removes gender as a potential source of variability and as a potential confounding variable.

LSD - Latin Square Design

A Latin square is an experiment design that is similar to CRD and RBD blocks, but contain rows and columns. It is an arrangement of N x N squares with an equal number of rows and columns which contain letters that occurs only once in a row. Hence the difference can be easily found with fewer errors in the experiment. Sudoku puzzle is also an example for LSD design.

For example - 4 X 4 arrangement, In this scheme each letter from A to D occurs only once in each row and also only once in each column. The balance arrangement, it may be noted that, will not get disturbed if any row gets changed with the other.

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

FD - Factorial Designs

This design allows the experimenter to test two or more variables simultaneously. It also measures interaction effects of the variables and analyzes the impacts of each of the variables. In a true experiment, randomization is essential so that the experimenter can infer cause and effect without any bias.

A experiment which involves multiple independent variables is known as a factorial design.

A factor is a major independent variable. In this example we have two factors: time in instruction and setting. A level is a subdivision of a factor. In this example, time in instruction has two levels and setting has two levels.

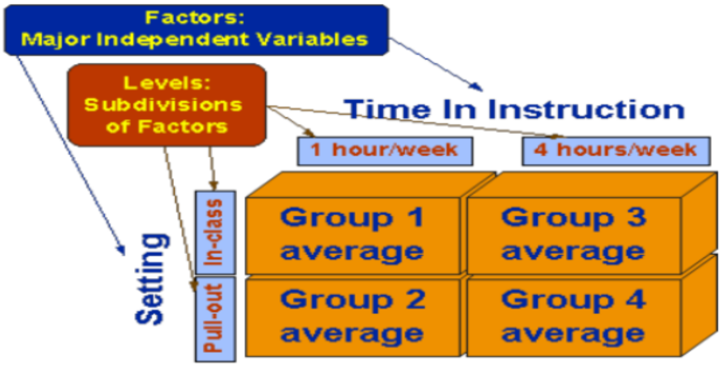
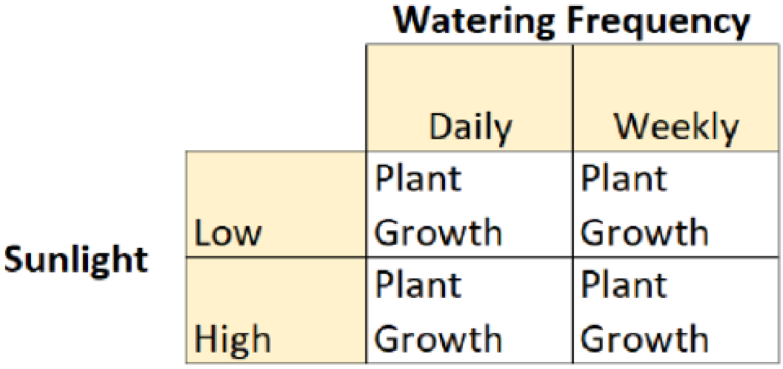


Figure 1.5

For example, suppose a botanist wants to understand the effects of sunlight (low vs. high) and watering frequency (daily vs. weekly) on the growth of a certain species of plant.



		Watering Frequency	
		Daily	Weekly
Sunlight	Low	Plant Growth	Plant Growth
	High	Plant Growth	Plant Growth

This is an example of a 2×2 factorial design because there are two independent variables, each with two levels:

Independent variable #1: Sunlight (Levels: Low, High)
Independent variable #2: Watering Frequency (Levels: Daily, Weekly)

Sources of Secondary Data

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

- Internal Sources - These are within the organization
- External Sources - These are outside the organization

Internal sources

If available, internal secondary data may be obtained with less time, effort and money than the external secondary data. In addition, they may also be more pertinent to the situation at hand since they are from within the organization. The internal sources include

- **Accounting resources-** This gives so much information which can be used by the marketing researcher. They give information about internal factors.
- **Sales Force Report-** It gives information about the sale of a product. The informationprovided is of outside the organization.
- **Internal Experts-** These are people who are heading the various departments. They can give an idea of how a particular thing is working
- **Miscellaneous Reports-** These are what information you are getting from operational reports. If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources.

External Sources of Data

External Sources are sources which are outside the company in a larger environment. Collection of external data is more difficult because the data have much greater variety and the sources are much more numerous.

Government Publications- Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data.

These are:

Registrar General of India- It is an office which generates demographic data. It includes details of gender, age, occupation etc.

Central Statistical Organization- This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

Director General of Commercial Intelligence- This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

Ministry of Commerce and Industries- This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc. It also generates All India Consumer Price Index numbers for industrial workers, urban, non manual employees and cultural labourers.

Planning Commission- It provides the basic statistics of Indian Economy.

Bank of India- This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

Labour Bureau- It provides information on skilled, unskilled, white collared jobs National Sample survey- This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.

Department of Economic Affairs- It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

State Statistical Abstract- This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

Non-Government Publications- These includes publications of various industrial and trade associations, such as The Indian Cotton Mill Association Various chambers of commerce.

1.2.1 Comparison of sources of data

Based on various features (cost, data, process, source time etc.) various sources ofdata can be compared as per table 1.

Comparison Feature	**Primary data**	**Secondary data**
Meaning	Data that is collected by a researcher.	Data that is collected by other people.
Data	Real time data	Past data.
Process	Very involved	Quick and easy
Source	Surveys, interviews, or experiments, questionnaire, etc..	Books, journals, publicationsetc..
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Specific to researcher need	May not be to researcher need
Available	Crude form	Refined form
Accuracy and reliability	More	Less

Table 1: Difference between primary data and secondary data.

1.3 Understanding Sources of Data from Sensor

Sensor data is the output of a device that detects and responds to some type of input from the physical environment. The output may be used to provide information or input to another system or to guide a process. Examples are as follows

- A **photosensor** detects the presence of visible light, infrared transmission (IR) and/orultraviolet (UV) energy.
- **Lidar, a laser-based method of detection**, range finding and mapping, typically uses a low-power, eye-safe pulsing laser working in conjunction with a camera.
- A **charge-coupled device** (CCD) stores and displays the data for an image in such a way that each pixel is converted into an electrical charge, the intensity of which is related to a color in the color spectrum.
- **Smart grid sensors** can provide real-time data about grid conditions, detecting outages, faults and load and triggering alarms.
- **Wireless sensor networks** combine specialized transducers with a communications infrastructure for monitoring and recording conditions at diverse locations. Commonlymonitored parameters include temperature, humidity, pressure, wind direction and speed, illumination intensity, vibration intensity, sound intensity, powerline voltage, chemical concentrations, pollutant levels and vital body functions.

1.4 Understanding Sources of Data from Signal

The simplest form of **signal** is a **direct current (DC)** that is switched on and off; this is the principle by which the early telegraph worked. More complex signals consist of an **alternating-current (AC)** or electromagnetic carrier that contains one or more data streams.

Data must be transformed into electromagnetic signals prior to transmission across a network. **Data and signals can be either analog or digital**. A signal is periodic if it consistsof a continuously repeating pattern.

1.5 Understanding Sources of Data from GPS

The Global Positioning System (GPS) is a **space based navigation system** that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. The system provides critical capabilities to military, civil, and commercial users around the world. The United States government created the system, maintains it, and makes it freely accessible to anyone with a **GPS receiver**.