07/06/20

L.1.1 :- Introduction and Types of Data :-

(Basic Definitions)

* **Statistics :-**

Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description and their analysis, which often leads to the drawing of conclusions.

* Major branches of statistics :-

1). Descriptive Statistics
2) Inferential statistics.

1). **Descriptive statistics :-**

The Part of statistics concerned with the description and summarization of Data is called Descriptive statistics.

- Summarization of Data means numerical / graphical summary of Data or to describe the main points of Data.
- A descriptive study may be performed either on a sample or on a population data.

2). **Inferential Statistics :-**

The Part of Statistics Concerned with the drawing of conclusions from data is called inferential Statistics.

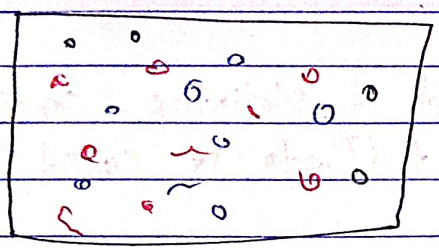
- To be able to draw a conclusion from data, we must take into account the possibility of chance. (Introduction to Probability).

↳ (Info from data) — chance

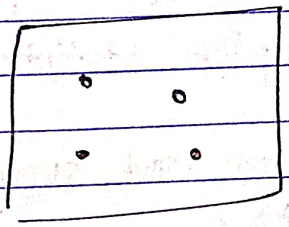
*. Population and Sample :-

- **Population :-**
  The collection of all the elements that we are interested in is called a population.

- **Sample :-** (Subset)
  A subgroup of the population that will be studied in detail is called a sample.



Population.



Sample

*. Purpose of statistical Analysis :-

- If the purpose of the analysis is to examine and explore information about the collected data only, then the study is descriptive. (for its own intrinsic interest only.)

e.g. A class of 10 student gave an exam. and the average marks of class is 51. This type of study is called descriptive statistics.

- If the information is obtained from a sample of a population and the purpose of the study is to use that information to

draw conclusion / inferences about the population. The study is inferential.

Eg: When a teacher wants to know the average marks of all student in the school. then he decides to take a sample of student because there are many student in school. and he calculate average from sample. and from statistical techniques he will know the avarage marks of student. This type of study is called inferential statistics because here we are making conclusion about population based on the sample data.

- A descriptive study may be performed either on a sample or on a population.
- When an inference is made about the population, based on information obtained from the sample, does the study become inferential.

(I ~~don't~~ want my data to say something)

L.1.2 You donot want the Data to say anything, You infact, the data is a fact you use data to extract information what data is telling for interpretation Purposes.

07/06/24

1.1.2 :- Introduction and types of data. — (Understanding data)

*. **Data :-**

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

- Statistics relies on data, information that is around us.

*. **Why do we collect Data :-**

- **Purpose :-** Generally, we collect the data when we are
- Interested in the characteristics of some group or group of people, places, things or events.

e.g. ① To know about temperatures in a particular month in Chennai.
② To know about the marks obtained by student in their class 12th.
③ To know how many people like a new song / Product / Video collected through Comments.

*. **Data Collection :-**

- Data available — Published Data
- Data not available — Need to collect, Generate Data.

We assume data is available and our objective is to do a statistical analysis of available data.

*. **Unstructured and structured Data :-**

→ It required more effort / Natured formate / Diverse datatypes

- **Unstructured data** :- Unstructured data is a dataset that is not organized in a predefined manner.
  Unstructured information is typically text-heavy, but may contain data such as dates, numbers and facts as well. Also, Unstructured data requires more work to process and understand.

- for example :- Youtube Comments, Image files, Social-media posts, Lyrics of a song etc. (Eg. Web Pages/ Video files, Image files)

- When data are scattered with no structure, ie. not in any standard formate, the information is of very little use.

(We need to organised data)

→ Easy to search / Clearly defined and searchable Dataset / Organised/Tabular for

- **Structured Data** :-
  Structured data is a standardized formate for providing information about a dataset and it is clearly define and searchable, as for the information in a dataset to be useful, we must know the context of the numbers and text it holds. Also. Structured data is easy to analyze and understand. Hence, we need to organize the data.

Ex :-

Name	Gender	DOB	Marks
Yash	M	29/06/2011	60
Mayank	M	15/12/2010	59
Sakshi	F	17/05/2009	70
Vikas	M	18/11/2011	44
Seema	F	25/01/2010	75
Rehan	F	01/01/2012	—

*. Variables and Cases :-

- Case (Observation) :- A case / observation is a unit for which data is collected. Cases should uniquely Identify each row in the dataset.

for example :- Previous above chart Yash, Mayank, Sakshi, Vikas, Seema etc are cases as data is collected for every student and all the names uniquely identify each row in the dataset.

- Variables :- A Variable is a characteristic or attribute that varies across all units. Intuitively, a Variable is that "Varies".

for example :- Previous chart, Name, Gender, DoB, Mark e.t.c are variables, as their value keeps on Varying.

Note :- The student dataset in tabular form. If we want to organise a data in a tabular form, then following point should take into Consideration :

*. Row represent Cases :- for each case, same attribute is recorded.

* Columns represent ~~Cases~~ Variables :- for each Variable, same type of Value for each case is recorded. (same unit of data)

* There is difference between in dataset have count 0 and count not available. If count is 0 then its mean figure of that activity is 0 mean that occur but it not available it mean activity is not performed.

* Summary :-

We have Organized data in a spreadsheet into a table.

Each Variable must have its own column.

Each Observation must have its own row.

08/02/24

L.1.3 : Introduction and Types of Data :-
(Classification of data)

* Classification of Data :-
Data is broadly classified into two categories :-

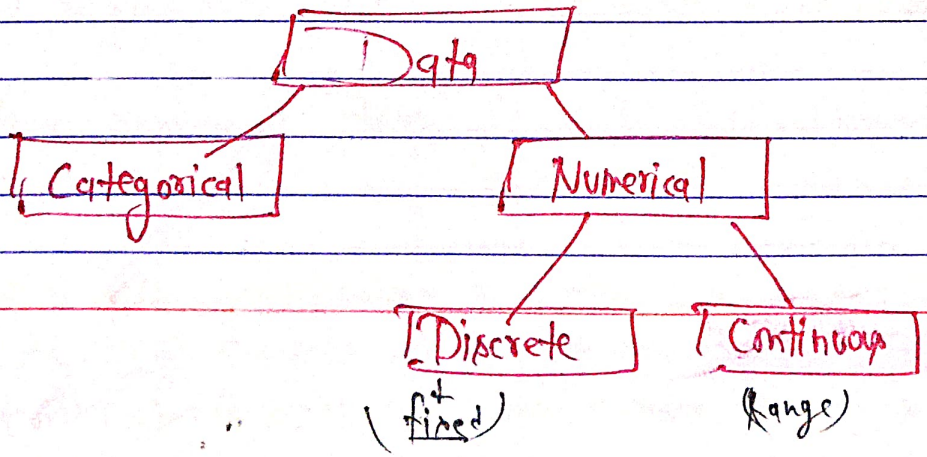
① Categorical Data
② Numerical Data

① Categorical Data :- Categorical Data are also called Qualitative Variables and it identifies the group membership. also, we cannot perform any meaningful mathematical operations on it.

② Numerical Data :- Numerical data are also called Quantitative Variables. It describes the numerical properties of the data. also we can perform mathematical operations on the data.

• Measurement Units :- (For Numerical Data) -
Scale that defines the meaning of numerical data, such as weights measured in kilogram, prices in rupees, heights in centimeter e.t.c.

• Also the data that make up a numerical variable in a data table must share a Common Unit.

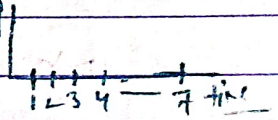


*. Time series and Cross-sectional Data :-

- Time series Data :- If the data is recorded over a period of time, then it called time-series data.

[The graph of a time series showing values in chronological order is known as time-plot.]

e.g. The data collected to observe the temperature in Delhi for seven different days is a time-series data. because data is recorded only for one place (Delhi). and it is recorded over a period of time. (i.e seven different days).



- Cross-sectional Data :- If the data is observed at the same time, then it is called cross-sectional data.

e.g. The data collected to observe the temperature of delhi, kanpur, Bhopal, Chennai on a particular day is a Cross-sectional data. Because, data is recorded at the same time and it is observed for several places.

02/08/24

L.1.4 - Introduction and Types of Data : - (Scales of measurement)

*. Scale of measurement :-

We have four scales of measurment called nominal, Ordinal, interval and ratio scale. Data collection requires any one of the scales of measurement.

*. Nominal scale of measurment :-

When the data for a variable consist of labels or names used to Identify the characteristic of an Observation, the scale of measurment is considered a Nominal Scale.

Ex :- Name, Board, Gender, Blood Group etc.

Note :- • Sometimes nominal variables might be numerically coded like as we might code men as 1 and women as 2 or 0

• There is no ordering in the variable.

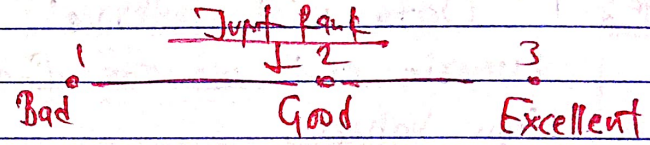
• In short, "Nominal scale is just categories or labels which does not contain any order."

*. Ordinal Scale of measurment :-

When data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurment a Ordinal Scale.

example: Each Customer who visits a restaurant provides a service rating of excellent, good or poor

- The data Obtained are the labels - excellent, good or poor - the data have the properties nominal data.
- In addition, the data can be ranked, or ordered with respect to the service Quality.



Note :-

(1). We can code an ordinal scale of measurment, as bad bad can be coded as 1, good can be coded as 2 and excellent can be coded as 3. There is an order 1, 2, 3 but one thing need to understand is the distance b/w bad and good need not be same as the distance b/w good and excellent. It is just an order.

As we know excellent is better than good, but we cannot say that the difference between good and excellent is same as the difference between good and bad. Thus we have just an order.

(2) In short, Ordinal scale is just categories or labels which can be ordered or contain an order.

*. Interval scale of measurment :-

If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale

Note :- ① Data with interval scale of measurement are always numeric and we can find out the difference between any two values.

② Ratios of values have no meaning here because the value of zero is arbitrary.

- Interval : [+/- only operate]

Numerical values that can be added / subtracted)

* Can't perform Multiplication and division, * (No absolute zero)

Temperature :-

Ex.- Suppose the response to a question, how hot the day is Comfortable and uncomfortable, then the temperature as a variable is nominal.

- Suppose the answer to measuring the temperature of a Liquid is Cold, warm, hot — the variable is ordinal.

- Consider an AC Room where temp is set at 20°C and the temp outside Room is 40°C. It is correct to say that the difference in temp is 20°C, but it is ~~not~~ incorrect to say that the outdoors is twice as hot as indoors. (Ratio have no meaning).

Also, temp in degrees Fahrenheit or degree Centigrade has

* Absolute Zero Property : → Absense of that Particular thing.

→ zero people in Ram - No people ✓

→ 0°C ⟶ No temp ✗ (Interval).

an interval scale of measurement, because it has no absolute zero. In the Celsius scale, 0 and 100 are set to be as the freezing point and the boiling point whereas, in Fahrenheit it is 32 and 212.

	Celsius	Fahrenheit
Freezing Pt	0	32
Boiling Pt	100	212

*. Ratio Scale of measurement :-

If the data have all the properties of interval data and ratio of two values is meaningful, then the scale of measurement is ratio scale.

*. Ratio scale of measurement has absolute zero property which is the key difference between ratio and Interval scale.

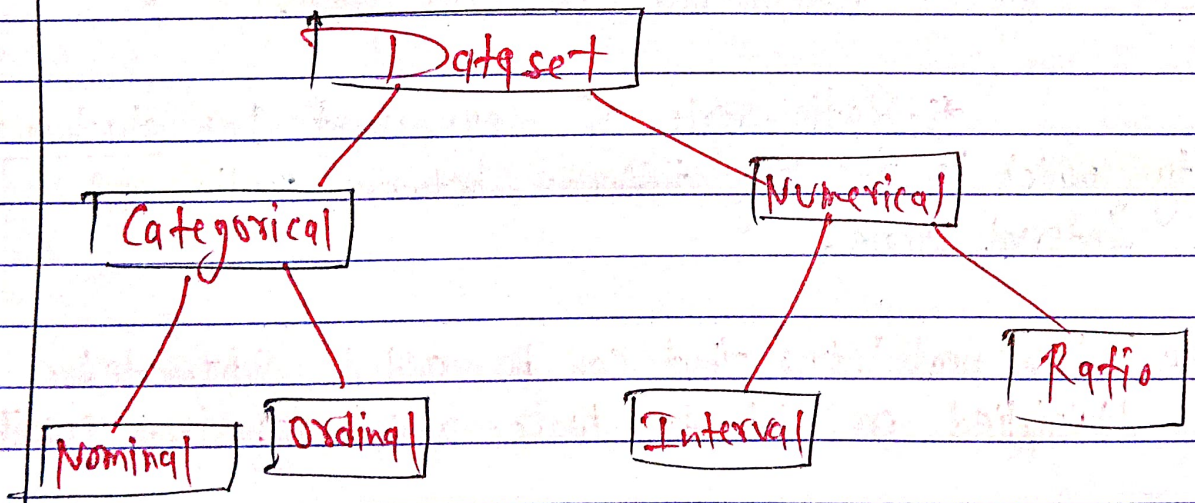
* Ratio : Numerical Values that can be added, Subtracted, Multiplied or divided (makes ratio comparisons possible).

Ex :- Height (cm), Weight (kg), age, marks e.t.c. all such types of data like height, weight and marks can be added, subtracted and multiplied or divided as it all have absolute zero property.

* Summary

- True zero exist → ratio possible → **Ratio Scale** → Age, height, Weight, marks
- No absolute zero, Difference exist → **Interval Scale** → Temperature, GPA etc
- Ratio Scale + Interval Scale → Numerical Data
- Named + Ordered Categories → **Ordinal Scale** → Ranking, rating etc
- Named categories → **Nominal Scale** → Name, blood Group etc.
- Ordinal Scale + Nominal Scale → Categorical Data

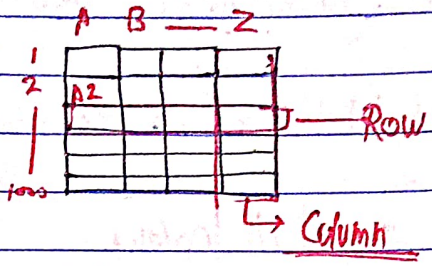
: Flow chart of summary of lecture :-



Week-1 (Tutorial-1) Google spreadsheets Introduction

* Google sheets
* Microsoft Excel

* Open spread on browser.
* each box is called Cell.



* Cell Name Given In this formate, ____ ____
  - ____ (Column Letter)
  - ____ (Row Number)

e.g. A2, E9

* Suppose you give 10000 Rs at interest of 0.5%/month Then Calculate it

	Month	Interest Per Month	Total Interest	Total Amount
①	March	0	0 *	10000
②	April	50	50 ↓	10050
③	May	50	100 ↓	10100
	Drag	Drag	Drag	

* Take any particular Cell and Cursor became pointer. and Click, Hold and Drag. pull it down. Spread sheet will fill after cell with same Value.

* Auto fill also catch pattern 0, 50, 100, Shift + ↓ ** and Now pull it down and drag. and auto fill will fill each cell in same pattern

Tutorial - 2 Formatting Google Spreadsheets

* Labelling the Columns.
- Add Row by go to month and right click and add Row.
  ↳ Rename as you want.
- Highlight the label row (or Heading) with different colour and Bold text.
  - Fill colour
  - Bold text (CTRL + B)
- Over type नहीं — Text wrapping option —
  * More option
  * Text wrapping → Wrap.
- Expand first Row — labelling Row —
  A, B, C, end ↔ cursor (↔) and Drag.
  ↳ edge of column.

* Horizonal align / Vertical Align —

* Number with Value. * Formate → Rupees.
  Select All

Tutorial – 3 (Spreadsheet) formulae

* for making calculation, write = in cell
  Asterisk * is for multiplication.

① 10000 * 0.5/100 = D$2 * 0.5/100 → Because D2 = 10000

* Total interest = interest this month + interest until last month.
  = B3+C2 , B4+C3 , B5+C4

* Total money = D2+C3 → Next D3+C4 , D4+C5
  [Autofill changes the cell numbers in the formula correspondingly.]

D$2+C3 , for D2 should be same.
C Column Row No should be change.
↓ $ symbol should be used
Add dollar symbol

* Interest Rate Per month | 0.5%.
  Ctrl+C → Ctrl+V

And Now fx = D$2 * G$1

* Tutorial – 4 (Downloading / Uploading S.Sheets
  File – Download – Choose formats.

* Openspeppion (Statistics 1)

*. Sample shald be a Good representation of population.

Type of Numerical Data :→

* Diresete Data :- Count and it taken the distinct Value in real line. E.g. No of people in a row.

* Continuous Data :- Not Countable ✗
Measurement ↵ and it contain take any Value b/w an Interval.
e.g. Weight of a person / Speed of a Vehical / Amount of milk.