Unit-1

Data and its Descriptive Analysis

* Qualitative Data:

- data are classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, region or caste which cannot be quantified. This is done in two ways:

- Simple classification: each class is subdivided into two sub-classes and only one attribute is studied.

- Manifold classification: a class is subdivided into more than two sub-classes which may be sub-divided further.

* ~~Qualitative data:~~
* Quantitative Data:

- data are classified on the basis of some characteristics which can be measured such as height, weight or sales.

- Variable: any quantity or attribute whose value varies from one investigation to another.

- Quantitative variables can be divided into the following two types:

- Continuous variable: the one that can take any value within the range of numbers. In which case data are obtained by measurements eg: height or weight of individuals.

- Discrete/discontinuous variable: the one where values change by steps or jumps and it can not assume a fractional value. In such a case data are obtained by counting. eg: number of children in a family.

* Variables: — Quantitative
— Categorical/Qualitative

- The characteristics which one intends to investigate and analyse while conducting a survey.

- They may vary from response to response

- Categorical variables are those which are expressed in numerical terms.

- Discrete variables are which can only take certain fixed integer numerical values.

- Continuous variables are which can take any numerical value.

* Scale of Measurement :

- **Nominal :-** (numerical values are either named or categorized in such a way that these values are mutually exclusive and collectively exhaustive.
  eg: a player wearing a shirt number 24 is not more of anything than a player wearing a shirt number 12.)

  - (It is said to be least powerful among four scales because this scale suggest no order or distance relationship) and have no arithmetic origin (It is usually used for obtaining personal data, where grouping of individuals or objects is useful.)

- **Ordinal :-** (numerical values are categorised to denote qualitative differences among the various categories as well as rank ordered the categories) in some meaningful way according to some preference

  - not only indicates the differences in the given items but also gives some information as to how respondents distinguish among these items by ranks ordering them. (However does not give any indication of the magnitude of the differences among the ranks.) The interval between values is not interpretable in an ordinal measure.

- (usually used to rate the preference or usage of various brands of a product by individuals) and to rank individuals, objects or events.

- Interval :- (allows us to perform certain arithmetical operations on the data collected from the respondents)

- nominal scale only allows to quantitatively distinguish groups by categorizing them into mutually exclusive and collectively exhaustive sets, the ordinal scale allows us to rank order the preferences and (the interval scale allows us to compute mean and standard deviation of the responses on the variables) i.e. not only classify according to certain categories and determines order of these categories, but also measure the magnitude of the differences in the preferences

- (the distance between attributes does have meaning)

- used when responses to various questions that measure a variable can be determined on a five-point or any other number of points scale

- Ratio: - (has an absolute measurement point, thus not only measures the magnitude of the differences between points) on the scale but (also provides the proportions in the differences)

  - (most powerful of the four scales because it has a unique zero origin)

  - The measure of central tendency of the ratio scale could either be arithematic or geometric mean and measure of dispersion could either be standard deviation or variance, or coefficient of variation.

  - (usually used in organizational research when exact figures on objective factors are desired.)

* Measures of Central Tendency

- Mean: - Arithmatic Mean
  - → Weighted
  - → Simple
  - Geometric Mean
  - Harmonic Mean

- Median

- Mode

- Notations :

m_i = mid point for the i^{th} class in the data set
f_i = number of observations in the i^{th} class
N = total number of observations
n = number of observations in the sample
L = lower limit of any class-interval
h = width of the class-interval
cf = cumulative frequency
Σ = summation of all values of observations

- Arithmetic Mean of Ungrouped Data :

- Direct Method : $\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N}$ (individual)

$$\bar{x} = \sum_{i=1}^{n} f_i x_i \cdot \frac{1}{n}, n=N \text{ (discrete)}$$

- Short-cut Method : $d_i = x_i - A$; A = assumed mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = A + \cdot \frac{1}{n} \sum_{i=1}^{n} d_i \text{ (individual)}$$

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^{n} f_i d_i \text{ (discrete)}$$

- Arithmetic Mean of Grouped Data

- Assumptions for grouped data :

- class intervals must be closed
- width of each class interval should be equal

* In individual case n = no. of observations whereas in discrete and continuous, $n = \Sigma f$

- Direct Method: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} f_i m_i$

- Step deviation Method: $\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i} \times h$

h = width of the class intervals

$d_i = \frac{m_i - A}{h}$, A is from m_i

- Step deviation is useful in those cases where mid-values and frequencies are in three or more digits.

* Special Cases:

- More than type or Less than type:

More than type are calculated by adding frequencies from bottom to top, so that the first class interval has the highest cumulative frequency and it goes on decreasing whereas in less than type it goes on increasing. Data ~~of~~ are first converted into inclusive class intervals or exclusive class intervals, then the calculations for $\bar{x}$ are done.

- Frequency distribution have open end intervals:

Since class intervals are uniform, therefore we can take same width for open-end class interval also.

- Weighted Mean: $\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$

* **Median:**

- called a middle value in an ordered sequence of data in the sense that half of the observations are smaller and half are larger than this value

- **Ungrouped Data:** arrange data in either ascending or descending order

  - when observations (n) is an odd number

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

  - when observations are even number

$$\text{Median} = \frac{\frac{n}{2}^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$$

- **Grouped data:** $\text{Median} = L + \frac{(n/2) - cf}{f} \times h$

L = lower limit of median class interval
cf = cumulative frequency of class prior to median class
f = frequency of median class
h = width of median class

- first identify the class interval which contains $(n/2)^{th}$ observation. To identify find cf of each class until the class for which value of cf is greater than or equal to $(n/2)^{th}$ observation

* **Mode**: that value of an observation which occurs most frequently in the data set.

- poor measure of central tendency when most frequently occuring values of an observation do not appear close to the centre of the data.

- convert ungrouped data in grouped data.

$$\text{Mode} = L + \frac{f_1 + f_0}{2f_1 - f_0 - f_2} \times h$$

f_1 = frequency of modal class
f_2 = frequency of class following the modal class
f_0 = frequency of class preceding the modal class

* **Relationship between Mean, Median, Mode:**

$$2\,\text{Mean} = 3\,\text{Median} - \text{Mode}$$

- **Introduction to Big Data:**

Big data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it efficiently.
e.g: NYSE generates 1 TB of data everyday.

Characteristics: (4)

- Volume: Size plays a crucial role in determining value out of data. Also, whether data is big or not. Resources required to process it. It needs to be considered while dealing with Big Data (5)

- Variety: It refers to heterogenous sources and nature of data. During earlier days, spreadsheets and databases were only sources. The variety of unstructured data poses certain issues for storage, mining and analyzing data.

- Velocity: It refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determine the real potential. Flows from sources like business processes, application logs. Flow of data is massive and continuous.

- Variability: This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage data effectively.

5 Stages

(1) Data Integration: Making data available to process from source
(2) Data Storage: Storing data using technology like data environment
(3) Data Analysis: Analyse data using DL and ML

to get meaningful interences.

(4) Visualization: data visualization techniques are used to analyse data or management decisions are based on them.

(5) Data Security: Protect data from leaking.