UNIT - II

GETTING INSIGHTS FROM DATA, DATA QUALITY AND PREPROCESSING

Descriptive statistics

→ It describe, show and summarize the basic features of a dataset found in a given study, Presented in a summary that describes the data sample and its measurement. It helps analyst to understand the data better.

→ It represent the available data sample and do not include theories, inferences, probabilities or conclusions.

Eg.,

→ stud Grade Point Avg (GPA).

→ Collect data points created through large selection of grades, classes, and exam, Avg them and present general idea of the stud's mean academic performance.

→ It doesn't Predict future or Conclusions.

→ It provides a straightforward summary of stud success based on collected data.

→ Consider, set of 2, 3, 4, 5 and 6 = 20

data set mean = '4' (20/5 = 4).

→ Analyst use chart and graphs to present descriptive statistics.

Eg:1

→ Movie - Theater

→ Ask '50' members collect the data from audience that whether they liked or not (Yes).

→ After collecting represent it into pie ...

Eg:2

Political Polling.

Types of Descriptive statistics:

Descriptive statistics break down into several types

characteristics or measures

→ Four scale types

Nominal

ordinal

interval

ratio

→ Frequency Distribution

→ Datasets Consists of a distribution of scores or values

→ statisticians use graphs or tables

to summarize.

Eg: choice of musician

Central Tendency of Measures

→ 3 methods (findings representation)

Mean - 'M/N' → Avg.

mode

median

Mean Eg:

how many hours will you sleep in a week.

6, 8, 7, 10, 8, 4, 9 = 52 (M)

7 (N) = 7.3

Mode:

→ Frequent Response value

→ Arrange the values in ascending order

4, 6, 7, 8, 8, 9, 10

→ Mode is '8'

Median:

→ Arrange the values in ascending order.

→ 4, 6, 7, 8, 8, 9, 10

→ Median is Eight

Variability

→ spreading info.

→ 3 aspects:

Range

Std deviation

Variance.

Range:

→ determine how far apart the most extreme values.

→ subtract lowest value from highest value from data set.

eg: 4, 6, 7, 8, 8, 9, 10

low: 4, high = 10 10 − 4 = 6

→ '6' is the range.

std deviation:

→ Avg amount of variability

→ 6 steps:

1. List the scores and their mean
2. Find the deviation by subtracting the mean from each score
3. Square each deviation
4. Total up all the squared deviation

5. Divide the sum of the squared deviation by N-1

6. find the results square root.

Raw data	Deviation from mean	Deviation squared.
4	4 - 7.3 = -3.3	10.89
6	6 - 7.3 = -1.3	1.69
7	7 - 7.3 = -0.3	0.09
8	8 - 7.3 = 0.7	0.49
8	8 - 7.3 = 0.7	0.49
9	9 - 7.3 = 1.7	2.89
10	10 - 7.3 = 2.7	7.29
M = 7.3	sum = 0.9	④ Square sum = 23.83

⑤ squared deviations = 6(N-1)

= 23.83/6 = 3.971

square root = 1.992.

Variance.

→ reflects the datasets degree spread.

→ greater the degree of data spread, the larger the variance relative ito the mean.

→ get the variance by just squarring the std deviation

Square $\sqrt{1.992} = 3.971$.

Descriptive Univariate statistics.

→ helpful when it comes to summarizing huge amounts of numerical data as well as revealing patterns in the raw data.

→ Patterns discovered in univariate data may be described using central tendency (mean, median & mode). as well as dispersion: variance, Range, quartiles, std deviations; Maximum and minimum.

→ When dealing with univariate data, we can numerous alternatives for defining it.

- Frequency distribution Table
- Histograms
- Bar charts
- Pie "
- Frequency polygon

Multivariate Analysis:

→ Many statistical techniques focus on just one or two variables

→ Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once.

→ Multivariate regression is not belong to this, but can be thought of as a MVA.

→ outline:

Why MVA is useful and important

Simpson's Paradox

Some Commonly used techniques

Principal Components

cluster Analysis

Correspondance Analysis

Market segmentation methods

MVA methods