JVP

Class – X

Subject – Data Science

Chapter -1 Use of Statistics in Data Science

What are subsets:

A **subset** is a **portion of data** taken from a **larger dataset**.
It contains **some (or all)** rows or columns from the original dataset, but **never anything outside** it.**Example**

If you have a Table of 100 rows and 100 columns and you want to perform certain actions on the first 5 rows and the first 5 columns, you can separate it from the main table. This small table of 5 rows and 5 columns is known as a "Subset" in Data Analytics.

How do we subset the data?

1. Row-based subsetting:
   Selecting specific **rows** (records) from a dataset based on some condition or criteria.
   Example:
   If you have a dataset of students and you only want students who scored 90 marks.
2. Column based subsetting:
   Selecting specific columns (features/variables) from a dataset.
   Example:
   If your dataset has many columns (name, age, marks, city) but you only need name and marks.
3. Data based subsetting:
   Selecting a **subset of data based on both rows and columns** — i.e., applying conditions to rows **and** selecting specific columns together.
   Example:
   Students who scored above 80 **and** only showing their names and marks:

Two-way frequency table:

A **two-way frequency table** shows how two categorical variables are related by counting how many times each pair of values occurs together.

Consider you are conducting a poll asking people if they like chocolates. You record the data in the below format.

Person name	Age	Like chocolates?
Person 1	6	Yes
Person 2	8	Yes
Person 3	13	Yes
Person 4	12	Yes
Person 5	18	No
Person 6	9	No
Person 7	16	Yes
Person 8	19	No
Person 9	14	No
Person 10	12	Yes

Age group	Like chocolates	Do not like chocolates
5 - 10	2	1
10 - 15	3	1
15 - 20	1	2

Two-way frequency tables show how many data points fit in each category. The row category in this example is “5-10 years”, “10-15 years” and “15-20 years”. The column category is their choice “Like chocolates” or “Do not like chocolates”. Each cell tells us the number (or frequency) of the people.

Two-way relative frequency table:

Two-way relative frequency table very similar to the two-way frequency type of table. Only difference here is we consider percentage instead of numbers.

Preference	Girls	Boys
Indoor games	70%	20%
Outdoor games	30%	80%
Total	100%	100%

Central Tendency:

Central tendency refers to a single value that attempts to describe a set of data by identifying the **central** value within that data set. The three most common measures of central tendency are: Mean, Mode, Median

Mean: Mean, also termed as the simple average, is an average value of a data set. Mean is a value in the data set around which the entire data is spread out.

Median The median like the mean is another form of central tendency. It is the middle point of a sorted data set. To calculate the median, we must order our data set in ascending or descending order, and then select middle value of the set.

Mean vs median So mean and median both represent the central tendency but Median is a more accurate form of central tendency especially in scenarios where there are some irregular values also known as outliers.

Mode: The mode is the number that appear most in a data set. A set of number may have one or more then one mode, or no mode at all.

Mean Absolute Deviation:

Mean Absolute Deviation (MAD) is the average of how far away all values in a data set are from the mean.

Let understand this with an example. Consider the below data set.
Example: Data = 12, 16, 10, 18, 11, 19
Step 1: Calculate the mean
Mean = (12 + 16 + 10 + 18 + 11 + 19) / 6 = 14 (rounded off)
Step 2: Calculate the distance of each data point from the mean. We need to find the absolute value. For example if the distance is -2, then we ignore the negative sign.

Value	Distance from mean (14)
12	2
16	2
10	4
18	4
11	3
19	5
Total	**20**

Fig 1.15 *Distance from mean*

Step 3: Calculate the mean of the distances.
Mean of distances = (2 + 2 + 4 + 4 + 3 + 5) / 6 = 3.33
So 3.33 is our mean absolute deviation, and the mean is 14

What is Standard Deviation?

The Standard Deviation is the measure of how spread out the numbers are. To be specific, standard deviation represents how much the data is spread out around the mean or an average.

1. Calculate the mean by adding up all the data pieces and dividing it by the number of pieces of the data.
2. Subtract mean from every value
3. Square each of the differences
4. Find the average of squared numbers calculated in point number 3 to find the variance.
5. Lastly, find the square root of variance. That is the standard deviation

For example, Take the values 1,2,3,5 and 8 to calculate Standard Deviation

Step 1: Calculate the mean 1+2+3+5+8 = 19

19/5 = 3.8 (mean)

Step 2: Subtract mean from every value

1- 3.8= -2.8
2- 3.8= -1.8
3- 3.8= -0.8
5- 3.8= 1.2
8- 3.8= 4.2

Step 3: Square each difference

-2.8*-2.8 = 7.84
-1.8*-1.8 = 3.24
-0.8*-0.8 = 0.64
1.2*1.2 = 1.44
4.2*4.2 = 17.64

Step 4: Sum of squared deviations
7.84 + 3.24 + 0.64 + 1.44 + 17.64 = 30.8
Step 5: Divide by total numbers (for Population SD)
30.8/5= 6.16