Statistics

(Session-3)

Data Measurements

Data measurement are divided into two types :

1. Central Tendency

- Mean
- Median
- Mode

2. Data Dispersion

- Range
- Mean deviation
- Absolute mean deviation
- Variance
- Standard deviation

MEAN:

- The average value , calculated by summing all values and dividing by the number of values.

- Sum of all observations divided by total number of observations.

- Generally , we can say the average of particular crop per a season as 10 bags

- or the virat kohli average in ODI is 50.

Mean = Sum of all observations / Total no.of observations

Let us take some observations , such as marks of a student

Marks :

Telugu = 91

Hindi = 81

English = 94

Maths = 89

Science = 90

Social = 92

Mean = Sum of all observations / Total no.of observations

$$\frac{91+81+94+89+90+94}{6} = \frac{537}{6} = 89.5$$

Here , we can say that

The student can get 90 marks in every subject as average

Average = $\frac{x_1+x_2+x_3+x_4+x_5+x_6}{6} = \frac{\sum_{i=1}^{n} x_6}{6}$

If there are N observations

Average = $\frac{x_1+x_2+x_3+x_4+\cdots+x_N}{N}$

$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$

Where , μ = Population mean

X = Sample mean

- Population mean is the sum of all the population values divided by the total number of population values.

- Sample mean is the sum of all the sample values divided by the total number of sample values

* For multiplication of observations :

$$x_1 * x_2 * x_3 * x_4 * x_5 * x_6 = \prod_{i=1}^{6} x_6$$

- For N observations

$$x_1 * x_2 * x_3 * \dots * x_n = \prod_{i=1}^{n} x_n$$

MEDIAN :

- Median is the middle value of the dataset , which is in order.

- The data should be in Ascending order or decending order , then find median.

- If there are even numbers of values , average of two middle values , we get median.

- Median is the 50 percentile of the data

Let us take some dataset

1,3,2,6,4,7,5

- Keep the dataset in order
- Ascending order of the dataset ==> 1,2,3,4,5,6,7
- Median = 4

Let us take even dataset

1,3,2,6,4,7,5,8

- Keep the dataset in order
- Ascending order of dataset ===> 1,2,3,4,5,6,7,8
- Two middle values ===> 4,5

Average = $\frac{4+5}{2} = 4.5$

Median = 4.5

MEAN Vs MEDIAN :

Imagine that USA friend asked Indian friend , What is the average Indian Income.

F1 : 1L , F2 : 2L , F3 : 3L , F4 : 4L , F5 : 5L

The above are the salaries of Indians

Average$= \frac{1+2+3+4+5}{5} = 3L$

Mean = 3L

Median of 1L,2L,3L,4L,5L =3L

If we add 200crs

1L , 2L , 3L , 4L , 5L , 200crs

Average$= \frac{1+2+3+4+5+200crs}{5} = 20crs$

Mean = 20crs

Median of 1L , 2L , 3L , 4L , 5L , 200crs = $\frac{3+4}{2}$ = 3.5

** If a data has very very huge values or very very less values , mean will be effect

Median doesnot effect

Outlier :

The above case of unusual observations is caleed as outlier.

- When we have huge observation either positive or negative.
  - mean will effected
  - median will not effected

MODE :

- Most repeated value from the raw data is called mode.
- Most frequently occurred value
- Raw data : 1,5,6,7,1,6,1,5,1,8,1,3,1.
- keep raw data in order

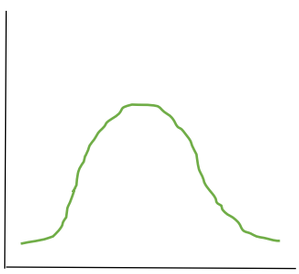
  1,1,1,1,1,1,3,5,5,6,6,7,8
- Most repeated value from raw data is '1'

Mode = 1

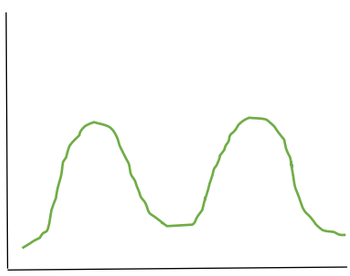
Data Distribution :

- In Distribution we can have many modes.

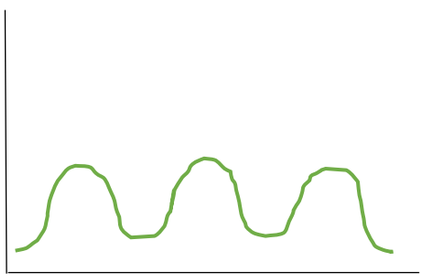
- If we have single mode , it is called as Unimode.
- If we have two modes , it is called as bimodes.
- If we have three modes , it is called as multimode.



UNIMODE



BIMODE

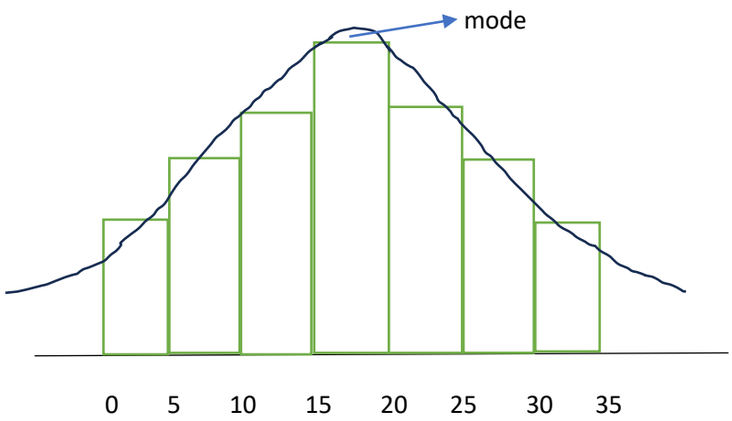


MULTIMODE

Raw data : 1,5,6,7,1,6,1,5,1,8,1,3,1

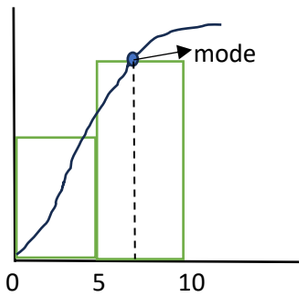
C - I	C I F
0 – 3	6
3 – 6	4
6 – 9	2

For Example :



- Data – 1,2,3,4,5,6,6,6,7,8,8,10

C – I	C I F
0 – 5	5
5 – 10	7



- Highest peak of data distribution is called as mode.

- Mode is available at that point.

- we know that distribution forms from histogram.

- Histogram forms from interval.

- If we are seeing highest peak value in the distribution means , that corresponding interval haas mode value.

- By seeing higest peak we cannot say Exact value of mode.

- we can say the mode available in the particular interval.

Mean – Median – Mode :

Mean :

- Mean will give average value of the data.
- Mean will affect by outlier.
- Mean gets pull by outlier , towards the Outlier.

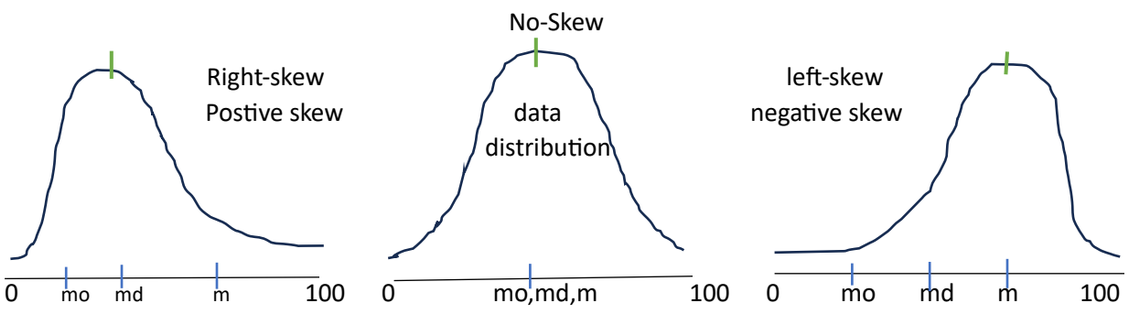
Median :

- Median will give middle value.
- Always 50 percentile of the data , Exactly half way.

Mode :

- Mode will give highest peak in the distribution

Types of Data Distribution :



Mo < Md < M

Mo = Md = M

Mo > Md > M

Skew ===> Pulling

Right Skewed or Negative Skewed :

- Because of positive Outliers.
- Mode < Median < Mean
- Assume that data ranges from 0 to 100
- Positive side or right side data is pulling which mean 100 side.
- So that mean value is high.

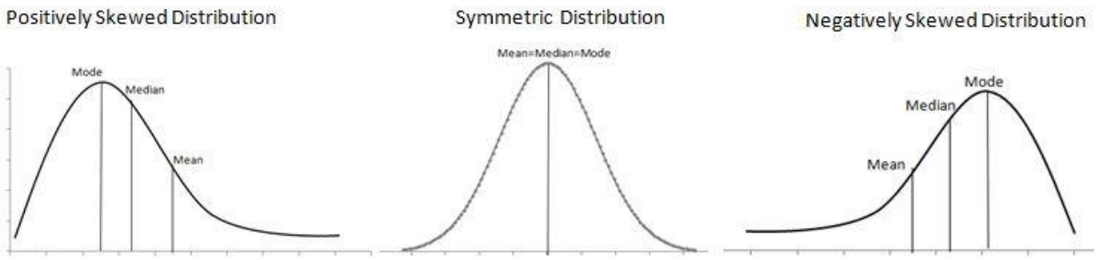
Left Skewed or Negative Skewed :

- Because of negative Outliers.
- Mode > Median > Mean
- Assume that data ranges from 0 to 100
- Negitive side or left side data is pulling which mean 0 side.
- So that mean value is low.

No Skew or Normal Distribution :

- No Outliers
- Mode = Median = Mean
- Bell shaped Curve
- 50% data is left side and 50% in right side

-



- Skew means pulling.
- The reason for the skew is Outliers
- If The outlier is :
    - Right side means – Maximum value , Based on the coordinate.
    - Left side means – Minimum value