Statistics -1

Statistics- Mathematical branch that uses data for finding useful information for making decisions.

- **Descriptive Statistics -** Summarization of Data
- **Inferential Statistics** - Uses a hypothesis to conclude the result.

Variables -

- **Categorical** - Can be placed into categories - e.g - Marital Status(Yes or No), Seasons(Summer, Winter, Rainy)
- **Numerical** - That represents quantities - 1) Discrete - Finite values. e.g - No. of children, no. of apples in a shop. 2) Continuous - infinite in range. e.g - the weight of apples.

Descriptive Statistics -

MODULE-1 Measure of Central Tendencies-

- **Mean - Average of data.**

1. Population Mean (μ) = $\Sigma X / N$, N number of the entire population
2. Sample Mean($\bar{x}$) = $\Sigma X / n$, n number of data in the sample taken
3. Weighted Mean - Category wise - average.

   $\bar{x} = \Sigma w_i x_i / \Sigma w_i$, where w_i is the weight. e.g - CGPA calculation -

Subject	Credits (Weight)	Points scored by the student
Maths	10	9
English	8	6
Science	9	9

The Cgpa will be calculated - (10*9)+ (8*6) +(9*9) / (10+8+9) = 8.11 Weighted mean = simple mean, when all the weights are equal.

4. Trimmed Mean - Removing the extreme values of the data before calculation of the mean. e.g - In a class of 40 students, the marks of the students are 100,65,622,63,68,59,....bet 70-20. So, here 100 marks is the extreme value, so while calculating trimmed mean, it will be excluded as it can give a wrong impression about the average marks of students.

5.Geometric Mean

- **x̄g**= n√(x1·x2···xn)

e.g - For calculating growth rate of a company.

- **Median - Middle value of the data.**

1. If n(number of data values) is odd - median = (n+1/2)th value.
2. if n is even - median = mean(n/2 th value & n/2 +1 th value).
3. e.g - 4 5 6 - n=3, median value is 3+1/2 = 2 nd value =" 5 "
4. 4,5,6,7 --> average of 5 and 6 = "5.5 "

- **Mode - Most frequent values.**

Module-2 - Dispersion Measures

- **Variance - Measure of variability of data.**

σ^2 = Σ(xi - μ)^2 / N

- **Standard Deviation**

The "**Population** Standard Deviation":

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The "**Sample** Standard Deviation**":**

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

e.g - Scores of Dhoni and Kohli in the entire series -

Dhoni	Kohli
60	**23**
10	**23**
20	**24**
5	**25**
Total -95	Total-95
σ = 21.61	σ=0.829

Thus, the standard deviation of Kohli's Score is much less than Dhoni. This implies that Kohli is more consistent.

Standard Deviation **inversely proportional** consistency of data.

- **Percentile -** Relative standing/measure.

The pth percentile is a value so that **roughly p% of the data are smaller and (100-p)% of the data are larger.** To find pth percentile:

1. Arrange the data in ascending order.
2. Calculate np/100, where n is the number of data values, and

a) if the result is an integer, say i, take the average of the ith ordered data value and the next value

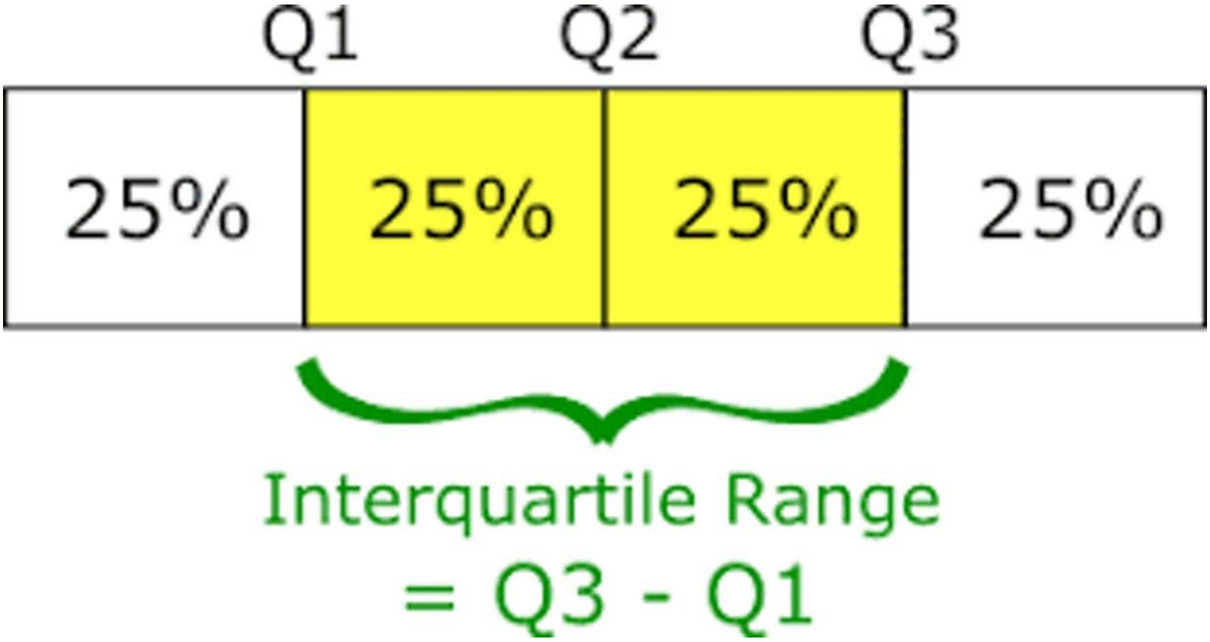
b) if not an integer, round the number up to the next integer, say j, take the jth ordered data value.

for eg **-** In an exam, if you score 90 percentile, that means you scored better than **90**% of people who took the **test**.

- **Quartile -** are values that divide your data into quarters.

1. First quartile: the lowest 25% of numbers
2. Second quartile: between 25.1% and 50% (up to the median)
3. Third quartile: 51% to 75% (above the median)

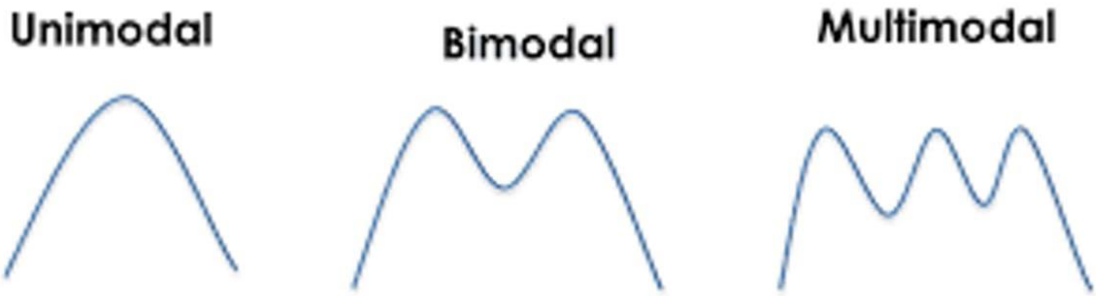
4. Fourth quartile: the highest 25% of numbers



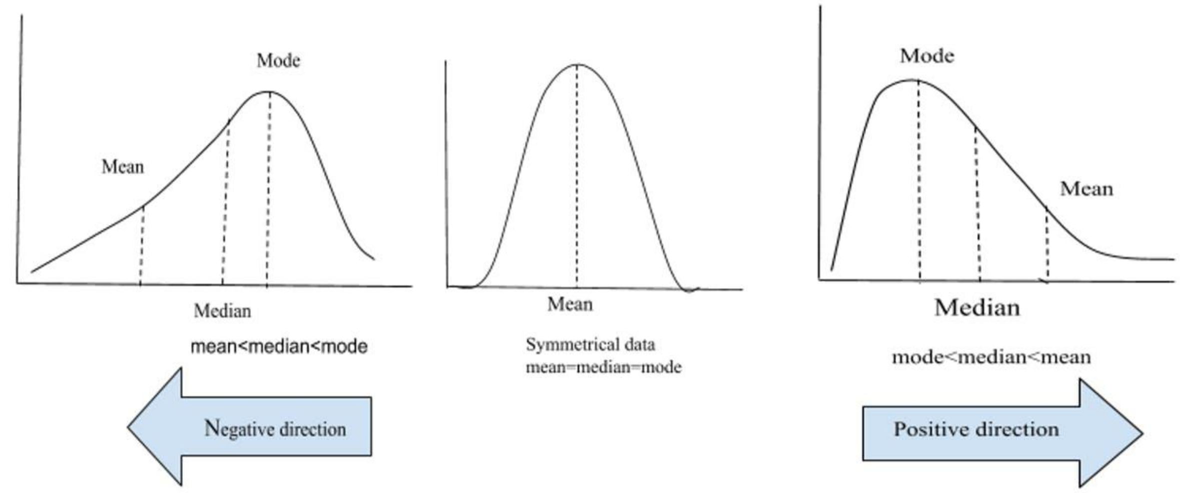
Q1 - 25 percentile, Q2 - 50 percentile and Q3 - 75 percentile.

MODULE -3 -Distribution Of Shape-

- **No. of peaks -**



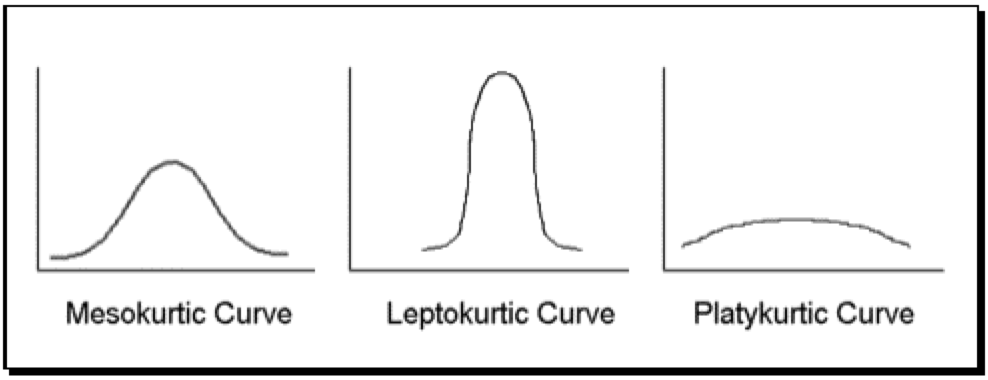
- **Skewness -** Measure of Symmetry



Skewness(S) = 3(μ-Median)/ σ

e.g - If we analyze the data of waiting time at a bank for cheque submission, for a month: The data will be positively skewed distribution as the first 10 days of the month the waiting time will be more has many people will come to issue their salaries.

- **Kurtosis - Measure peakness of data.**



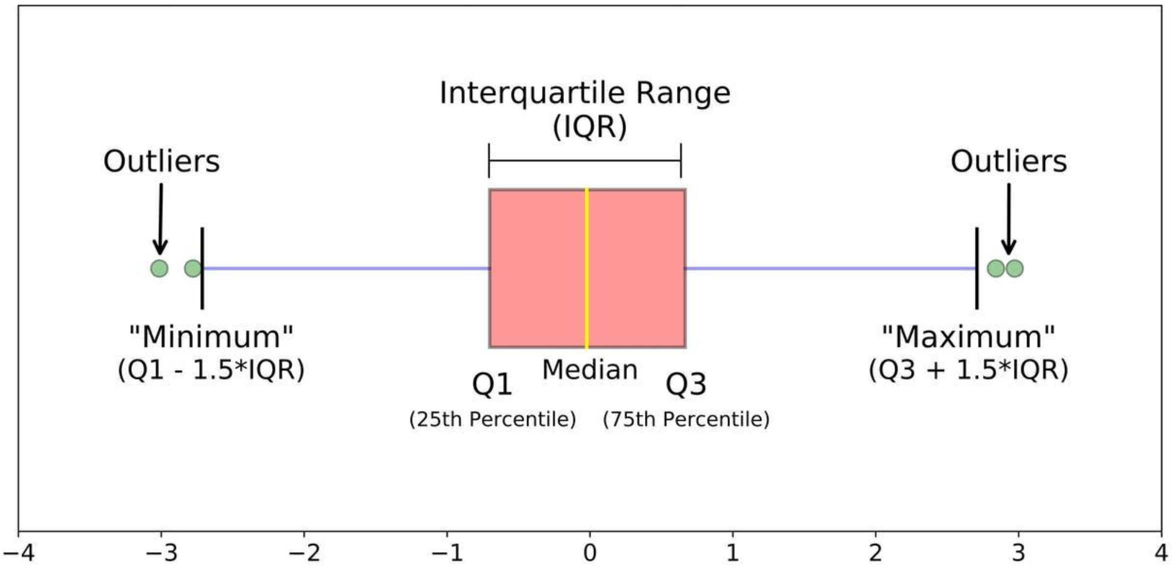
Leptokurtic- High and Thin, positive kurtosis

Mesokurtic - normally distributed. 0 kurtosis

Platykutic- flat and spread out, negative kurtosis

e.g - if the distribution of salaries of employees is Leptokurtic-> More people have higher salaries, if it is Mesokurtic -> The salaries are equally distributed, Platykutic-> more people have fewer salaries.

- **Box Plot -** Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).



- Coefficient of variation- measure of the dispersion of data points in a data series around the mean.

CV= *σ /μ*

MODULE-4 - OUTLIERS- data points that are far from other data points.(Extreme values). ***We need to identify outliers and discard it from the data series*** before making any further observation so that the conclusion made from the study gives more accurate results not influenced by any extremes or abnormal values.

WAYS OF FINDING OUTLIERS-

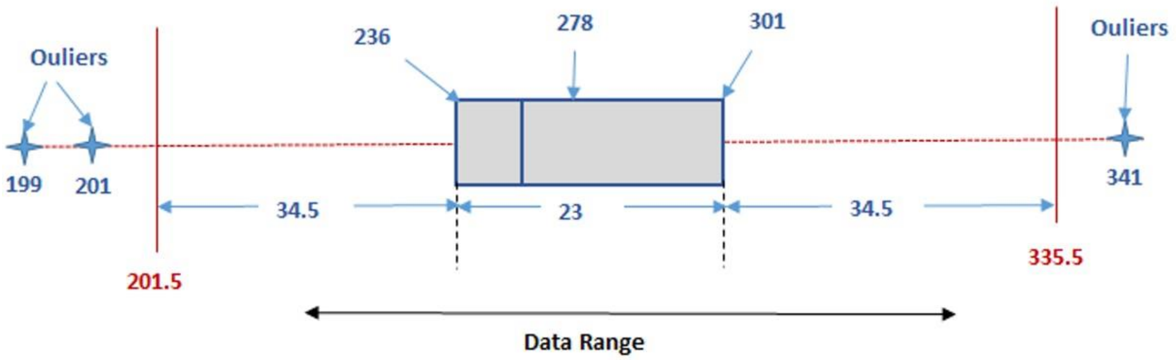
- Boxplots - Max limit - Q1-1.5xIQR, Min limit - Q3 +1.5*IQR

So any value that will be more than the upper limit or lesser than the lower limit will be the outliers.

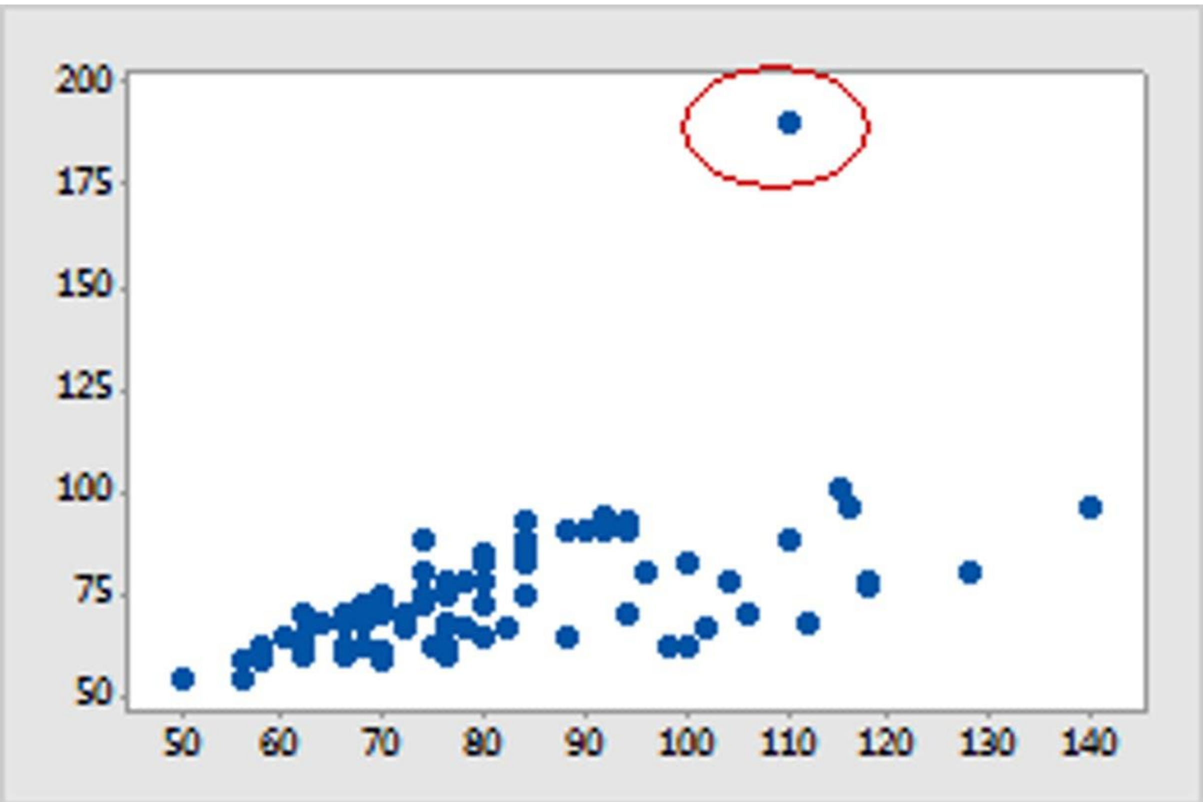
e.g -Let the data range be 199, 201, 236, 269,271,278,283,291, 301, 303, and 341

$$\textbf{Lower Quartile (Q1)} = \frac{1}{4}\ (n+1)\ th\ term$$

$$\textbf{Upper Quartile (Q3)} = \frac{3}{4}\ (n+1)\ th\ term$$



- Scatter plots -



This is used when x and y-axis values are continuous variables. The far aways data points are outliers.