DESCRIPTIVE STATISTICS

(3) measure of central tendency:-

measures of central tendency are statistical metrics that describe the center point or typical value of a dataset. They provide a single value that summarizes a set of data by identifying the central position within the dataset.

some measures of central tendency are:-

(i) mean / Average

(ii) median

(iii) mode

(i) mean:-

→ mean is the sum of all values divided by the number of values.

→ Population is denoted by N and sample is denoted by n and since sample is a subset of population data, we can say $\boxed{n \leq N}$

→ Population mean is denoted by μ.

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

where x is a random variable.

→ Sample mean is denoted by $\bar{x}$

$$\boxed{\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}}$$

Characteristics of mean:-

→ Affected by extreme outliers

→ used for interval and ratio data.

Illustration to show that mean is affected by an outlier:-

Let us take a sample data, $X = \{1,2,3,4,5\}$

Sample (n) = Number of elements in $X = 5$

$$\text{sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{5} \frac{x_i}{5}$$

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

Now, suppose there is an outlier in data i.e number that doesn't belong to the distribution

$X = \{1,2,3,4,5,100\}$

sample $(n) = 6$

$$\text{sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1+2+3+4+5+100}{6}$$

$$= \frac{115}{6} = 17.50$$

Just adding one outlier to the sample data increased the mean from 3 to 17.50. So, we can clearly say that the mean is affected by extreme outliers.

We can remove outlier before finding the mean but it leads to loss of data. Since every data point in our dataset is very important, it is

better to choose another method of measuring central tendency instead of removing outliers and losing data. So, we have median.

(ii) Median :-

The median is the middle value in a dataset when the values are arranged in ascending or descending order.

Illustration to show affect of outliers on median:-

$X = \{1, 5, 4, 2, 3\}$

Arranging in ascending order,

$X = \{1, 2, 3, 4, 5\}$

$n = 5 \Rightarrow$ odd number

Since n is odd, median is the middle element of the dataset, so median = 3

Now, let us add an outlier,

$X = \{1, 3, 5, 4, 100, 2\}$

Arranging in ascending order,

$X = \{1, 2, 3, 4, 5, 100\}$

$n = 6 \Rightarrow$ even number

Since n is even, median will be the average of middle elements.

$$\text{median} = \frac{3+4}{2} = \frac{7}{2} = 3.5$$

Before addition of outlier, median was 3 and after adding, median is 3.5

So, median is not impacted much by the outliers.

Characteristics of median:-

→ Not affected by extreme outliers.

→ used for ordinal, ration and interval data.

⇒ Mode:-

The mode is the value that appears most frequently in a dataset.

For example, dataset = {2, 4, 4, 6, 6, 7, 7, 7, 9}

mode = 7 (most frequently repeated)

Bimodal example, dataset = {2, 4, 4, 6, 6, 7}

mode = 4, 6 (most frequently repeated)

Characteristics of mode:-

→ Not affected by extreme outliers.

→ used for all scales of measurement of data i.e nominal, ordinal, interval and ratio data.

⇒ choosing most appropriate meesure of central tendency:-

(i) mean:-

mean can be best used when data is symmetrically distributed without outlrers. By providing a mathematical average, it is further useful for statistical calculations.

(ii) Median:-
Median is best used when data is skewed. It can also be used if data contains outliers. By providing the middle value, it represents the center of skewed dataset.

(iii) Mode:-
Mode is best used for categorical data. It is used to identify most common category and also to identify most frequent value of ordinal / interval / ratio data.

→ Application of measure of central tendency in feature engineering:-

In feature engineering, we can use measure of central tendency to replace missing values.

If the data is symmetrical and contains no outliers, then we can replace the missing values with mean.

If the data is skewed and contains outliers, then we can replace the missing values with the median.

If the data is nominal, then we can replace the missing values with mode.

(a) Measure of dispersion :-

Measures of dispersion describe the spread or variability of the dataset. They indicate how much values in the dataset differ from central tendency.

Some common measures of dispersion are:-

(i) Range
(ii) variance
(iii) standard deviation
(iv) Inter Quartile Range (IQR)

(1) Range :-

Range is the difference between maximum and minimum values in a dataset.

Range = maximum value - minimum value

Affect of outliers on range:-

Ages = {10, 14, 18, 17, 16}

Range = maximum value - minimum value

= 18 - 10 = 8

Suppose we have an outlier

Ages = {10, 14, 18, 17, 16, 100}

Range = maximum value - minimum value

= 100 - 10 = 90

Just addition of one outlier changed the value of range from 8 to 90. So, range is affected

by extreme outliers in data.

characteristics of Range:-

→ simple to calculate
→ sensitive to outliers
→ just provides a rough measure of dispersion.

(ii) variance:-

→ variance measures the average square deviation of each value from the mean. It provides a sense of how much the values in a dataset vary.

→ Population variance is denoted by σ^2

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

where, $\mu \Rightarrow$ population mean
$N \Rightarrow$ population

→ sample variance is denoted by s^2

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

where $\bar{x} \Rightarrow$ sample mean
$n \Rightarrow$ sample

we will discuss further about why the population data is divided by N and why sample variance is divided by $n-1$.

Now, let us understand the affect of outliers on variance.

Affect of outliers on variance:-

Size of flower petals = {5, 8, 12, 15, 20}

Population (N) = 5

Population mean (μ) = $\frac{5+8+12+15+20}{5} = \frac{60}{5} = 12$

Population variance (σ^2) = $\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$

$$= \sum_{i=1}^{5} \frac{(x_i-12)^2}{5}$$

$$= \frac{1}{5}\left[(5-12)^2+(8-12)^2+(12-12)^2+(15-12)^2 + (20-12)^2\right]$$

$$= \frac{1}{5}(49+16+0+9+64) = \frac{1}{5}(138)$$

$$= 27.6$$

Now, let us add an outlier

Size of flower petals = {5, 8, 12, 15, 20, 102}

Population (N) = 6

Population mean (μ) = $\frac{5+8+12+15+20+102}{6}$

$$= \frac{162}{6} = 27$$

Population variance (σ^2) = $\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$

$$= \sum_{i=1}^{6} \frac{(x_i-\mu)^2}{6}$$

$$= \frac{1}{6}\left[(5-26)^2+(8-26)^2+(12-26)^2+(15-26)^2+(20-26)^2+(102-26)^2\right]$$

$$= \frac{1}{6}(441+324+196+121+36+5776)$$

$= \frac{1}{6}(6698) = 1116.33$

Addition of one outlier increased the variance from 27.6 to 1116.33. So, it is very sensitive to outliers.

Characteristics of variance:-

→ It provides a precise measure of variability.
→ units of variance of squared of original data units.
→ variance is very sensitive to outliers than range.

(iii) Standard deviation:-

Standard deviation is the square root of variance.

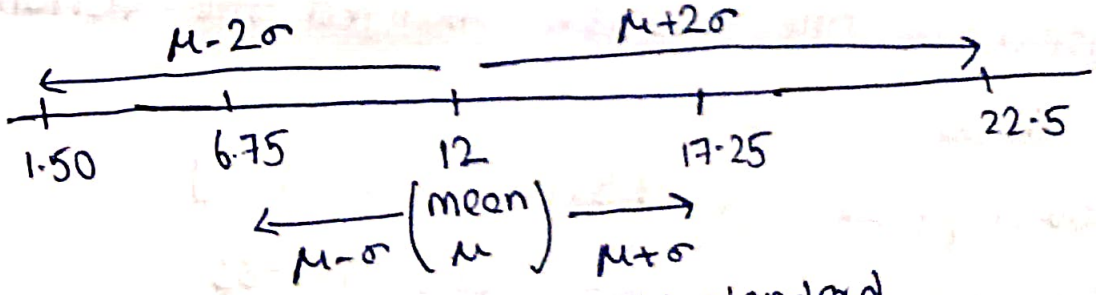
Affect of outliers on standard deviation:-

Let us take the example we have taken in variance where size of flower petals = {5, 8, 12, 5, 20}

Population variance (σ^2) was 27.6

Population standard deviation (σ) = $\sqrt{\sigma^2} = \sqrt{27.6}$

$\Rightarrow \sigma = 5.25$

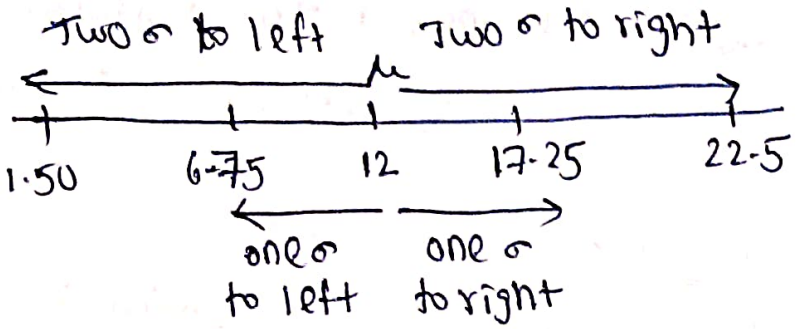


one standard deviation to the left.

one standard deviation to right

Let us see where each data point falls in.

Size of flower petals = {5, 8, 12, 15, 20}



5 ⇒ Lies in range 1.50 to 6.75 so it falls in second standard deviation towards left.

8 ⇒ Lies in range 6.75 to 12. So, it falls in second standard deviation towards left.

12 ⇒ It is the mean so it has zeroth standard deviation from the mean.

15 ⇒ It lies in the range 12 to 17.25. So, it falls in first standard deviation towards the right.

20 ⇒ It lies in range 17.5 to 22.5. So, it lies in the second standard deviation towards the right.

Addition of one outlier changed the variance to 1116.33.

Flower petals size = {5, 12, 8, 15, 20, 100}

Population variance (σ^2) was 1116.33

standard deviation of population = $\sqrt{\sigma^2} = \sqrt{1116.33}$

$= 33.41$

Addition of one outlier increased the value of standard deviation from 5.25 to 33.41. So, standard deviation is also affected by outliers.

Characteristics of standard deviation:-

→ standard deviation provides a clear measure of spread of data in the same units of data.

→ standard deviation is sensitive to outliers.

⑥ Why is sample variance divided by n-1

Population variance $(\sigma^2) = \sum_{i=1}^{N} \frac{(x-\mu)^2}{N}$

sample variance $(s^2) = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$

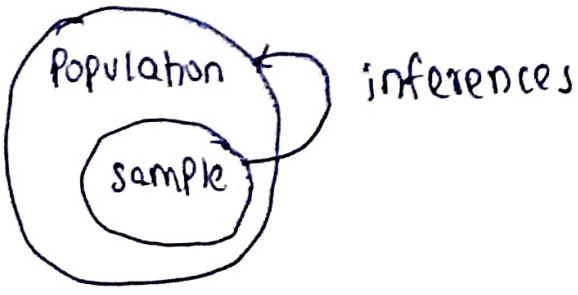
Population variance is divided by N (population) while sample variance is divided by n (sample)-1. This is due to Bessel's correction.

⑦ Bessel's correction :-

Bessel's correction is an adjustment used when calculating sample variance where we divide by n-1 instead of n to obtain an unbiased estimator of the population variance. This correction compensates for the bias introduced when estimating population mean from a sample.

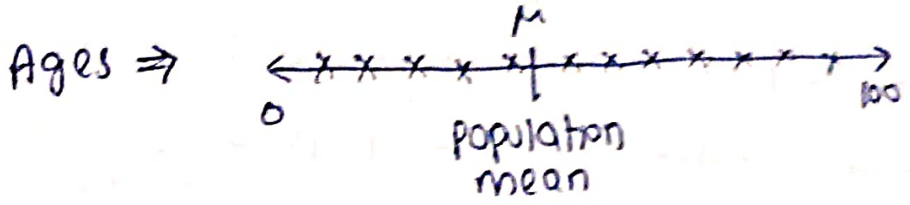
Illustration for why Bessel's correction is required :-

we acutally use sample data to obtain inferences about population data.

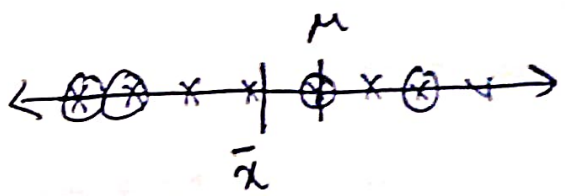


Let us temporarily use the formula

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \quad \text{(temporarily)}$$



If sample data is obtained in a good manner, then

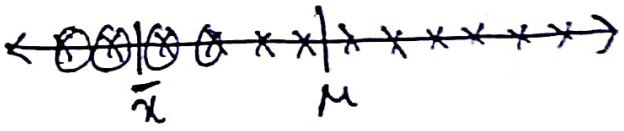


$\bar{x} \approx \mu$

$\sigma^2 \approx s^2$

If sampling is done properly, then the sample mean and sample variance will be somewhere near the population mean and population variance.

But if data is not sampled properly,



In this case

$\bar{x} \lll \mu$

$s^2 \lll \sigma^2$

So, when we use sample variance $(s^2) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$,

we are under-estimating the true value of the population variance. So, instead of dividing the sample variance by n, we divide it by n-1 so that atleast the gap between population variance and sample variance reduces.

so, since n-1 is lesser than n, driding by n-1 increase the value of sample variance to reduce the gap between sample and population variance.

so, the division by n-1 just for unbiased estimation so that we donot under-estimate population variance.

① Random variables:-

Random variables are denoted by X.

Random variables are functions that has values derived from processes or experiments.

Example of random variables:-

→ Tossing a coin

Possibilities ⇒ Head or Tail

$$X = \begin{cases} 0, \text{ if head is obtained} \\ 1, \text{ if tail is obtained} \end{cases}$$

Here we are obtaining values based on an experiement that is tossing a coin.

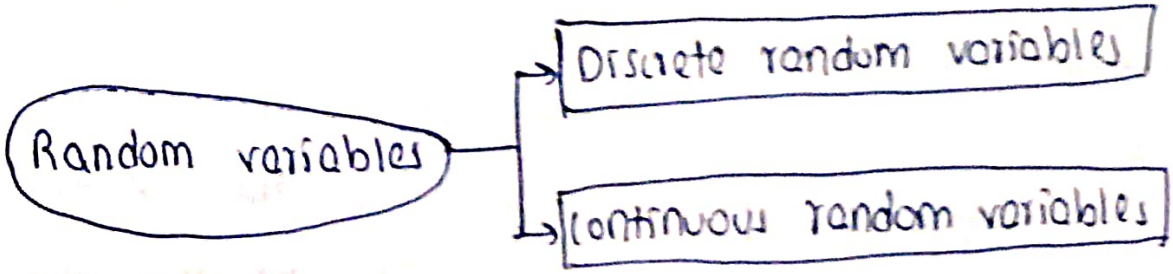
→ Rolling a fair dice

Possibilities ⇒ 1,2,3,4,5,6

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$

Rolling a die is an experiment we are performing and random variable can hold values 1 to 6.

② Types of random variables:-



Let us understand the difference between discrete and continuous random variables.

Discrete random variable:-

→ A discrete random variable can take on a countable number of distinct values. So, it can typically hold countable number of whole numbers.

→ The value which discrete random variables hold are seperated and there cannot be any value between them.

→ Examples of discrete random variables:-

① Number of students absent in the class
It can be 0/1/2.. etc. but cannot be 1.5 or 2.5.

② Number of cars in the parking
It can be 0/1/2.. but cannot be a fractional/decimal value like 1.50, 2.5 etc.

Continuous random variables:-

→ A continuous random variable can take on infinite number of values within a given range.

→ The values held by continous random variables are typically measured and not counted.

→ Examples of continuous random variables:-

① Height of students in the class
It can hold any value like 140.5 cm, 167cm, 168.85cm etc. It can hold both decimal and non-decimal values.

(ii) Temperature in a day

It can hold values like 25°c, 20.5°c, 20.31°c etc. The values can be as much precised as required.

So, we can say that the values of discrete random variables are seperated and countable while the values of continuous random variables can be any fractional number within the range and typically measured but not countable.

⑧ Percentiles and Quartiles :-

→ A percentile is the value below which a certain percentage of observations lie:

Example :-

Numbers = {2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

Percentile of value 9 = $\frac{\text{no. of values below 9}}{\text{Total no. of values}} \times 100$

$= \frac{11}{14} \times 100 = 78.57$ percentile

Percentile of value $x = \frac{\text{No. of values below } x}{\text{sample size}} \times 100$

78.57 percentile of 9 means that 78.57% of the elements are less than 9.

→ To find n^{th} percentile value in a distribution

Example :-

Numbers = {2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

Finding 25th percentile value of numbers.

25^{th} percentile value = $\frac{\text{Percentile}}{100} \times (\text{sample size} + 1)$

$= \frac{25}{100} \times (14+1) = \frac{15}{4} = 3.75$

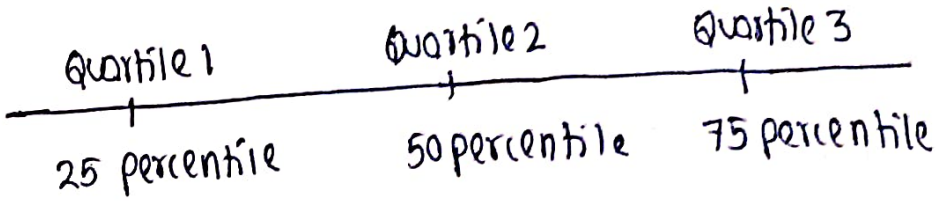
There is no position as such in the elements that is 3.75. So, the value will be the mean of 3rd and 4th element in the elements.

3rd element value = 3
4th element value = 4

mean = $\frac{3+4}{2} = 3.50$

So 25th percentile value in the numbers distribution is 3.50.

→ Quartiles represent every 25th percentile value of a distribution.



Quartile 1 ⇒ 25th percentile value in the distribution

$= \frac{25}{100} \times (n+1)^{th}$ value $= \frac{1}{4}(n+1)^{th}$ value

Quartile 2 ⇒ 50th percentile value in the distribution

$= \frac{50}{100}(n+1)^{th}$ value $= \frac{1}{2}(n+1)^{th}$ value in distribution

Quartile 3 ⇒ 75th percentile value in the distribution

$\Rightarrow \frac{75}{100}(n+1)^{th}$ value $= \frac{3}{4}(n+1)^{th}$ value in distribution.

Quartile 1 = $\frac{1}{4}(n+1)^{th}$ value in the distribution.

Quartile 2 = $\frac{2}{4}(n+1)^{th}$ value in the distribution

Quartile 3 = $\frac{3}{4}(n+1)^{th}$ value in the distribution

(*) Five number summary:-

→ Five number summary is a quick way to describe a set of data using just five key values:-

(i) minimum
(ii) Q_1 (First Quartile)
(iii) median
(iv) Q_3 (Third Quartile)
(v) maximum

→ Five number summary quickly shows the spread and center of the data, helps spot outliers and understand the range of values.

Example of five number summary:-

Data points = {1,2,2,2,3,3,4,5,5,5, 6,6,6,6, 7,8,8,9, 27}

we first define lower fence and higher fence values

Lower fence ⟷ Higher fence

All the values lower than lower fence and higher than higher fence are removed as an outliers.

lower fence = $Q_1 - 1.5(IQR)$
Higher fence = $Q_3 + 1.5(IQR)$

IQR is Inter-Quartile Range

$IQR = Q_3 - Q_1$

Finding values of Q_1 and Q_3

$Q_1 = \frac{25}{100}$ (sample size $+1)^{th}$ value of the distribution

$= \frac{1}{4}(19+1) = 5^{th}$ value of the distribution $= 3$

$Q_3 = \frac{75}{100}$ (sample size$+1)^{th}$ value of the distribution

$= \frac{3}{4}(19+1) = 15^{th}$ value of the distribution $= 7$

Finding value of Inter-Quartile range,

$IQR = Q_3 - Q_1 = 7-3 = 4$

Finding values of lower and higher fence,

Lower fence $= Q_1 - 1.5(IQR) = 3 - 1.5(4) = 3-6 = -3$

Higher fence $= Q_3 + 1.5(IQR) = 7 + 1.5(4) = 7+6 = 13$

[Lower fence $\leftrightarrow$ Higher fence] $\Rightarrow$ [$-3 \leftrightarrow 13$]

In our data points, there are no values lesser than -3 but there are values greater than the higher fence(13). So, they can be removed as outliers

Data points = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,(17)} ← X

more than higher fence → outlier

Post removal, our data points are

Data points = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9}

Now let us findout our 5 values in 5 number summary that are minimum, maximum, median, Q_1, Q_3

Minimum = 1

Maximum = 9

Median ⇒ central value of the distribution

since sample size is even, median is the mean of middle elements

$\Rightarrow \frac{5+5}{2} = 5$

$Q_1 = \frac{25}{100}(n+1)^{th}$ value $= \frac{1}{4}(19) = 4.75^{th}$ value

= Average of 4^{th} and 5^{th} value

$= \frac{2+3}{2} = 2.50$

$Q_3 = \frac{75}{100}(n+1)^{th}$ value $= \frac{3}{4}(19) = 14.25^{th}$ value

= Average of 14^{th} and 15^{th} value

$= \frac{6+7}{2} = 6.50$

So, the five number summary is

minimum = 1, $Q_1 = 2.50$, median = 5, $Q_3 = 6.50$, maximum = 9

Plotting a box plot helps us to visualize outliers. We will learn how to plot further.

(A) Histogram and skewness :-

→ A histogram is a graphical representation of the distribution of numerical data. It is an estimate of probability distribution of continuous variables and is used to visualize the shape, central tendency and variability of the dataset.

Example :-

Ages = {11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50}

Range = 0 to 50 (maximum age)

If we need 10 bins in the histogram, then the number of bins will be $(50-0)/10 = 5$

Bin sizes :-

0-5 ⇒ no. of elements ≥ 0 and < 5 ⇒ 0

5-10 ⇒ no. of elements ≥ 5 and < 10 ⇒ 0

10-15 ⇒ no. of elements ≥ 10 and < 15 ⇒ 3 {11, 12, 14}

15-20 ⇒ no. of elements ≥ 15 and < 20 ⇒ 1 {18}

20-25 ⇒ no. of elements ≥ 20 and < 25 ⇒ 1 {24}

25-30 ⇒ no. of elements ≥ 25 and < 30 ⇒ 1 {26}

30-35 ⇒ no. of elements ≥ 30 and < 35 → 1 {35}

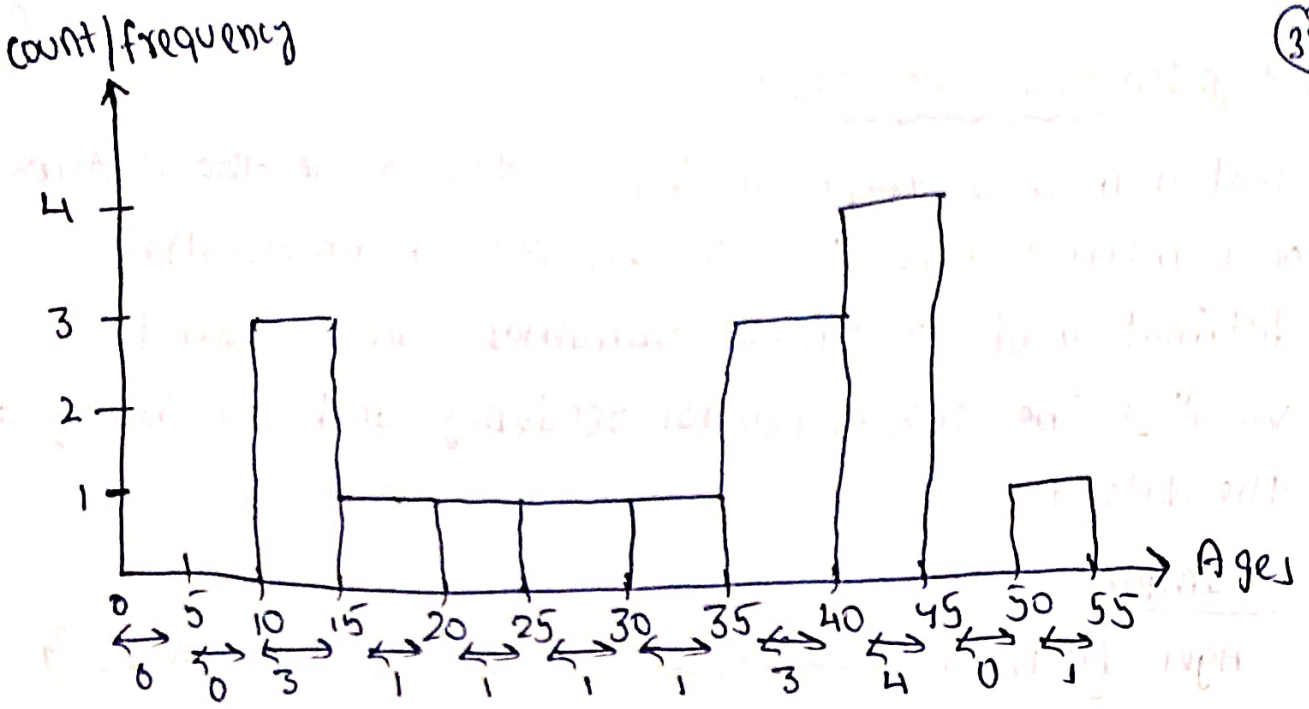
35-40 ⇒ no. of elements ≥ 35 and < 40 ⇒ 3 {35, 36, 37}

40-45 ⇒ no. of elements ≥ 40 and < 45 ⇒ 4 {40, 41, 42, 43}

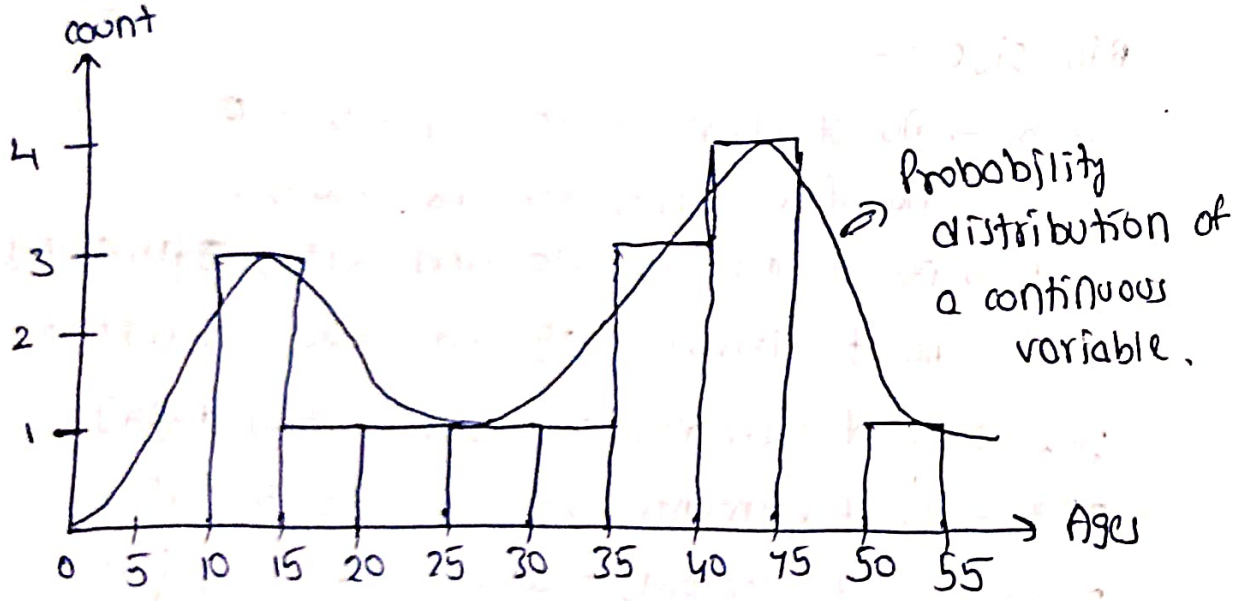
45-50 ⇒ no. of elements ≥ 45 and < 50 ⇒ 0

50-55 ⇒ no. of elements ≥ 50 and < 55 ⇒ 1 {50}

Since we now have the count, let us plot the histogram



we can obtain the probability distribution of this ages by smoothening the histogram.



To perform this smoothening of histogram to obtain probability distribution of continuous variable using a technique called <u>Kernel Density Estimator</u>.

→ Skewness is a statistical measure that describes the asymmetry of a distribution about how much data is spread on one side of the mean compared to other.

There are different types of skewness cases :-

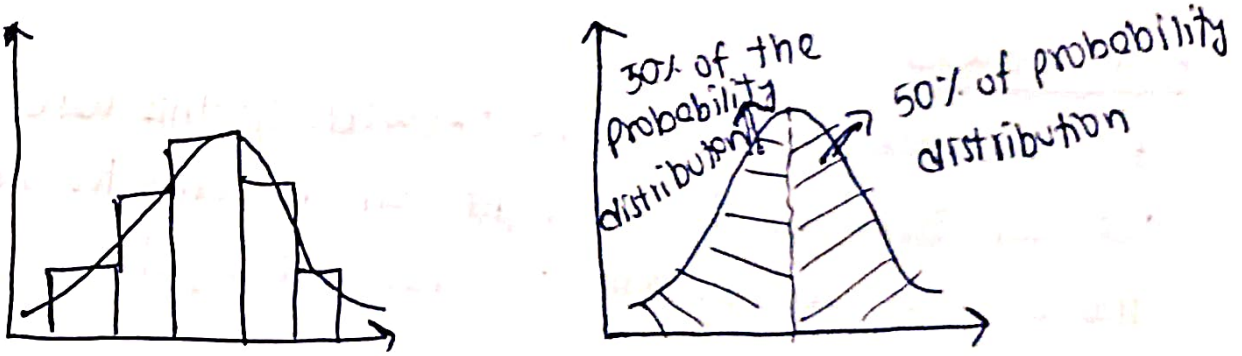
(i) If data is symmetric (like a perfect bell curve), then the skewness is zero.

(ii) If the tail (extreme values) stretches further to the right, it is called positive/right skewness.

(iii) If the tail stretches further to the left, it is called a left/negative skewness.

→ Symmetric data :-

Here we obtain a perfect bell curve. The probability distribution of such data is called normal/Gaussian distribution.

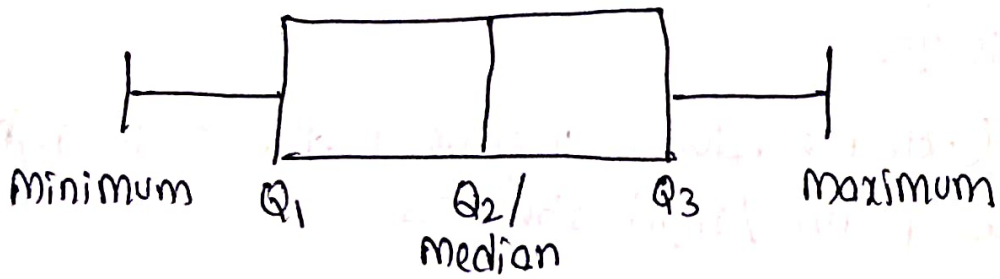


Normal/Gaussian distribution

In such distribution,

- Skewness is zero
- 50% of probability distribution is present on either sides of the mean i.e it is symmetric.
- The mean, median and mode of the data in such distribution perfectly lie at the centre.
- The differences between third and second quartile will be similar to difference between second and first quartile.

Box plot of gaussian / normal distribution looks like:.



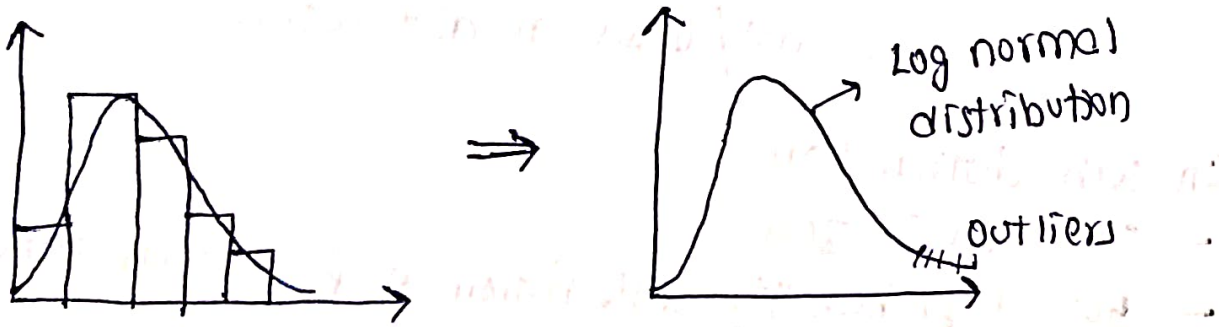
$$Q_3 - Q_2 \approx Q_2 - Q_1$$

The mean, median and the mode are exactly at the center in symmetric data.

$$\text{Mean} = \text{Median} = \text{Mode}$$

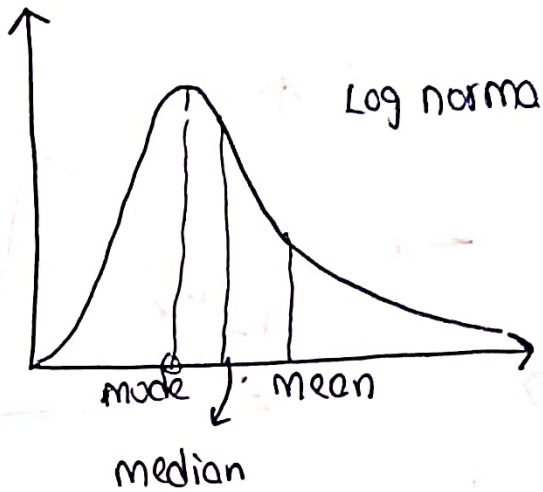
⇒ Right skewed :-

It is also known as positive skewed. In this case the tail stretches to the right which indicates that the outliers are present towards the right.



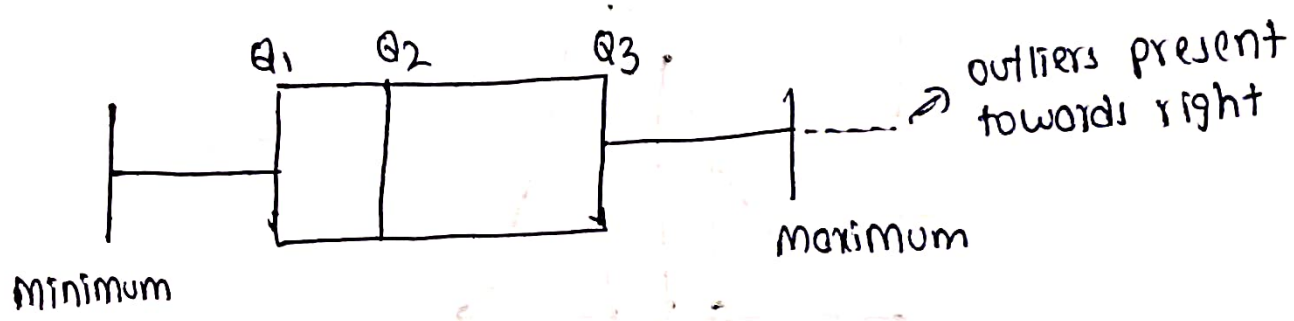
The probability distribution obtained after smoothening such kind of data is also called as log normal distribution.

understanding relation between mean, median and mode in right skewed distribution.



$$\text{Mean} \geq \text{median} \geq \text{mode}$$

The box plot of such distribution looks like

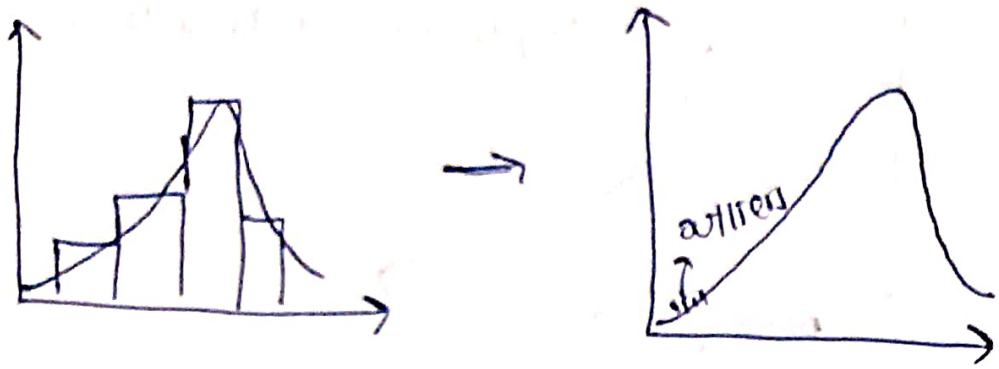


$$Q_3 - Q_2 \geq Q_2 - Q_1$$

This is because most of the values in the distribution are at the lower values and very few values extend the tail on the higher side. so, the number of values between Q_1 and Q_2 inspite of less difference will be equal to number of values between Q_3 and Q_2 even though the difference between Q_3 and Q_2 is very high.

→ Left skewed :-

It is also known as negative skewed. In this case, the tail (outliers) extend towards the left.

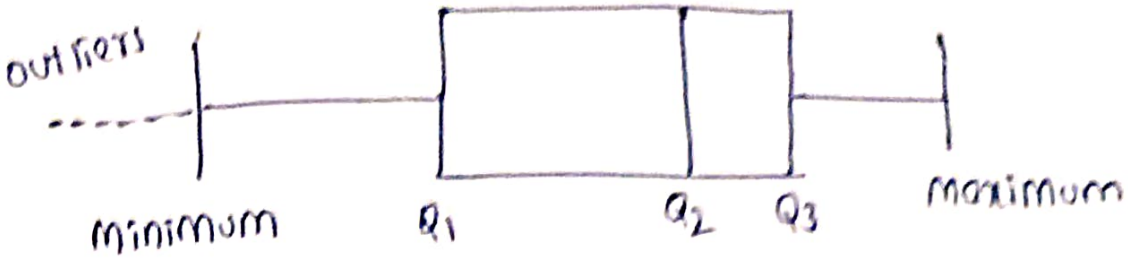


In this, the outliers have lower value and hence significantly lower the mean.

For left skewed / negtive skewed distribution

$$\text{mean} \leq \text{median} \leq \text{mode}$$

In box plot,
since outliers are stretched towards the left, the difference between Q_2 and Q_1 will be more than that of Q_3 and Q_2



$$Q_2 - Q_1 \geq Q_3 - Q_2$$

Example of left skewed data can be the retirement age data where we can find most people retiring between 55 to 60 while very less number of people retire at age less than 25.

⑧ Correlation and covariance

Correlation and covariance are two statistical measures used to determine the relationship between two variables. Both are used to understand how change in one variable are associated with change in another variable.

⇒ Covariance:-

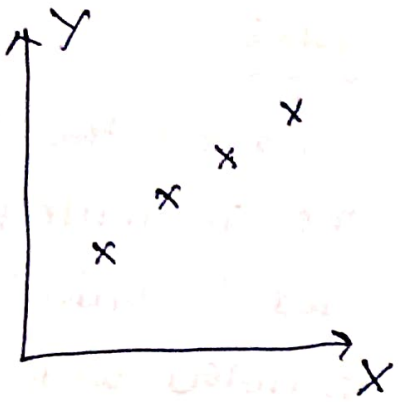
Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive and if one tends to increase while other decreases, the covariance is negative.

Example:-

size of the house	Price of the house
1200 sq.ft	40 lakhs
1300 sq.ft	45 lakhs
1500 sq.ft	55 lakhs

when size of house increases, the price also increases and when size decreases, the price also decreases. So, we can say that the covariance between the variables size of the house and price of the house is positive.

X	Y
2	3
3	4
4	5
5	6

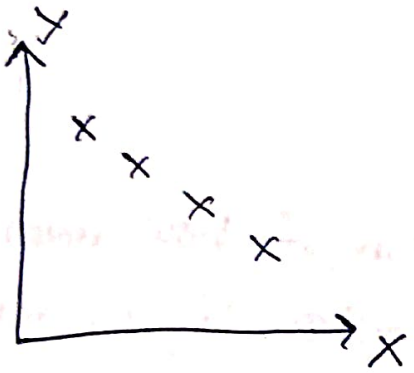


$X\uparrow \Rightarrow Y\uparrow$

$X\downarrow \Rightarrow Y\downarrow$

Negative (x)
Positive covariance (V)

X	Y
2	7
3	6
4	5
5	4



$X\uparrow \Rightarrow Y\downarrow$

$X\downarrow \Rightarrow Y\uparrow$

Negative covariance

Formula for covariance is

$$\text{covariance}(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

covariance between the same random variable is its sample variance.

$$\text{covariance}(X,Y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$\text{cov}(X,X) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(x_i-\bar{x})}{n-1} = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$$

$$= \text{variance}(X)$$

$$\boxed{\text{cov}(X,X) = \text{variance}(X)}$$

$x_i \Rightarrow$ Datapoints of random variable X

$\bar{x} \Rightarrow$ sample mean of x

$y_i \Rightarrow$ datapoints of random variable Y

$\bar{y} \Rightarrow$ sample mean of Y.

$\Rightarrow$ Example:-

Number of students study hours (x)	marks obtained (y)
2	50
3	60
5	70
6	80

Since, $X\uparrow \Rightarrow Y\uparrow$, $X\downarrow \Rightarrow Y\downarrow$ } The covariance should be positive.

Let us verify with formula.

$$\boxed{cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}}$$

$$\bar{x} = \frac{2+3+5+6}{4} = \frac{16}{4} = 4$$

$$\bar{y} = \frac{50+60+70+80}{4} = \frac{260}{4} = 65$$

$$cov(X,Y) = \sum_{i=1}^{4} \frac{(x_i-4)(y_i-65)}{4-1} = \frac{1}{3}\sum_{i=1}^{4}(x_i-4)(y_i-65)$$

$$= \frac{1}{3}\left[(2-4)(50-65)+(3-4)(60-65)+(5-4)(70-65)+(6-4)(80-65)\right]$$

$$= \frac{1}{3}\left[-2(-15)+(-1)(-5)+(1)(15)+2(25)\right]$$

$$= \frac{1}{3}\left[30+5+15+50\right] = \frac{1}{3}(100) = 33.33 > 0$$

Positive covariance indicates that the number of hours studied increases the exam score also.

Advantages and disadvantage of covariance:-

→ we can quantify the relationship between X and Y i.e we can identify how X and Y move together. This is a good advantage.

→ Disadvantage is that the covariance doesnot have a specific limit value i.e

$$cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

its value can lie anywhere between $-\infty$ and $+\infty$ so, if covariance values between A and B is more than covariance value of B and C, we cannot say that covariance of A and B is more than covariance of B and C i.e if $cov(A,B) = 50$ and $cov(B,C) = 20$, we cannot compare that $cov(A,B) > cov(B,C)$.

since we cannot compare solely based on values in covariance, we use another statistical technique called correlation.

⇒ **correlation :-**

There are two types of correlation :-

(i) Pearson correlation coefficient

(ii) spearman correlation coefficient

(A) Pearson correlation co-efficient :-

It limits the values from [-1 to +1]

It is represented by $\rho_{x,y}$

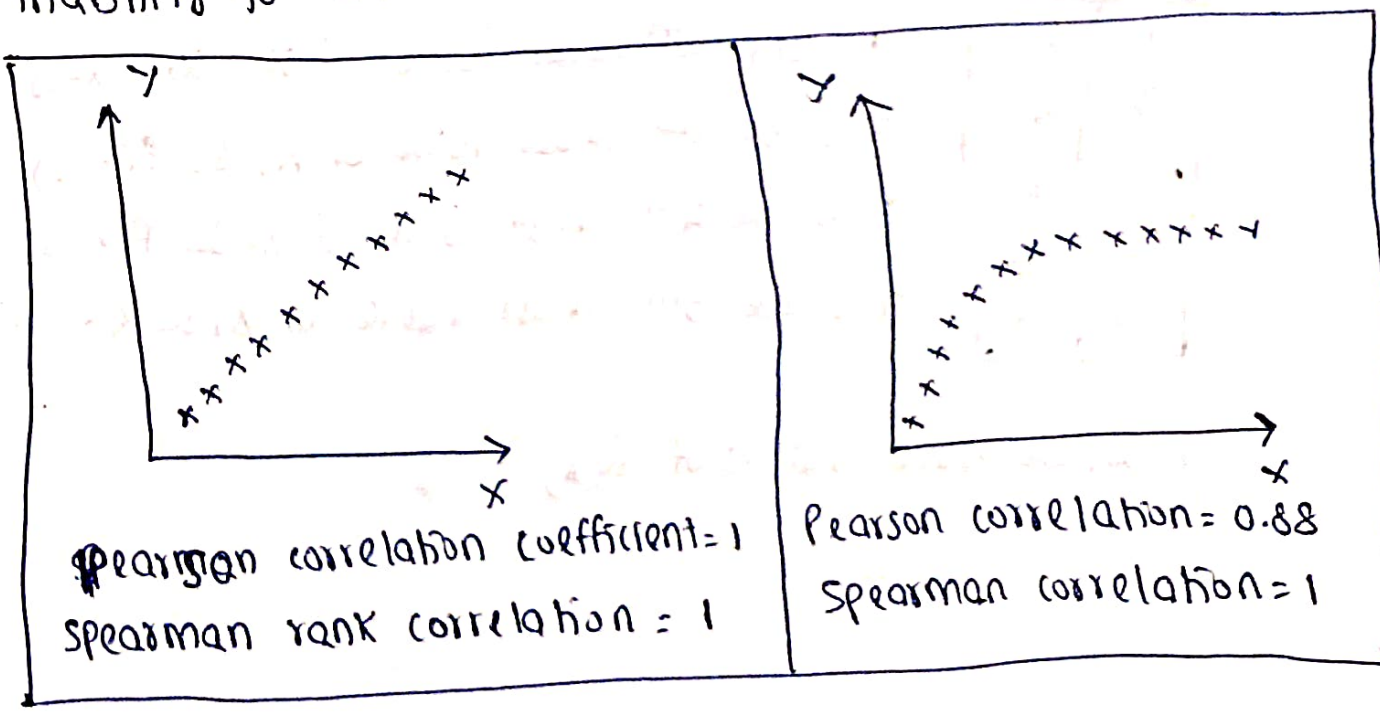
$$\rho_{x,y} = \frac{\text{covariance}(x,y)}{\sigma_x \cdot \sigma_y}$$

→ The more the value towards +1, the more is the positive correlation between x and y.

→ The more the value towards -1, the more is the negative correlation between x and y.

(A) Spearman rank correlation :-

It fixes the issue of pearson correlation co-efficient's inability to deal with non-linear data.



Actually, in both the cases with an increase in Y, there is an increase in X. So, ideally the correlation must be 1. But, in case of non-linear data, the value of pearson correlation coefficient is 0.88.

This is because even though values of X and Y increase at the same time, at certain points the amount of increase is less. So, we use spearman rank correlation to fix this.

spearman rank correlation is denoted by r_s

$$r_s = \frac{\text{covariance}(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

where $R(x)$ = Rank of x

$R(y)$ = Rank of y

Example:-

X	Y
1	2
3	4
5	6
7	8
0	7

0 is the 1st least value of X. so $R_{(x=0)} = 1$

1 is the 2nd least value of X, so $R_{(x=1)} = 2$

3 is the 3rd least value of X, so $R_{(x=3)} = 3$

5 is the 4th least value of X. So $R_{(x=5)} = 4$

7 is the 5th least value of X, so $R_{(x=7)} = 5$

Now, we have values of $R(x)$

X	Y	R(x)
1	2	2
3	4	3
5	6	4
7	8	5
0	7	1

similarly for R(y),

2 is the 1st least element of $Y \Rightarrow R(Y=2) = 1$

4 is the 2nd least element of $Y \Rightarrow R(Y=4) = 2$

6 is the 3rd least element of $Y \Rightarrow R(Y=6) = 3$

7 is the 4th least element of $Y \Rightarrow R(Y=7) = 4$

8 is the 5th least element of $Y \Rightarrow R(Y=8) = 5$

X	Y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

So, to overcome the disadvantage of pearson correlation coefficient not doing well for non-linear relationship between x and y, we use spearman rank correlation.