COVARIANCE AND CORRELATION 21.4.24

COVARIANCE

Covariance in statistics refers to the measure of the relationship between two random variables and how they change together. It defines the changes between two variables, indicating whether they move in the same (positive covariance) or opposite (negative covariance) directions. The covariance formula calculates this relationship, considering the mean of the variables and the number of data values. To calculate the coefficient of covariance for a specific dataset, you need to find the mean of X and Y, then apply the covariance formula using the given data values

CORRELATION

Correlation is a statistical measure of the relationship between two variables, indicating the degree to which they move together, either in the same direction (positive correlation) or in opposite directions (negative correlation). The strength of the relationship is measured by the correlation coefficient, which ranges from -1 to +1. A correlation coefficient of +1 indicates a perfect positive correlation, while a coefficient of -1 indicates a perfect negative correlation. A coefficient of 0 indicates no correlation between the variables.

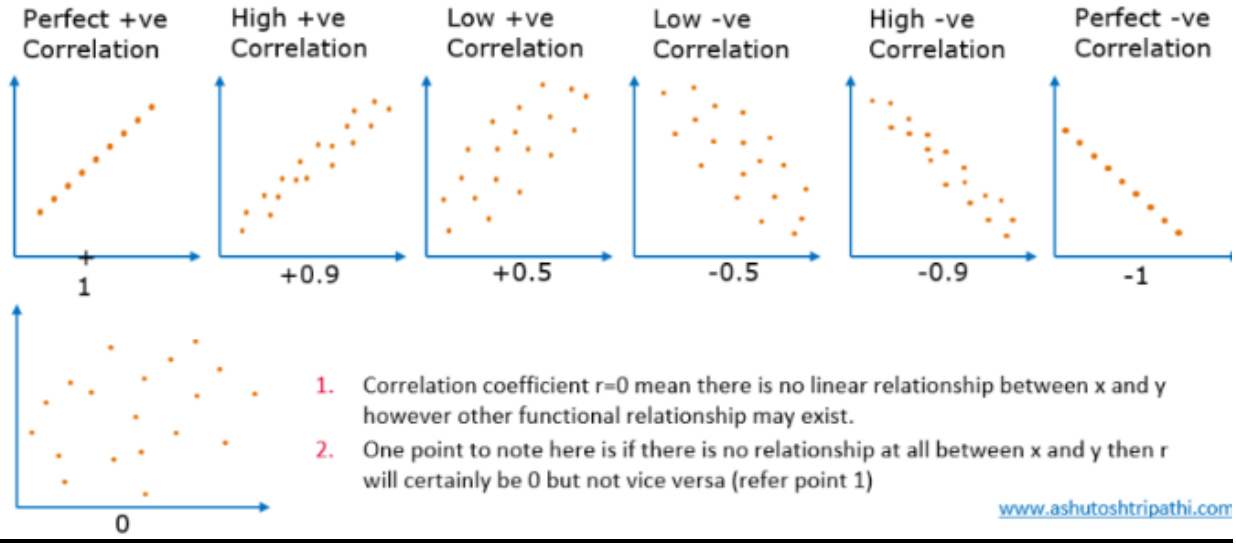
Covariance	Correlation
Indicates the direction of the linear relationship between variables	Indicates both the strength and direction of the linear relationship between two variables
Covariance values are not standard	Correlation values are standardized
Positive number being positive relationship and negative number being negative relationship	1 being strong positive correlation, -1 being strong negative correlation
Value between positive infinity to negative infinity	Value is strictly between -1 to 1

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.



1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

www.ashutoshtripathi.com

COVARIANCE OF PREPLACEMENT DATA

[8]:
```
dataset[numeric_columns].cov()
```

[8]:

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
ssc_p	117.228377	58.853253	42.702550	37.659225	24.535952	9.088585e+05
hsc_p	58.853253	112.063731	33.684453	33.838355	21.517688	7.310079e+05
degree_p	42.702550	33.684453	53.604710	22.078774	17.185200	4.663363e+05
etest_p	37.659225	33.838355	22.078774	176.251018	16.886973	3.727004e+05
mba_p	24.535952	21.517688	17.185200	16.886973	34.028376	1.239934e+05
salary	908858.485818	731007.850848	466336.264888	372700.449468	123993.387361	2.259185e+10

A **positive covariance** would indicate that as the marks in senior secondary increase, the marks in higher secondary also tend to increase, showing a direct relationship.

Conversely, a **negative covariance** would suggest an inverse relationship, where higher marks in senior secondary are associated with lower marks in higher secondary.

A covariance of **zero** would indicate no linear relationship between the pass marks in senior secondary and higher secondary education

---

SummaryDifferentiation

A covariance of **degree_p** with **etest_p** is a (**POSITIVE COVARIANCE**). Showing a direct relationship. It Indicates the marks in degree_p increase, the marks in e_test also increase.

A covariance of **e_test** with **mba_p** is a (**POSITIVE COVARIANCE**). Showing a direct relationship. It Indicates the marks in e_test increase, the marks in mba_p also increase.

CORRELATION

```
dataset[numeric_columns].corr()
```

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
ssc_p	1.000000	0.513478	0.538686	0.261993	0.388478	0.558475
hsc_p	0.513478	1.000000	0.434606	0.240775	0.348452	0.459424
degree_p	0.538686	0.434606	1.000000	0.227147	0.402376	0.423762
etest_p	0.261993	0.240775	0.227147	1.000000	0.218055	0.186775
mba_p	0.388478	0.348452	0.402376	0.218055	1.000000	0.141417
salary	0.558475	0.459424	0.423762	0.186775	0.141417	1.000000

The correlation between **senior secondary pass** marks and **higher secondary pass** marks refers to the **degree** to which these two variables are linearly related.

A positive correlation would indicate that as the senior secondary pass marks increase, the higher secondary pass marks also tend to increase.

A negative correlation would indicate that as the senior secondary pass marks increase, the higher secondary pass marks tend to decrease.

A correlation coefficient of 0 would indicate no linear relationship between the two variables.

SummaryDifferentiation

- A correlatoin of **ssc_p** with **hsc_p** is a Positive Covariance. Showing a direct relationship. It Indicates**(students who perform well in their ssc_p are also likely to perform well in hsc_p)** the marks in degree_p increase, the marks in e_test also increase. But the Strength of this correlation is 0.513478 (**LOW POSITIVE CORRELATION**)

- A correlation of **mba_p** with **ssc_p** is a Positive Covariance. Showing a direct relationship. It Indicates the marks in mba_p increase, the marks in ssc_p also increase. But the Strength of this correlation is 0.388478 (**LOW POSITIVE CORRELATION**)

- A correlation of **mba_p** with **salary** is a Positive Covariance. Showing a direct relationship. It Indicates the marks in mba_p increase, the marks in salary also increase. But the Strength of this correlation is 0.141417 (**MINIMAL CORRELATION or ZERO CORRELATION**)