Correlation:-

* Correlation is the statistical term that tells the degree of the relationship.

* It means whether x and y have a strong relationship or weak relationship.

* Correlation analysis deals with the association between 2 or more Variables.

* Association means x and y may have a causal relationship, means x is the cause of y that is change in y is the effect of change in x.

* If there is a mutual dependence among x and y, we cannot say which is the cause and which one is effect.

* For example, price of commodity is affected by demand and supply.

(*) If x and y are correlated then x and y may or may not have casual relationship

(*) If x and y have a casual relationship then x and y must be correlated.

* Any third common factor can also influence the correlation between 2 variables x and y for example between production of tea and rice per hectare, here they are not directly correlated, instead the cause is the good rainfall well in time.

* There are types of correlation, and the types were categorised.

i) Based on degree of correlation:-

i) Positive correlation:-

* Values of 2 variables move in same direction either $x\uparrow$ $y\uparrow$ or $x\downarrow$ $y\downarrow$

* ex:- Age and income
Amount of rain and yield of crop.

ii) Negative correlation:-

* Values of 2 variables move in opposite direction either $x\uparrow$ $y\downarrow$ or $x\downarrow$ $y\uparrow$

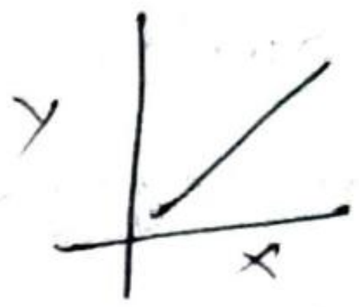
* ex:- Height above sea level and temperature
- Sales of woolen clothes and temp.

II) Based on change in proportion:-

i) Linear correlation:-

* change in amount of one variable tends to preserve a constant ratio to amount of change in other Variable.

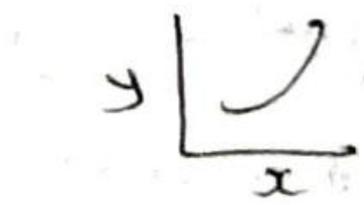
*ex:- whenever price rises by 10·1· then supply rises by 20·1·



ii) Non-Linear correlation:-

* Amount of change in on Variable does not tends to preserve a constant ratio.

*ex:- whenever price rises by 10·1·, sometimes it rise by 10·1·, sometime by 20·1· etc --



III) Based on no. of Variables studied:-

i) Simple correlation - using 2 variables

ii) Multiple correlation - more than 2 Variables

ex:- relation b/w yield of rice per hectare and both the amount of rainfall along with the no of fertilizers were used to find rice production.

iii) Partial correlation -

* when one or more var kept constant and relationship studied b/w other 2 variables.

eg:- Relationship b/w rainfall and rice yields under constant temp.

Pearson correlation coefficient:- (r)

*Pearson correlation coefficient is the linear correlation coefficient that returns a value of between −1 and +1.

*In other words it gives the strength and direction of the linear data.

* Formula

$$r = \frac{cov(x,y)}{\sigma_x * \sigma_y}$$

* correlation quantifies the strength of the relationship

*And there is something called confidence which tells how good the correlation is.

* P-value which tells how confident is our correlation.

* Here the correlation should reject the null hypothesis.

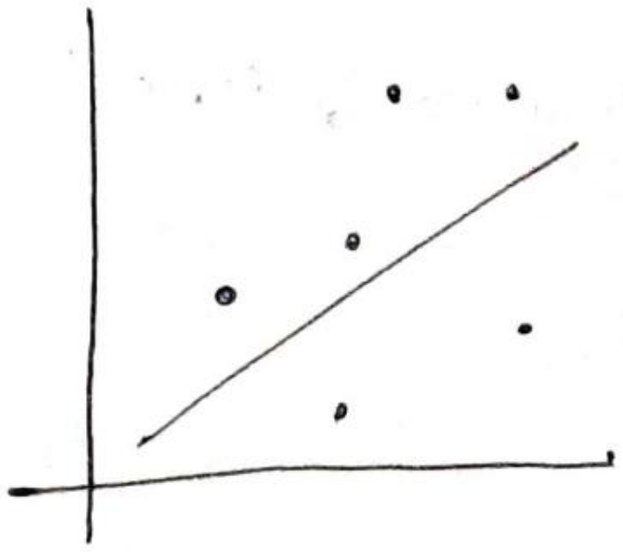
* The null hypothesis is 2 variables x, y are not correlated means $r = 0$

* So we should reject this null hypothesis in order to accept the Alternate hypothesis, that is 2 variables are correlated.

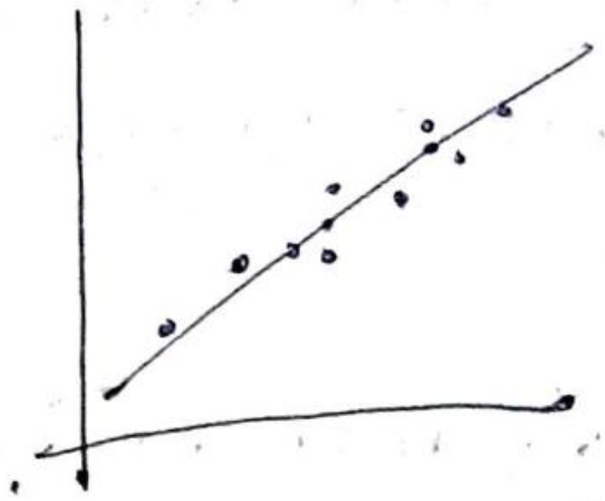
* If the P-value is very low, or it is less than the level of significance then the null hypothesis will be rejected.

* If there is more data, P-value will be low, but if the correlation value is less- then, we should not accept this.

* what i am trying to say is if the P-value is very low, we cannot blindly decide that x and y were strongly correlated.



→ weakly correlated.



→ Strongly correlated

→ no Correlation

* If the Variables were strongly correlated then we can make a educated guess for new input for x and predicted the y value for it.

* If it is weakly correlated there will be narrow range of values, we cannot make a correct guess.

* R^2 Value tells how the relationship between the 2 variables explains the variation in data.

* It is the square of correlation coefficient (r)

* If Coefficient R is 0.9 then R^2 is 0.81.

* So, it tells The relationship blw 2 variables explains 81.1. of variation in the data..

* remaining done by some other else.

* The disadvantage of Pearson correlation is it does not work for non-linear correlation between Variables.

* It also doesn't work for monotonic and non-monotonic data.

Spearman Rank correlation coefficient

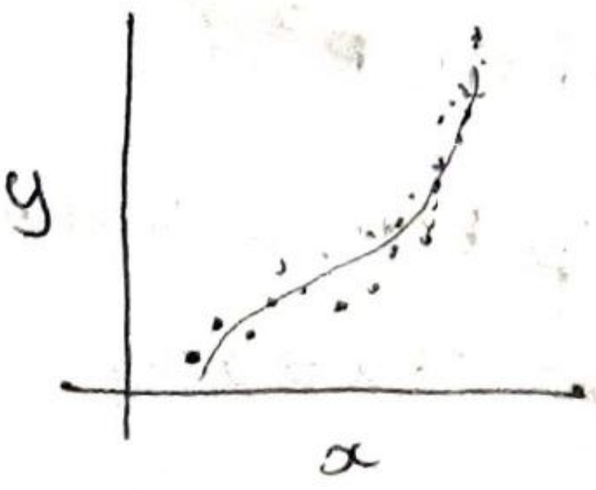
* Spearman rank correlation is the non-parametric version of Pearson correlation

* Spearman rank correlation determines the strength and direction of 2 Variables using rank Variables for x and y.

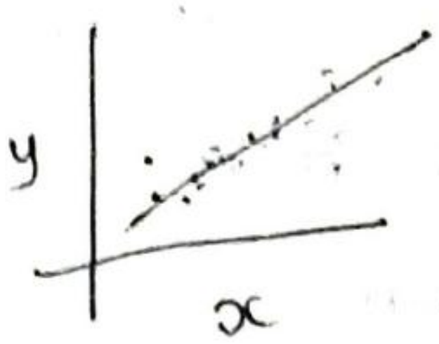
* It works for non-linear data (increasing or decreasing) and also for linear data.

* It works for monotonic relationship between the Variables.

* Monotonic relationship means the variables tend to move in same direction but the change will not be in constant rate.



→ • monotonic relation,
• both were increasing, but changing in different rate

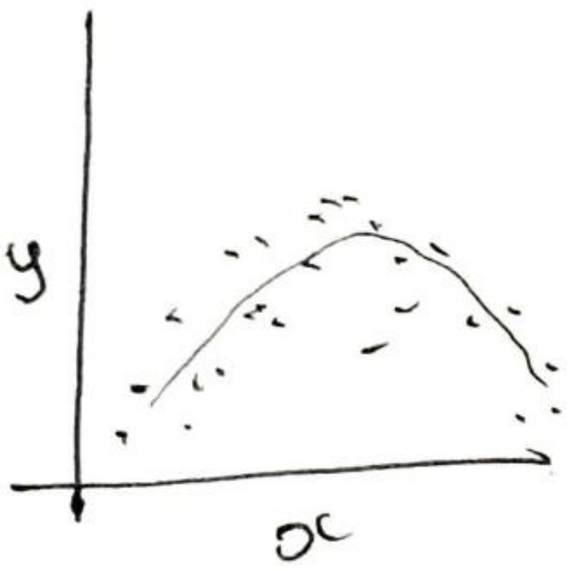


→
- Linear relation
- Both were increasing
- changing in constant rate

* It works well for ordinal data, which automatically defines the ranked data.

* It is less sensitive to outliers

* If spearman coefficient of variables is close to 0, it means there is no monotonic relationship between them.



→ non-monotonic relation

* Formula

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$ (If no tied ranks)

d_i is difference b/w r_x and r_y.

For tied ranks

$$\rho = \frac{\sum_{i=1}^{n} (rx_i - \bar{r}x)(ry_i - \bar{r}y)}{\sqrt{\sum_{i=1}^{n} (rx_i - \bar{r}x)^2 \sum_{i=1}^{n} (ry_i - \bar{r}y)^2}}$$

Hypothesis for spearman correlation

null hypothesis: x and y does not have monotonic relation

Alternate hypothesis: x and y have monotonic relation.

So, if P-value is less then level of significance, then null hypothesis is rejected. And it is ~~shown~~ statistically significant.