5. CORRELATION

THERE ARE TWO TYPES OF DISTRIBUTIONS:

1) Univariate distribution 2) Bivariate Distribution

UNIVARIATE DISTRIBUTION:

1) A distribution involves only one variable.

BIVARIATE DISTRIBUTION:

1) A Distribution involving two variables.
2) For (m x n) bivariate frequency distribution,
   a) No. of Marginal Distributions = 2
   b) No. of Conditional Distributions = m + n
   c) No. of Cell Frequencies = m n

CORRELATION:

a) Correlation studies the relationship between two variables.

b) It can determine the degree of association or relationship between two or more variables.

TYPES OF CORRELATION:

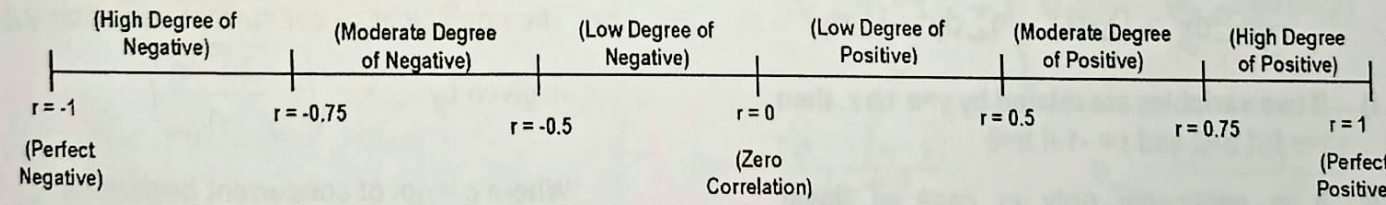
1) Positive Correlation (Direct): If two variables move in same direction, then correlation said to be positive.
2) Negative Correlation (Inverse): If two variables move in opposite direction then correlation said to be negative.
3) Spurious Correlation: The correlation between two variables which mathematically proved and fail to prove logically.

EXAMPLES FOR DIFFERENT TYPES OF CORRELATION:

POSTITIVE CORRELATION	NEGATIVE CORRELATION
Amount of rainfall and yield of crop'	Price and Demand
Sales of cold drinks and day temperature'	Sales of woolen garments and day temperature
Income and expenditure	Volume and pressure
Employment and purchasing power'	Unemployment and purchasing power
Age of applicants and the premium of insurance	Insurance companies profits and the no of claims

CORRELATION COEFFICIENT: It is the numerical value of correlation. It is generally denoted by r.

1) 'r' value lies between -1 and +1
2) If $r = +1$ there is perfect positive correlation and $r = -1$ there is perfect negative correlation.
3) If $r = 0$ there is no correlation and variables are independent.
4) r is independent of the units of measurement, and it is a pure number.
5) r is independent of both change of origin and scale.



MEASURES OF CORRELATION:

a) Scatter diagram
b) Karl Pearson's product moment correlation coefficient
c) Spearman's rank correlation coefficient
d) Coefficient of concurrent deviation.

1) Scatter Diagram:
   a) It gives the nature of correlation and distinguishes different types of correlation.
   b) It fails to measure the numerical value of correlation.
   c) It can be applied for linear and non-linear correlation
   d) If the points are in upward direction, i.e., Lower left to upper right ,it shows positive correlation.
   e) If the points are in downward direction i.e., Upper right to lower right , it shows negative correlation.

f) If the points are evenly distributed, then no correlation between variables.

g) If the plotted points very close to each other, it shows high correlation otherwise poor correlation.

2) Karl Pearson's Correlation Coefficient:

a) Covariance measures common or joint variation between variables.

b) $\text{Cov}(x, y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \frac{\Sigma xy}{n} - \bar{x}.\bar{y}$

c) Covariance is zero when variables are Independent.

d) $\text{Cov}(x\,y) \leq \sigma_x.\sigma_y$

e) Using Covariance: $r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y}$

f) Deviations from mean, $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

g) No deviations,

$$r = \frac{n\Sigma xy - \Sigma x.\Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

h) Deviations from assumed mean,

$$r = \frac{n\Sigma dx.dy - \Sigma dx.\Sigma dy}{\sqrt{n\Sigma dx^2 - (\Sigma dx)^2}\sqrt{n\Sigma dy^2 - (\Sigma dy)^2}}$$

i) If two variables are related by y=a +b x, then r=+1 if b>0 and r = -1 if b<0

j) It is applicable only in case of linear relationship between the two variables

k) It is independent of changes in origin and scale, but the direction of the correlation will be depending on the signs of scale factors.

l) It holds symmetry property i.e., $r_{xy} = r_{yx}$.

3) SPEARMAN'S RANK CORRELATION:

a) It is used for qualitative characteristics also.

b) Spearman's rank correlation coefficient (R) is given by

When ranks are not repeated

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

When ranks are repeated

$$R = 1 - \frac{6\left[\Sigma d^2 + \Sigma\frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

c) When repeated ranks are given , the correction factor is $\frac{m(m^2-1)}{12}$, (where m is rank frequency)

m:	2	3	4
$m\left(\frac{m^2-1}{12}\right)$:	0.5	2	5

n:	10	9	8	7	6
$n(n^2-1)$:	990	720	504	336	210

d) Sum of difference of ranks is always zero i.e., $\Sigma d = 0$

e) If the ranks are given in same order, then $\Sigma d^2 = 0$ and $r = +1$.

f) If the ranks are given in reverse order, then rank of x + Rank of y = n + 1 and

$r = -1$.

4) CONCURRENT DEVIATION METHOD:

a) The coefficient of concurrent deviation (r_c) is given by $r_c = \pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$

Where c = no. of concurrent deviations

n = no. of pairs of observations – 1

b) If (2c-n) > 0, then we take the positive sign both inside and outside the radical sign.

c) If (2c-n) < 0 then we take the negative sign both inside and outside the radical sign.

d) When c = 0 then r = -1

e) When c = n then r = +1

f) When $c = \frac{n}{2}$ then r = 0.

PROBABLE ERROR:

1) It is used to study the reliability and significance of r.

2) $P.E.(r) = 0.6745.\frac{1-r^2}{\sqrt{n}} \Rightarrow P.E.(r) = 0.6745\,SE(r)$

3) $P.E.(r) = \frac{2}{3}SE(r)$, where $S.E.(r) = \frac{1-r^2}{\sqrt{n}}$

4) If $r < 6$ P.E, then r is insignificant.

5) If $r \geq 6$ P.E, then r is significant

6) Limits of population correlation coefficient are $r \pm P.E.$

r	Correlation coefficient
r^2	Coefficient of determination (Explained variance or accounted variation)
$1-r^2$	Coefficient of non- determination (Unexplained variance and unaccounted variation)
$\sqrt{1-r^2}$	Coefficient of alienation