Module-4

Correlation

In statistics we have two types of distribution of data:- (I) Univariate (II) Multivariate

Univariate Population or Univariate Distribution
:→ A Population Contains only one variable. The distribution of data for which so far we have obtained the Mean, Median, Mode, Moments, Skewness, Kurtosis etc. were the Univariate population.

Multivariate population or Multivariate Distribution
:→ A population contains more than one variable. If it Contains two variable exactly then it is known as Bivariate population.

Now we have to study Bivariate Population or Distribution. In a given data distribution the two variables are related to each other then we say there is a correlation between these two independent variables otherwise we say these two variables are independent or Uncorrelated.

In formal way we can define the correlation as follows:→

If the change in one variable affects the value of another variable in a bivariate distribution then we say there is a Correlation between them.

Obviously there may be two types of these changes.

(I) Positive correlation :→ If an increment in first variable increases the value of second variable then correlation is said to positive.

or

If a decrement in first variable decreases the value of second variable then correlation is said positive.

(II) Negative correlation :→ If an increment in first variable decreases the value of second variable then correlation is said to Negative.

or

If a decrement in first variable increases the value of second variable then correlation is said Negative.

The mathematical formulation of correlation is given by Karl Pearson and it is known as Pearson's Coefficient of Correlation or Sometimes it is known as correlation coefficient only. It is denoted by r generally. Where

$$\boxed{r = \frac{\Sigma XY}{n\sigma_x \sigma_y}}$$, where $X = x - \bar{x}$, $Y = y - \bar{y}$ | $\bar{x}$ and $\bar{y}$ are the means of x-series and y-series respectively

And σ_x & σ_y are the standard deviations of x-series and y series.

Date : / /

This can be simplified further also.

$$r = \frac{\Sigma XY}{n\sigma_x \sigma_y}$$

$$= \frac{\Sigma XY}{n \cdot \sqrt{\frac{\Sigma X^2}{n}} \cdot \sqrt{\frac{\Sigma Y^2}{n}}}$$

$$= \frac{\Sigma XY}{\cancel{n}\frac{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}{\cancel{n}}}$$

We know $\sigma_x = \sqrt{\frac{\Sigma f_i (x_i - \bar{x})^2}{n}}$

Here all the frequencies are 1 so $f_i = 1$

$\therefore \sigma_x = \sqrt{\frac{\Sigma X^2}{n}}$

Similarly, $\sigma_y = \sqrt{\frac{\Sigma Y^2}{n}}$.

$$\boxed{\therefore r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}}$$

We can use both of the above formulae to find Coefficient of Correlation.

If we simplify it further again then we get

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \Sigma (y-\bar{y})^2}}$$

$$= \frac{\Sigma (xy - \bar{x}y - \bar{y}x + \bar{x}\bar{y})}{\sqrt{\Sigma (x^2 - 2\bar{x}x + \bar{x}^2) \Sigma (y^2 - 2\bar{y}y + \bar{y}^2)}}$$

Numerator is $\Sigma xy - \bar{x}\Sigma y - \bar{y}\Sigma x + \bar{x}\bar{y}\Sigma 1$

$= \Sigma xy - \frac{\Sigma x}{n}\Sigma y - \frac{\Sigma y}{n}\Sigma x + \frac{\Sigma x}{n} \cdot \frac{\Sigma y}{n} \cdot n$ | $\sum_{i=1}^{n} 1 = n$

Scanned by CamScanner

Date :

$$= \sum xy - \frac{\sum x \sum y}{n} - \cancel{\frac{\sum x \sum y}{n}} + \cancel{\frac{\sum x \sum y}{n}}$$

$$= \frac{n\sum xy - \sum x \sum y}{n}$$

And denominator is $\sum (x^2 - 2\bar{x}x + \bar{x}^2)$

$$= \sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum 1$$

$$= \sum x^2 - 2\frac{\sum x}{n}.\sum x + \left(\frac{\sum x}{n}\right)^2 . n$$

$$= \sum x^2 - \frac{2(\sum x)^2}{n} + \frac{(\sum x)^2}{n}$$

$$= \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= \frac{n\sum x^2 - (\sum x)^2}{n}$$

$$\therefore r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

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

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

This formula to find correlation is useful when we don't find the means and the standard deviations for the given distribution.

And lastly if we simplify it further again then we find

$$\boxed{r = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\{n\sum d_x^2 - (\sum d_x)^2\}\{n\sum d_y^2 - (\sum d_y)^2\}}}}$$

where $d_x = \frac{x-A}{h}$ and $d_y = \frac{y-B}{K}$

with A and B are assumed means and h & K are arbitrary numbers related to data.

This formula is useful when the given data is complex i.e. complicated, not simple.

Magnitude of the correlation coefficient always lies between -1 and 1. i.e. $|r| \leq 1$. i.e. $-1 \leq r \leq 1$.