Paper : MTHM – 601 (Statistics)
UNIT – III : Correlation and Regression

1. What is bi-variate data ?

Solution : In case of univariate data, the data are collected by studying only one characteristic, as for example, the mean height of students in a college, each observation represents the height of each student. But, in case of bi-variate or multi-variate data, the data are collected more than one characteristics under the study, as for example, when the students of a college are considered with respect to height and weight. When the heights of some students are same then weight may be different or, when the weights of another group of students are same then height may be different. For bivariate data, we consider (x_i, y_i) where $i = 1,2,3, \ldots, n$ and the frequency distribution is called bivariate frequency distribution.

2. What is correlation ?

Solution : When there exists a relation between two variables in such a way, if there is a change in the value of one variable then there corresponds a change in the value of the other variable, this relation between the variables is known as correlation.

The correlation is studied between two variables is called simple correlation.

The correlation is studied between more than two variables is called multiple correlation.

3.Write the types of correlation with example.

Solution : On the basis of nature of relationship between two variables, correlation is divided into the following categories.

(i) Positive Correlation : If increase (or decrease) in one variable results a corresponding increase (or decrease) in the other variable, i.e., two variables move in the same direction, variables are said to be positively correlated, as for example, heights and weights of children, price and supply of a set of commodities etc., income and expenditure of a group of families, etc.

(ii) Negative Correlation : If increase (or decrease) in one variable results a corresponding decrease (or increase) in the other variable, i.e., two variables deviate in the opposite direction, variables are said to be negatively correlated, as for example, price and demand of a set of commodities, speed and time, etc.

(iii) Perfect Positive Correlation : When changes in two related variables are exactly proportional there is perfect correlation between them.

If increase (or decrease) in one variable results a corresponding and proportional increase (or decrease) in the other variable, variables are said to be perfect positively correlated, as for example, correlation between radius and circumference of a circle, temperature and volume, etc.

(iv) Perfect Negative Correlation : If increase (or decrease) in one variable results a corresponding and proportional decrease (or increase) in the other variable, variables are said to be perfect negatively correlated, as for example, pressure and volume, etc.

(v) Zero Correlation : If the change in one variable does not result any change in the other variable, variables are said to be uncorrelated or independent or having zero correlation, as for example, price of commodities and weight of individuals, etc.

4. What is Karl Pearson’s Co-efficient ? Derive its result from two variables.

Solution : The statistician Karl Pearson presented a co-efficient to measure the degree of linear relationship between two variables which is known as Karl Pearson’s co-efficient of correlation (or product-moment correlation co-efficient). It is denoted by the symbol ‘r’. The correlation co-efficient between two variables x and y is

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}\frac{1}{n}\sum_{i=1}^{n} y_i + \frac{1}{n}n\bar{x}.\bar{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

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

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\,\bar{x} + \bar{x}^2$$

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

Similarly, $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}$

Thus, r_{xy} can also be written in the following forms

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

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

$$r_{xy}=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

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

when actual mean in decimal, the calculations become very tedius and in such cases we may take help of the following formula.

$$r=\frac{n\sum uv-(\sum u)(\sum v)}{\sqrt{n\sum u^2-(\sum u)^2}\sqrt{n\sum v^2-(\sum v)^2}}\ ;\ u=x-A, v=y-B$$

where, A and B are assumed means.