Correlation & Regression

1. Definitions

1.1 Univariate Distribution: These are the distributions in which there is only one variable is involved.

For example

(i) The height of students in a class.

(ii) The marks obtained by students in a class.

1.2 Bivariate Distribution: Distributions involving two discrete variables is called a bivariate distribution

For example

(1) The heights and weights of the students for a class in a school.

(2) The marks obtained by (2) The marks obtained by students of a class in two subjects.

1.3 Bivariate frequency distribution: This is a distribution in which two variables are involved. Let x and y be two variables. Suppose x takes the values x_1, x_2,.....,x_n and y takes the values y_1, y_2,.....y_n then we record our observations in the form of ordered pairs (x_i, y_j), where $1 \le i \le n$, $1 \le j \le n$, If a certain pair occurs f_{ij} times, we say that its frequency is f_{ij}.

The function which assigns the frequencies f_{ij}'s to the pairs (x_i, y_j) is known as a **bivariate frequency distribution.**

Two way frequency tables : In such tables, the top row consists of the values of the variable x and the left hand column consists of the values of the y. The frequencies corresponding to a pair of values are written in the cell at the intersection of the relevant row and column.

The column total provide the univariate frequency distribution of x and the row totals provide the univariate frequency distribution of y. These column totals and the row totals are known as the **marginal frequency distributions** of x and y respectively.

Conditional frequency and frequency distribution : From the bivariate frequency distribution we can study relationship of two variables and their degree. Once we study the degree of relationship we can estimate the value of one variable while the value of the other variable is given for some fixed values of x the frequencies with which the various y values occur, when listed gives the conditional frequency distributions of y on x.

Similarly for some fixed values of y, the frequencies with which various x values occur, when listed gives the conditional frequency distribution of x on y.

2. Co-variance

Before we study correlation, let us introduce the concept of covariance between two quantitative variables.

Definition : Covariance is the arithmetic mean of the products of the corresponding deviations of two series from their respective means.

Let two variables x and y takes the values x_1, x_2, x_3.....x_n and y_1, y_2, y_3 .....y_n then covariance is defined as-

$$\text{Cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x and y series respectively.

3. Correlation

The relationship between two variables such that a change in one variable results in a positive or negative change in the other variable is known as **correlation.**

3.1 Types of correlation :

(i) Perfect Correlation : If the two variables vary in such a manner that their ratio is always constant, then the correlation is said to be perfect.

(ii) Positive or direct correlation : If an increase or decrease in one variable corresponds to an increase or decrease in the other, the correlation is said to be positive.

For example

Income and expenditure are positively (or directly) correlated because expenditure increases as income increases. Expenses are curtailed with the decrease in income.

(iii) Negative or indirect Correlation : If an increase or decrease in one variable corresponds to a decrease or increase in the other, the correlation is said to be negative

For example

Pressure and volume of a gas are negatively (or inversely) correlated because with the increase of pressure on a gas there is decrease in volume & viceversa.

(iv) Zero Correlation : If the variation in one has no relation with that in the other then the variable have no correlation or there is zero correlation between the variables.

4. Coefficient of correlation

Karl Pearson gave the following formula for the calculation of correlation coefficient between two variables x and y-

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

where x, y, $\bar{x}$ and $\bar{y}$ have usual meanings

$$\text{or} \quad r_{xy} = r = \frac{\sum dx\ dy}{\sqrt{\sum dx^2 \times \sum dy^2}}$$

where $dx = (x - \bar{x})$, $dx^2 = (x - \bar{x})^2$
$dy = (y - \bar{y})$ and $dy^2 = (y - \bar{y})^2$

Modified formula :

$$r = \frac{\sum dx\,dy - \frac{\sum dx.dy}{n}}{\sqrt{\left\{\sum dx^2 - \frac{(\sum dx)^2}{n}\right\}\left\{\sum dy^2 - \frac{(\sum dy)^2}{n}\right\}}}$$

$$\text{Also} \quad r_{xy} = \frac{Cov.(x,y)}{\sigma_x\sigma_y} = \frac{Cov.(x,y)}{\sqrt{Var(x).Var(y)}}$$

5. Rank Correlation

Rank correlation is the correlation between different ranks or grades of the two characteristics

It is given by $1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Here $d^2 = \sum_{i=1}^{n}\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$

where Σd^2 = sum of the squares of the difference of two ranks and n is the number of pairs of observations.

6. Properties of Correlation coefficient (r)

(a) r lies between – 1 and + 1
(b) the correlation is
(i) perfect and positive if r = + 1
(ii) perfect and negative if r = – 1
(iii) not correlated if r = 0
(iv) positive if r > 0
(v) negative if r < 0
(c) It is independent of the change of origin and scale.
(d) It is a pure number and hence unitless
(e) If x and y are independent then r = 0

7. Regression Analysis

In the previous article we have seen that correlation is merely a tool of ascertaining the degree of relationship between two variables. If does not tell any thing about the functional relationship or nature of relationship between two variables but regression analysis attempts to study the functional relationship between the variables so that one can predict the value of one variable for the given value of the other variable, so

Regression analysis is a statistical device with the help of which we can estimate or predict the unknown values of one variable from the known values of the other variable.

8. Line of Regression

The regression line is a graphical method, which describes the average of relationship between the two variables. Let us take the case of two variable x and y we shall have two lines of regression because there are two variable.

(i) Y on X (ii) X on Y

8.1 Line of regression of y on x :

The line of regression of y on x gives most probable values of y for given values of x and so it is used to estimate y for any given value of x. Its equation is -

$$y - \bar{y} = \frac{cov.x,y}{\sigma_x^2}(x - \bar{x})$$

$$\text{or} \quad y - \bar{y} = r.\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

8.2 Line of regression of x on y:

The line of regression of x on y gives the most probable values of x for given values of y and so it is used to estimate the value of x for given value of y. Its equation is-

$$x - \bar{x} = \frac{cov.(x,y)}{\sigma_y^2}(y - \bar{y})$$

$$x - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}\,(y - \bar{y})$$

9. Regression Coefficient

(i) The regression coefficient of y on x is denoted by b_{yx} and is given by

$$b_{yx} - r.\,\frac{\sigma_y}{\sigma_x} = \frac{cov.(x,y)}{\sigma_x^2}$$

This represents the change in the values of y corresponding to a unit change in x.

(ii) The coefficient of regression of x on y is denoted by b_{xy} and is given by

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{cov.(x,y)}{\sigma_y^2}$$

This represents the change in the value of x corresponding to a unit change in y.

10. Properties of regression coefficients

(i) $r = \sqrt{b_{yx}.b_{xy}}$ i.e. the coefficient of correlation is the geometric mean between the two regression coefficients.

(ii) If byx > 1, then bxy < 1, i.e. If one of the regression coefficient is greater then unity then the other will be less than unity.

(iii) If the correlation between the variables is not perfect then the regression lines intersect at $(\bar{x}, \bar{y})$

(iv) b_{yx} is called the slope of regression line y on x & b_{xy} is called the slope of regression line x on y.

(v) $b_{yx} + b_{xy} > 2\sqrt{b_{yx}.b_{xy}}$ or $b_{yx} + b_{xy} > 2r$ i.e the arithmetic mean of the regression coefficient is greater than the correlation coefficient.

(vi) Regression coefficients are independent of change of origin but not of scale.

(vii) The product of lines of regression's gradients is given by $\frac{\sigma_y^2}{\sigma_x^2}$.

(viii) If the angle between lines of regression is θ then

$$\tan\theta = \left(\frac{1-r^2}{r}\right).\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

(ix) If both the lines of regression coincide, then correlation will be perfect linear

(x) If both b_{yx} & b_{xy} are positive, then r will be positive and if both b_{yx} & b_{xy} are negative then r will be negative.

11. Important points on regression lines

(I) If r = 0, then tan θ is not defined i.e. $\theta = \frac{\pi}{2}$.

Thus If two variables are not correlated, then the lines of regression are perpendicular to each other.

(ii) If r = ±1, then tan θ = 0 i.e. θ = 0 Thus the regression lines are coincident .

(iii) If regression lines are y = ax + b & x = cy + d then $\bar{x} = \frac{bc+d}{1-ac}$ and $\bar{y} = \frac{ad+b}{1-ac}$

12. Standard error of prediction

The deviation of the predicted value from the observed value is known as the **standard error of prediction** and is defined as

$$S_y = \sqrt{\left\{\frac{\Sigma(y-y_p)^2}{n}\right\}}$$

where y is actual value and y_p is predicted value.

In relation to coefficient of correlation, it is given by -

(i) Standard error of estimate of x is

$$S_x = \sigma_x\sqrt{1-r^2}$$

(ii) Standard error of estimate of y is

$$S_y = \sigma_y\sqrt{1-r^2}$$