2020 - 2021

M.A.L.D GOVT ARTS AND SCIENCE COLLEGE - GADWAL

DEPARTMENT OF STATISTICS

PROJECT WORK

STUDENTS SOCIO ECONOMIC CONDITIONS

Certified that the project work has been submitted by the student R. Akhila Regd. Number 19033024467010 during the year

Head

Department of statistics

CONTENTS

OBJECTIVE OF THE PROJECTIVE WORK

The main objective of the project work is to analyses statistical data and use the statistical tools for finding the about the students marks of FIRST YEAR B.Sc students of statistics Dept. AND also find the CORRELATION between Marks in statistics and Mathematics.

Tools and Techniques

CORRELATION AND REGRESSION

Correlation :-

Defination :-

In a bivariate distribution, if the change in one variable affects a change in the other variable, the variables are said to be correlated.

The correlation may be classified in to the following heads.

1. positive correlation :-

In a bivariate distribution, if the two variables are deviated in the same direction, then the two variables are (may be carried as may be) said to be positively correlated.

2. Negative correlation :-

In a bivariate distribution, if the two variables are deviated in the opposite direction then the two variables may be said to be -vely correlated.

Scatter diagram :-

It is the simplest way of the diagramatic representation of bivariate data. Thus for the bivariate distribution $(x_i, y_i;$ $\forall i = 1, 2, ... n)$, if the values of the variables x axis and y axis be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of data so obtained is known as scatter diagram.

METHODS OF CORRELATION

karlpearson's coefficient of correlation :-

As a measure of intensity of linear relationship between two variables, karl pearson (1867 - 1936) a British biometrician developed a formula called correlation coefficient.

correlation coefficient between two random variables X-axis and Y-axis usually denoted by $r(X,Y)$ (or) r_{xy} or r as defined as

$$r = \frac{cov(X,Y)}{\sqrt{u(x)}\ \sqrt{v(y)}}$$

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

Where $cov(x,y) = E[x - E(x)]\ E[y - E(y)] \rightarrow ①$

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

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

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

$$= \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \rightarrow 3$$

$$v(x) = E[x - E(x)]^2 \rightarrow ①$$

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

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

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

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

$$= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \rightarrow 3$$

III'ly $V(y) = E[y - E(y)]^2 \rightarrow 1$

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

$$V(y) = \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \bar{y} \rightarrow 3$$

Calculation of the correlation coefficient for a bi-variate frequency distribution

When the data are considerably large, they may be summarized by using a two way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariable case. If there are m claised for x and n classes for y, there will be in all m×n cells in the two ways table. The whole set of sell frequencies will then defined a bi-variate frequency distribution

Y/X	x_1	x_2 -------	x_i ------	x_m	Total
y_1	f_{11}	f_{21}	f_{i1}	f_{m1} ----	$f_{.2}$
y_2	f_{12}	f_{22} -------	f_{i2}	f_{m2}	f
y_j	f_{ij}	f_{2j} ------	f_{ij}	f_{mj} - ..	f_{nj}
y_n	f_m	f_{2n}	f_{in}	f_{nj}	f_n
Total	f_i	f_2 ------	f_{in}	f_{mn} ...	$f=N$

in the above table $\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} i.j = f.. = N$

Here $f_{1.}, f_2 \ldots f_i \ldots f_m$ are called the mariginal frequencies of the variable x and $f_1, f_2, \ldots f_j \ldots f_n$ are called marginal frequencies of the variable y' and f_{ij} is the frequency of $(ij)^{th}$ cell where.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{n} \times x_i$$

$$= \frac{1}{N}\sum_{i=1}^{m} f_i \cdot x_i$$

$$\bar{y} = \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, x_j$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_j\, y_j$$

$$\sigma^2 x = V(x) = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, x_i^2 - \bar{x}^2$$

$$= \frac{1}{N}\sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2$$

$$\sigma^2 y = V(y) = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \cdot y_j^2 - (\bar{y})^2$$

$$= \frac{1}{N}\sum_{i=1}^{m} f_j\, y_j^2 - \bar{y}^2$$

$$\mathrm{covu}(x) = \frac{1}{N}\sum_{i=1}^{m}\sum_{y=j} f_{ij}\, x_i y_j - \bar{x}\,\bar{y}$$

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{m}\sum_{y=j}^{n} f_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{N}\sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2}\ \sqrt{\frac{1}{N}\sum_{j=1}^{n} f_j y_j - \bar{y}^2}}$$

$$= \frac{\frac{1}{4}\left[\sum_{i=1}^{m}\sum_{y=j}^{n} f_{ij}\, x_i y_j - \left(\sum_{i=1}^{m} f_i \cdot x_i\right)\left(\sum_{j=1}^{n} f_j\, y_j\right)\right]}{\frac{1}{N}\left[\sqrt{\sum_{i=1}^{m} f_i\, x_i^2 - \left[\frac{\sum_{i=1}^{m} f_i \cdot x_i^2}{N}\right]}\ \sqrt{\sum_{j=1}^{n} f_j\, y_j^2 - \left(\frac{\sum_{j=1}^{n} f_j\, y_j}{N}\right)^2}\right]}$$

$$r_{xy} = \frac{\sum_{i=1}^{m}\sum_{y=1}^{n} f_{ij}\, x_i\, y_j - \frac{\left[\sum_{i=1}^{m} f_i x_i\right]\left[\sum_{j=1}^{n} f_j x_j\right]}{N}}{\sqrt{\sum_{i=1}^{m} f_i x_i^2 - \left[\frac{\sum_{i=1}^{m} f_i x_i^2}{N}\right]}\;\sqrt{\sum_{j=1}^{n} f_j y_j^2 - \sum_{j=1}^{m} f_j \cdot y_j^2 - \frac{\left[\sum_{j=1}^{n} f_{ij} y_j\right]^2}{N}}}$$

STATISTICAL DATA

Marks in Mathematics :-

92, 80, 65, 120, 97, 85, 91, 70, 77, 96, 129, 90, 113, 142, 78, 95, 101, 128, 123, 1, 39,

106, 112, 107, 100, 146, 126, 93, 98, 82, 105, 82, 85, 147, 121, 104, 125, 101, 96, 146, 115, 115, 129, 114, 78, 134, 136, 81, 85, 143, 113, 118, 117, 136, 113, 143, 150.

Marks in statistics :-

78, 89, 80, 85, 91, 101, 108, 95, 99, 98, 103, 85, 107, 95, 89, 92, 92, 135, 138, 147, 122, 89, 118, 127, 150, 110, 110, 130, 92, 126, 87, 78, 100, 92, 107, 100, 106, 94, 143,

107, 134, 80, 107, 90, 129, 121, 113, 129, 143, 94, 93, 111, 112, 86, 117, 134.

The above data can be converted into bivariate frequency data.

Bi variate data :

	71-80	81-90	91-100	101-110	111-120	121-130	131-140	141-150
51-70	1	-	1	-	-	-	-	-
71-90	1	5	2	1	1	1	-	-
91-110	1	-	5	4	1	4	-	-
111-130	1	3	4	5	1	-	3	-
131-150	-	-	2	-	2	2	1	4

Statistical Analysis :-

* Calculations for correlation bivariate table:

dx/dy	-4	-3	-2	-1	0	1	2	3	Total	fdx	$f(dx)^2$	fdxdy
-2	1	–	1	–	–	–	–	–	2	-4	8	12
-1	1	5	2	1	1	1	–	–	10	-11	11	23
0	1	–	5	4	1	4	–	–	15	0	0	0
1	1	3	4	5	1	–	3	–	17	17	17	-20
2	–	–	2	–	2	2	1	4	11	22	44	24
Total	4	8	14	10	5	7	4	4	56	24	80	39
fdy	-16	-24	-28	-10	0	7	8	12	-51			
fdy^2	64	72	56	10	0	7	16	36	261			
fdx.dy	8	6	-8	-4	0	3	10	24	39			

CONCLUSION :-

The correlation coefficient is $r = 0.4975$

Result :-

There is a positive correlation betwee[n] marks in mathematics and statistics.