CORRELATION AND REGRESSION ANALYSIS

1. What is Correlation?

If two quantities vary in such a way that movements in one are accompanied by movements in the other, these quantities are correlated. For example, there exist some relationship between price of commodity and amount demanded.

The measure of correlation called the correlation coefficient. A correlation coefficient is a numerical measures of some type of correlation.

According to Ken Black, "Correlation is a measure of degree of relatedness of variables."

According to Lathrop, "correlation is a joint relationship between two variables."

2. The value of correlation coefficient of correlation.

The value of the coefficient of correlation shall always lie between ± 1. When $r = +1$, it means there is perfect positive correlation between the variables. When $r = -1$, it means there is perfect negative correlation between the variables.

When $r = 0$, it means there is no relationship between the variables.

3. Unvariate distribution.

If the distributions involving only one variable, called unvariate distribution.

4. Bivariate distribution.

The frequency distribution of two variables is known as bivariate distribution.

Example:

Price	quantity sales
100	1
90	3
40	5
10	10

5. Multivariate distribution?

If the distribution involve three variable, called multivariate distribution.

6. Types of Correlation.

There are three types of correlation, they are as follows:-

(I) Positive or Negative correlation.

(II) Simple, partial and Multiple

(III) Linear and Non Linear correlation.

(I) Positive or Negative correlation:

Positive correlation: Correlation is said to be positive when the two variable moves in same direction, i.e., if one variable is increase, then other variable also increase and vice-versa.

Forexample: X : 10 12 15 18 20 and X : 80 70 60 40

Y : 15 20 22 25 37 Y : 50 45 40 30

Negative correlation: If both the variables are varying in the oposit direction, i.e., if one variable is increasing then the other variable is decreasing and vice-versa, correlation is said to be negative.

Forexample: X : 20 30 40 50 and X : 100 90 80 70

Y : 40 30 25 15 Y 10 30 40 50

(ii) Simple, partial and Multiple correlation:

Simple correlation: When only two variables are studied, it is a simple problem of correlation.

example: x : 20 30 40 50

y : 10 15 20 25

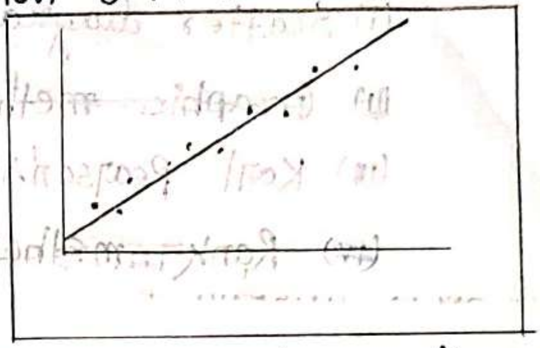
Partial correlation: In partial correlation, there more than two variable, but consider only two variable to be influencing each other, the effect of other variable being kept constant.

Multiple Correlation: When there are more variables are studied, it is a problem of either multiple or partial correlation. In multiple correlation there are more variables are studied simultaneously.

Forexample, when we study the relationship between the yield of rice per acre and both the amount of rainfall and fertilizer, it is a problem of multiple correlation.

(III) Linear and Non-Linear correlation:

Linear Correlation: If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable, then correlation said to be linear.

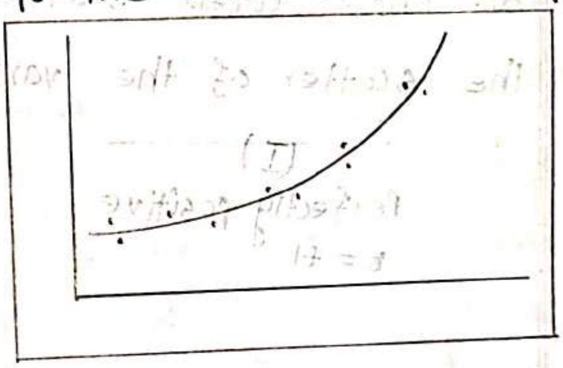


For example:

X: 10 20 30 40 50

Y: 70 140 210 280 350

Non-Linear Correlation: Correlation would be called non-linear or curvalinear if the amount of change in one variable does not bear constant ratio to the amount of change in the other variable.



For example:

X: 10 20 30 40 50

Y: 20 50 90 140 300

7. Method of Studying Correlation.

There are verious methods of studying correlation are as follows: —

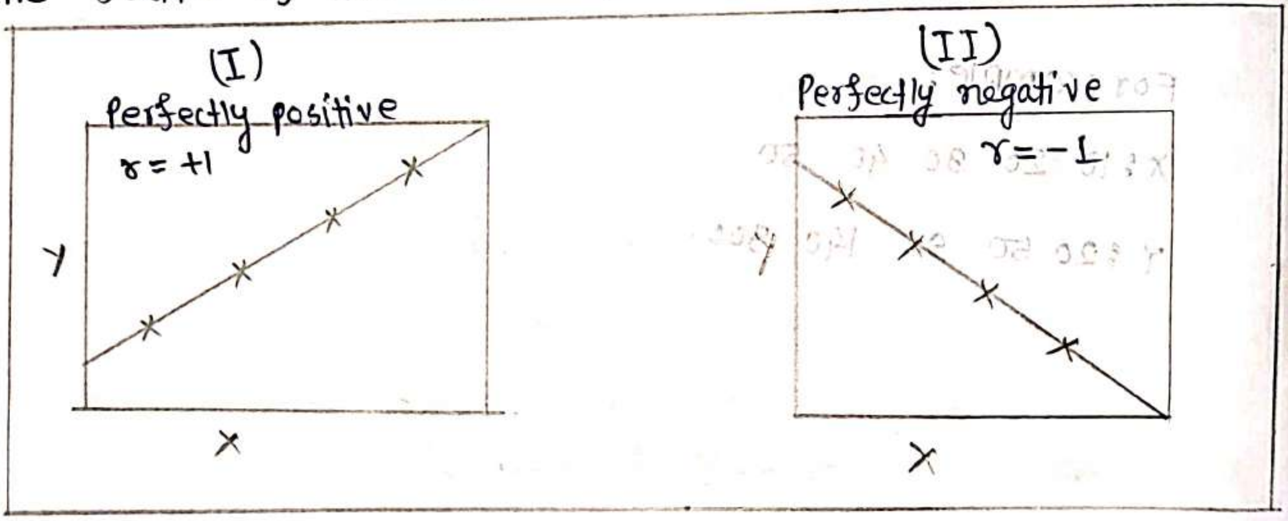
(I) Scatter diagram method. (graphical method)

(II) Karl Pearson's coefficient correlation,

(III) Spearman Rank correlation.

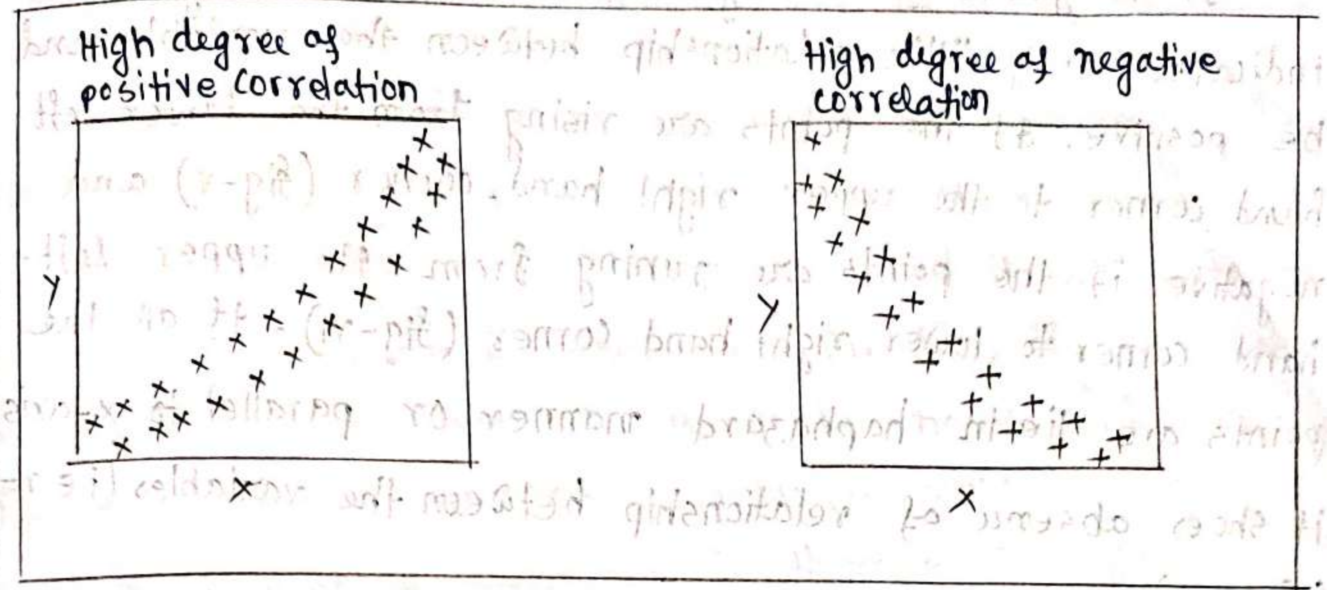
(I) Scatter Diagrame method:

When two variables are related is to prepare a dot chart called scatter diagram, because its indicates the scatter of the various points.



If all the points lie on a streight line rising from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (i.e, $r = +1$) (I fig)

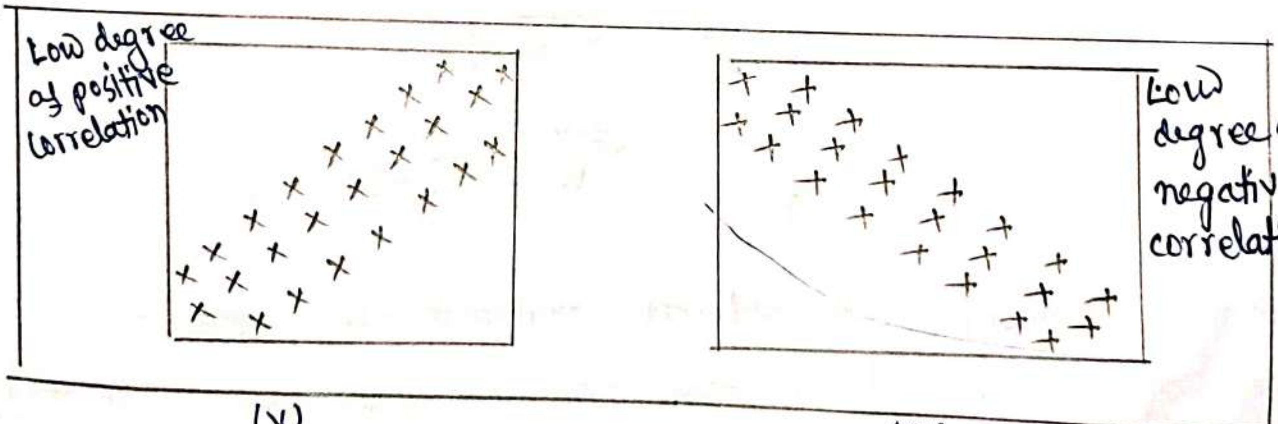
On the other hand, if all the points are lying on a straight line falling from the upper left-hand corner to lower right-hand corner, the correlation is said to be perfectly negative (i.e. $r = -1$) (fig - II)



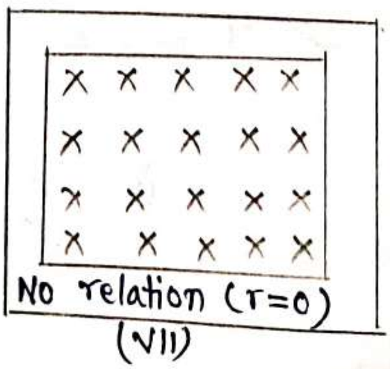
(III) (IV)

If all the plotted points fall in a narrow band and show a rising tendency from the lower left-hand corner to the upper right-hand corner, correlation is said to be high degree of positive correlation (fig: III).

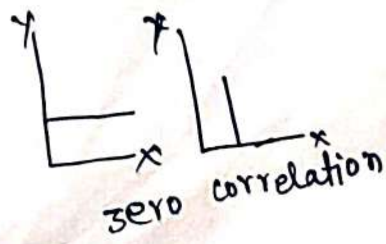
If the points shows a decreasing tendency from upper-left corner to the lower right-hand corner, shows the high degree of negative correlation (fig: IV).



(V) (VI)



(VII)



If the points are widely scattered over the diagram, it indicates very little relationship between the variables and be positive. If the points are rising from the lower left-hand corner to the upper right hand corner (fig-v) and negative if the points are runing from the upper left-hand corner to lower right hand corner (fig-vi). If all the points are lie in haphazard manner or parallel to x-axis, it shoes absence of relationship between the variables (i.e. $r=0$) (fig-vii).

Merits:

(i) It is a simple and non-mathematical method of studying correlation between the variables.

(ii) It is not influenced by the size of extreme items.

(ii) Karl Pearson's product moment correlation coefficient :

It is a numerical measure of assessing the nature and strength of correlation between two variables. It is developed by prof. Karl pearson and it is suitable for measuring correlation between two variables if the relationship between them is linear. It is to be noted that Karl pearson's product moment correlation coefficient can not be used for measuring correlation if there exist a linear relationship between the variables.

If x and Y are two non-random variables, then product moment correlation coefficient between them is denoted by r_{xy} or $r(x,y)$ or simply by r, and is defined by –

$$r_{xy} = \frac{b_{xy}}{b_x \cdot b_y}$$

where, $b_{xy} = \frac{1}{n}\Sigma(X-\bar{X})(Y-\bar{Y})$

$$b_x = \sqrt{1/n\,\Sigma(X-\bar{X})^2}$$

$$b_y = \sqrt{1/n\,\Sigma(Y-\bar{Y})^2}$$

* Pearson correlation is widely used in statistics to measure the degree of the relationship between linear related variables.

If x and Y are random variable then product moment correlation coefficient between them is denoted by r_{xy} or $r(x,y)$ or simply by r and is defined as –

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}$$

Here, $$Cov(X,Y) = E\left[\{X - E(X)\}\{Y - E(Y)\}\right]$$

$$V(X) = E\left[\{X - E(X)\}^2\right]$$

$$V(Y) = E\left[\{Y - E(Y)\}^2\right]$$

Note:

For calculating correlation coefficient in case of both random and non-random variables we can also use the following simplify formulas —

(1) $b_{xy} = \frac{1}{n}\Sigma XY - \bar{X}.\bar{Y}$

(2) $b_x = \sqrt{\frac{1}{n}\Sigma X^2 - (\bar{X})^2}$

(3) $b_y = \sqrt{\frac{1}{n}\Sigma Y^2 - (\bar{Y})^2}$

(4) $Cov(X,Y) = E(X,Y) - E(X) \cdot E(Y)$

(5) $V(X) = E(X^2) - \{E(x)\}^2$

(6) $V(Y) = E(Y^2) - \{E(Y)\}^2$

* Properties of correlation coefficient :

(i) Correlation coefficient is a pure number.

(ii) The coefficient of correlation lies between -1 and $+1$.

(iii) Correlation coefficient is indipendent of change of origin and of scale.

(iv) Two independent variables are uncorrelated.

(v) Correlation coefficient is symetric, i.e., $r_{xy} = r_{yx}$

* Merit of correlation coefficient.

(i) Exact measurment of correlation is possible.

(ii) It indicates the degree and direction of correlation.

* Limitation of correlation coefficient.

(i) Correlation coefficient cannot measure correlation between two variables if the variables are having a non-linear relationship.

(ii) Correlation coefficient cannot be used for finding correlation between qualitative variables.

(iii) Correlation coefficient cannot give us any idea about the cause-effect relationship between two variables.

Exercise

Q. Calculate correlation coefficient from the following data and interpret the result.

$n = 10$ $\Sigma X = 100$ $\Sigma Y = 80$ $\Sigma XY = 125$

$\Sigma X^2 = 120$ $\Sigma Y^2 = 114$

Soln

Given that,

$n = 10$, $\Sigma X = 100$; $\Sigma Y = 80$

$\Sigma XY = 125$, $\Sigma X^2 = 120$, $\Sigma Y^2 = 114$

We know that, $r_{XY} = \frac{cov(X, Y)}{b_X \cdot b_Y}$ $\frac{b_{XY}}{b_X \cdot b_Y}$

again, $\underline{cov(X, Y)}_{b_{XY}} = \frac{1}{n} \Sigma XY - \bar{X} \bar{Y}$

$= \frac{1}{10} \times 125 - \frac{\Sigma X}{n} \cdot \frac{\Sigma Y}{n}$

$= \frac{125}{10} - 10 \times 8$

$= \frac{125}{10} - 80$

$= 12.5 - 80$

$= -67.5$

$b_X = \sqrt{\frac{1}{n} \Sigma X^2 - (\bar{X})^2}$

$= \sqrt{\frac{1}{10} \times 120 - \left(\frac{100}{10}\right)^2}$

$= \sqrt{\frac{120}{10} - 10^2}$

$= \sqrt{12 - 100}$

$= \sqrt{-88}$

$= \sqrt{-1 \times 800} = \sqrt{i^2 \times 88}$

$= i\sqrt{88}$

$$by = \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{Y})^2}$$

$$= \sqrt{\frac{1}{10} \times 114 - \left(\frac{80}{10}\right)^2}$$

$$= \sqrt{\frac{114}{10} - \frac{80}{10}}$$

$$= \sqrt{14.4 - 64}$$

$$= \sqrt{6.4}$$

Again,

$$bxy = \frac{1}{n} \Sigma x y - \bar{X} \cdot \bar{Y}$$

$$= \frac{1}{10} \times 125 - 10 \times 8$$

$$= 12.5 - 80$$

$$=$$

$$\therefore rxy = \frac{bxy}{bx \cdot by}$$

$$= \frac{-67.5}{}$$

8. Rank correlation coefficient.

The correlation coefficient (r) is calculated on the basis of value of the variables. But quite often, situation arise in which the data under consideration are not capable of quantitative measurement but can be arranged in serial order.

Charles Edward Spearman (1904) developed a formula which helps in obtaining the correlation coefficient between the ranks of N individuals in two characterstics:-

(I) Non-tied rank (Non-repead Rank).

(II) Tied Rank (Repead Rank).

(I) Spearmen's Rank correlation coefficient formula:- in case of Non-tied Rank:

$$R/\rho = 1 - \frac{6\Sigma d_i^2}{n^3 - n} \quad , \quad -1 \leq R \leq 1$$

here,

$d_i = x_i - y_i$ (difference between the rank of two series).

n = Number of observation.

Q. Calculate Rank correlation for following data and interprite result.

x_i	5	3	2	1	4	6
y_i	2	3	1	5	6	4

Soln

Since there is no repetation of ranks, we use spearman's Rank correlation coefficient formula for non-tied rank.

$$\therefore R = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

For calculating rank correlation coefficient value, we construct the following table:

x_i	y_i	$d_i\ (x_i - y_i)$	d_i^2
5	2	3	9
3	3	0	0
2	1	1	1
1	5	-4	16
4	6	-2	4
6	4	2	4

$\Sigma d_i^2 = 34$

$$\therefore R = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

$$= 1 - \frac{6(34)}{6^3 - 6}$$

$$= 1 - \frac{204}{210}$$

$$= 1 - \cancel{0.0048}\ 0.971$$

$$= 0.029$$

Interpretation of result:

$R = 0.029$, means that there is a positive but very low association between two series of rank.

9. Calculate Rank correlation coefficient for the following data and interprete the ~~data~~. result

Mark in maths (X)	84	80	90	64	46	30
Mark in Statistics (Y)	73	48	35	80	36	39

Sol^n

In order to calulate rank correlation coefficient we construct the following table and ranking the marks :

X	Y	x_i	y_i	d_i	d_i^2
84	73	2	2	0	0
80	48	3	3	0	0
90	35	1	6	-5	25
64	80	4	1	3	9
46	36	5	5	0	0
30	39	6	4	2	4

$\Sigma d_i^2 = 38$

We know that,

$$R = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

$$= 1 - \frac{6(38)}{6^3 - 6}$$

$$= 1 - \frac{228}{210}$$

$$= 1 - 1.086$$

$$= -0.086$$

Since the value of correlation of coefficient ~~lies between~~ is $r = -0.086$, it means that there is a negative and very law association between two series of rank.

(ii) Spearman Rank correlation Coefficient – for tied Rank :

$$R = 1 - \frac{6\left\{\Sigma d_i^2 + \Sigma \frac{m^3 - m}{12}\right\}}{n^3 - n}$$

here, $d_i = x_i - y_i$

$\frac{m^3 - m}{12}$ = Adjustment factor where m denots number of times a value occurs for which we have a tie.

n = Number of observation.

Step:

(i) Assining common rank to the number which are repeated.

(ii) Addition of adjustment factor.

(iii) How many times have the adjustment factor will depend up on the number of type present in problem.

Excercise:

1. Given the data on the marks obtain by five students in a P.G class in the papers Micro and Macro economics are given below, calculate rank correlation coefficient and enterprate the result.

X	10	20	35	35	40
Y	25	15	30	30	46

Sol^n

The given problem, there are two types of cases, hence we have use the formula of Rank corrdation coefficient for tied rank, which is follows –

$$\rho = 1 - \frac{6\left\{\Sigma d_i^2 + \Sigma \frac{m^3 - m}{12}\right\}}{n^3 - n}$$

Let us construct the following table for calculation:

X	Y	x	y	di (x−y)	di²	
10	25	5	4	1	1	
20	15	4	5	−1	1	
35	30	2.5	2.5	0	0	
35	30	2.5	2.5	0	0	
40	46	1	1	0	0	

$\Sigma d_i^2 = 2$

Now the adjustment factor $= \dfrac{m^3 - m}{12}$

$$= \frac{m^3 - m}{12} + \frac{m^3 - m}{12}$$

$$= \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12}$$

$$= \frac{6}{12} + \frac{6}{12}$$

$$= 1$$

$$\therefore \rho = 1 - \frac{6\left\{\Sigma d_i^2 + \Sigma \frac{m^3 - m}{12}\right\}}{n^3 - n}$$

$$= 1 - \frac{6\{2 + 1\}}{5^3 - 5}$$

$$= 1 - \frac{18}{120}$$

$$= 1 - 0.15$$

$$= 0.85$$

Interpretation:

$\rho = 0.85$ means that, rank correlation coefficient is high and positive. Meaning that, students who have perform well in Micro economics, is also perform more or less in the similler manner in macro economics.

2. Calculate rank correlation coefficient from the following data and interprete the result.

X	85	90	85	85	70	50	60
Y	66	63	80	80	50	71	63

Soln

Given the ~~formula~~ problem, there are three type of cases, hence we have use the fromula of Rank correlation coefficient for tied rank, which is follows —

$$\rho = 1 - \frac{6\left\{\Sigma d_i^2 + \Sigma \frac{m^3 - m}{12}\right\}}{n^3 - n}$$

Let us construct the following table for calculation:

X	Y	x	y	di (x−y)	di²
85	66	3	4	−1	1
90	63	1	5.5	−4.5	20.25
85	80	3	1.5	1.5	2.25
85	80	3	1.5	1.5	2.25
70	50	5	7	−2	4
50	71	7	3	4	16
60	63	6	5.5	.5	.25

$\Sigma d_i^2 = 46$

Now the adjustment factor $= \dfrac{m^3 - m}{12}$

$$= \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12}$$

$$= \frac{24}{12} + \frac{6}{12} + \frac{6}{12}$$

$$= 2 + 0.5 + 0.5$$

$$= 3$$

$$\therefore \rho = 1 - \frac{6\left\{\sum d_i^2 + \frac{m^3 - m}{12}\right\}}{n^3 - n}$$

$$= 1 - \frac{6\{46 + 3\}}{7^3 - 7}$$

$$= 1 - \frac{\cancel{298}\ 294}{336}$$

$$= 1 - \cancel{0.768}\ 0.875$$

$$\cancel{= 0.232}$$
$$= 0.125$$

Interpretation:

$\rho = 0.125$, this means there is a very law but positive correlation of coefficient between the variable.

Q. ~~Spurious~~ Spearman correlation and zero correlation.

$$\sigma = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

convert to zero correlation $\frac{6\sum d_i^2}{n^3 - n} = 1$

$$\therefore 6 = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

$$= 1 - 1 \quad \left[\frac{6\sum d_i^2}{n^3 - n} = 1\right]$$

$$= 0$$

= zero correlation

$\therefore$ ~~Spurious~~ Spearman correlation same with zero correlation when $\frac{6\sum d_i^2}{n^3 - n} = 1$.

* Spurious correlation / Non-Sense correlation.

Sometimes, we may come across two variables which do not have any logical relationship between them but technically when calculated with emperical data, their correlation value may very high. Such kinds of correlation is called Spurious or Non Sense correlation.

For example: Correlation between National Income of a country and the height of Citizens.

8. Regression analysis.

The dictonary meaning of the word regression is 'stepping back' or 'Going back'. The term regression was first used by Sir Francis Galton. Regrassion is the measure of the average relationship between two or more variables in terms of the original units of the relationship between variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting.

9. Use of Regression analysis.

(I) It provides estimate of values of dependent variables from values of independent variable.

(II) It can be extended to two or more variables, which is known as multiple regression.

(III) It shows the nature of relationship between two or more variables.

(IV) With the help of regression coefficient we can calculate the correlation coefficient.

10. Line of regression.

In a simple linear regression, there are two regression lines constructed for the relationship between two variable, say X and Y.

(i) Line of Regression of Y on X :

$$Y-\bar{Y} = r.\frac{\sigma_Y}{\sigma_X}(X-\bar{X})$$

$$\text{or } Y-\bar{Y} = \frac{\sigma_{XY}}{\sigma_X^2}(X-\bar{X})$$

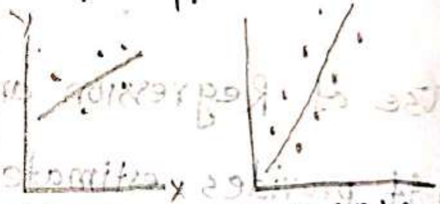
(ii) Line of Regression of X on Y :

$$X-\bar{X} = r.\frac{\sigma_X}{\sigma_Y}(Y-\bar{Y})$$

$$\text{or } X-\bar{X} = \frac{\sigma_{XY}}{\sigma_Y^2}(Y-\bar{Y})$$

Q. Why two Regression line?

In regression analysis, there are usually two regression line two show the average relationship between X & Y variables. It means that if there are two variables X and Y then one line represent regression of Y upon X and other shows the regression of X upon Y.



Y on X   X on Y

~~13. Properties of Regression equation line~~.

11. ~~Pro~~ Note :

* If $r = \pm 1$ → Two regression line become identical
* if $r = 0$ → Two regression line become perpendicular
* If $r \neq 0, r \neq \pm 1$ → Regression line intersect each other.

12. Regression coefficient.

Regression coefficient is defined as the slop of the line of regression. It shows the change in dependent variable due to per unit change in independent variable. For instance, regression coefficient of Y on X shows the change in Y with respect to unit change in X. It is denoted by b_{YX} and is given by:

$$b_{yx} = \frac{\sigma_{xY}}{\sigma_{x}^{2}}$$

$$\text{or } b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

Regression coefficient of X on Y, on the other hand, shows the change in X with respect to unit change in Y. It is denoted by b_{xy} and given by:

$$b_{xy} = \frac{\sigma_{xY}}{\sigma_{y}^{2}}$$

$$\text{or } b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

5.0 13. Properties of Regression coefficient.

(i) Regression coefficient is the geometric mean between the regression coefficients, i.e. $r = \pm\sqrt{b_{yx} \cdot b_{xY}}$

(ii) If one of the regression coefficients is greater than unity then the other must be less than unity, i.e. if $b_{yx} > 1$ then $b_{xy} < 1$.

(iii) Arithmetic mean of the regression coefficients is greater than the correlation coefficient provided $r > 0$,

i.e. $$\frac{b_{yx} + b_{xy}}{2} > r$$

(iv) Regression coefficients are independent of change of origin but not of Scale.

Examples

1. Obtain the line of regression of Y on X and of X on Y for the following data.

$\Sigma X = 100$, $\Sigma Y = 115$, $n = 10$, $\Sigma X^2 = 2250$, $\Sigma Y^2 = 1600$

$\Sigma XY = 4000$.

Also find the value of Y when $X = 7$ and the value of X, when $Y = 8$.

Sol^n

Given that,

$$\Sigma X = 100, \quad \Sigma Y = 115, \quad n = 10$$

$$\Sigma X^2 = 2250, \quad \Sigma Y^2 = 1600, \quad \Sigma XY = 4000.$$

Now,

we know that,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{100}{10} = 10$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{115}{10} = 11.5$$

and

$$\sigma_x^2 = \frac{1}{n}\Sigma X^2 - (\bar{X})^2$$

$$= \frac{1}{10} \times 2250 - 100$$

$$= 22{,}5 - 100$$

$$= 125$$

$$\sigma_y^2 = \frac{1}{n}\Sigma Y^2 - (\bar{Y})^2$$

$$= \frac{1}{10} \times 1600 - 132.25$$

$$= 160 - 132.25$$

$$= 27.75$$

and $\sigma_{xy} = \frac{1}{n} \Sigma xy - \bar{x}\bar{Y}$

$= \frac{1}{10} \times 4000 - (10 \times 11.5)$

$= 400 - 115$

$= 285$

Now,
we know that,
Regression line of y on X $\Rightarrow Y - \bar{Y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

$\Rightarrow Y - 11.5 = \frac{285}{125}(x - 10)$

$\Rightarrow Y - 11.5 = 2.28(x - 10)$

$\Rightarrow Y - 11.5 = 2.28X - 22.8$

$\Rightarrow Y = -11.3 + 2.28X$ ——— (1)

Again,
Regression line of x on y $\Rightarrow x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$

$\Rightarrow X - 10 = \frac{285}{27.75}(y - 11.5)$

$\Rightarrow X - 10 = 10.27(y - 11.5)$

$\Rightarrow x - 10 = 10.27Y - 118.105$

$\Rightarrow X = -108.105 + 10.27Y$ ——— (11)

$\therefore$ when $x = 7$, from eq^n (1) $\Rightarrow Y = -11.3 + 2.28(7)$
$= 4.66$

when $y = 8$, from eq^n (11) $\Rightarrow X = -108.105 + 10.27(8)$
$= -25.945$

2. Given the following data, obtain the line of regression of Y on X and also predict the value of Y for X=10.

$\Sigma X = 200$, $\Sigma Y = 140$, $n = 10$, $\Sigma XY = 5000$, $\Sigma(X-\bar{X})^2 = 6400$.

Soln

Given that,

$$\Sigma X = 200, \quad \Sigma Y = 140, \quad n = 10$$

$$\Sigma XY = 5000, \quad \Sigma(X-\bar{X})^2 = 6400$$

We know that,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{200}{10} = 20$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{140}{10} = 14$$

And,

$$\sigma_x^2 = \frac{1}{n}\Sigma(X-\bar{X})^2$$

$$= \frac{1}{10} \times 6400$$

$$= 640$$

$$\sigma_{xy} = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y}$$

$$= \frac{1}{10} \times 5000 - (20 \times 14)$$

$$= 500 - 280$$

$$= 220$$

We know that,

Regression line of Y on X $\Rightarrow Y - \bar{Y} = \frac{\sigma_{xy}}{\sigma_x^2}(X-\bar{X})$

$$\Rightarrow Y - 14 = \frac{220}{640}(X-20)$$

$$\Rightarrow Y - 14 = 0.34(X-20)$$

$$\Rightarrow Y - 14 = 0.34X - 6.875$$

$$\Rightarrow Y = 7.125 + 0.34X$$

Again,
The value of Y, when $x = 10$,

$$Y = 7.125 + 0.34(10)$$
$$= 7.125 + 3.4$$
$$= 10.525$$

1.3 Simple / Total, Partial and Multiple correlation

S. Correlation is a measure used to determine the strength and the direction of the relationship between two variable x and Y. $-1 < r < 1$

14. Simple correlation coefficient.

→ Karl Pearson's correlation coefficient.

15. Partial correlation coefficient.

In particular correlation, there more than two variable, but consider only two variable to be influencing each other, the effect of other variable being kept constant.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Example:

1. Find the partial correlation coefficient $r_{12.3}$ by using the following data.

$r_{12} = 0.77$, $r_{13} = 0.72$, $r_{23} = 0.52$

Also interprate the result.

Sol^n

Given that,

$r_{12} = 0.77$

$r_{13} = 0.72$

$r_{23} = 0.52$

we know that,

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.77 - (0.72 \times 0.52)}{\sqrt{(1-0.72^2)(1-0.52^2)}}$$

$$= \frac{0.77 - 0.3744}{\sqrt{(1-0.5184)(1-0.2704)}}$$

$$= \frac{0.3956}{\sqrt{0.4816 \times 0.7296}}$$

$$= \frac{0.3956}{0.5928}$$

$$= 0.6678$$

Interpretation:

$r_{12.3}$ means that correlation between first two variable after eleminating linear effect of 3rd variable upon them is positive and moderate.

16. Multiple correlation coefficient.

When there are more variables are studied, it is a problem of either multiple or partial correlation. In multiple correlation there are more variable are studied simultaneously.

* $$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

* $$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12} \cdot r_{23} \cdot r_{13}}{1 - r_{13}^2}}$$

<u>Note:</u>

Multiple correlation coefficient lies between 0 to 1.

i.e., $\boxed{0 \leqslant R \leqslant 1}$

<u>Note:</u> Relation among simple, partial and Multiple correlation.

Multiple correlation = (Simple correlation) × (partial correlation)

* Advance measure of correlation

Phi coefficient:

Phi coefficient is define as a correlation coefficient or assessing strangth and nature of relationship between two dichotomous variable.

Forexample, correlation between marital status and smoking.

Point Bi-serial Correlation:

Point Bi-serial correlation include both continuous and dichotomus variable.

Forexample, correlation between income and gender.

* Scale of measure.

(I) Interval Scale } Continuous variable
(II) Ratio Scale }
(III) Ordinal scale – (qualitative approach) (rank).
(IV) Nominal scale (Dichotomous in nature)
→ 2 Category.

Proof

1. Methametically derive / prove spairman's coefficient correlation in case of non- tied rank.

$$\left[r_s = 1 - \frac{6\Sigma d_i^2}{N^3 - N} \right]$$

Proof:

Let $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ be the rank of 'n' individuals ~~in the two~~ characterstics A and B respectively.

Let $d_i = x_i - y_i$ denote the difference between the rank of 'i'th individual in the two characterstics. If we assume that there is no tie, then each of variables x and y takes the values $1, 2, 3, \ldots n$.

$$\therefore \bar{x} = \bar{y} = \frac{n+1}{2}$$

and $\sigma_x^2 = \sigma_y^2 = \frac{n^2-1}{12}$ —————— (i)

we have,

$$d = x - y$$

$$\Rightarrow d = (x - \bar{x}) - (y - \bar{y})$$

$$\Rightarrow d^2 = \left[(x-\bar{x}) - (y-\bar{y})\right]^2$$

$$= (x-\bar{x})^2 - 2(x-\bar{x})(y-\bar{y}) + (y-\bar{y})^2$$

Summing both side over n values and dividing by n, we get.

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

$$= \sigma_x^2 + \sigma_y^2 - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n}$$ —————— (ii)

Spearman's rank correlation coefficient raw is given by,

$$\rho = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\,\sigma_y}$$

$$\Rightarrow \rho\,\sigma_x\,\sigma_y = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} \qquad \text{——— (III)}$$

Now, from eq^n(II),

$$\frac{\Sigma d^2}{n} = \sigma_x^2 + \sigma_y^2 - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$\Rightarrow \frac{\Sigma d_i^2}{n} = \sigma_x^2 + \sigma_y^2 - 2\rho\,\sigma_x\,\sigma_y \quad [\text{from eq}^n\ (\text{III})]$$

$$\Rightarrow \frac{\Sigma d^2}{n} = 2\sigma_x^2 - 2\rho\,\sigma_x^2 \quad [\because \sigma_x = \sigma_y]$$

$$\Rightarrow \frac{\Sigma d^2}{n} = 2\sigma_x^2(1-\rho)$$

$$\Rightarrow 1-\rho = \frac{\Sigma d^2}{2n\,\sigma_x^2}$$

Now, from eq^n (1) $\Rightarrow 1-\rho = \frac{\Sigma d^2}{2n\cdot\sigma_x^2}$

$$\Rightarrow 1-\rho = \frac{\Sigma d^2}{\cancel{2}n\left(\frac{n^2-1}{\cancel{12}_{\,6}}\right)}$$

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

$$\Rightarrow \rho = 1 - \frac{6\Sigma d^2}{n^3-n}$$

proved.

2. Proof that Karl pearson Correlation coefficient are lies between -1 and $+1$. i.e., $-1 \leq r \leq 1$.

Proof.

Let x and y be, deviation of X and Y series from their Mean and σ_x and σ_y by their standard deviations.

Expand the function,

$$\Sigma\left(\frac{X}{\sigma_x}+\frac{Y}{\sigma_y}\right)^2 = \Sigma\left(\frac{X^2}{\sigma_x^2}+\frac{2XY}{\sigma_x\sigma_y}+\frac{Y^2}{\sigma_y^2}\right)$$

$$= \frac{\Sigma X^2}{\sigma_x^2}+\frac{2\Sigma XY}{\sigma_x\sigma_y}+\frac{\Sigma Y^2}{\sigma_y^2} \quad \text{———— (1)}$$

But, $\frac{\Sigma X^2}{\sigma_x^2} = N$ and Similarly,

$$\frac{\Sigma Y^2}{\sigma_y^2} = N$$

$\therefore$ from eqn (1) $\Rightarrow$ $\frac{\Sigma X^2}{\sigma_x^2}+\frac{2\Sigma XY}{\sigma_x\sigma_y}+\frac{\Sigma Y^2}{\sigma_y^2}$

$$= N + 2Nr + N$$

$$= 2N + 2Nr$$

$$= 2N(1+r)$$

$\because \sigma_x^2 = \frac{\Sigma X^2}{N}$

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

$= N$

Again,

$\frac{2\Sigma XY}{\sigma_x\sigma_y} = 2\frac{\Sigma XY}{N\cdot\sigma_x\sigma_y}\cdot N$

$= 2Nr$

Now,

$\Sigma\left(\frac{X}{\sigma_x}+\frac{Y}{\sigma_y}\right)$ is the sum of squares of real quantities and it cannot be negative, at the most it can be zero.

$\therefore \{2N(1+r) \geq 0\}$

$\Rightarrow (1+r) \geq 0$

$\Rightarrow r \geq -1$, and it cannot be less than -1, at the most it can be -1.

Similarly,

By expand $\Sigma\left(\frac{X}{\sigma x} - \frac{Y}{\sigma y}\right)$, it can be show that this is equal to $2N(1-r)$, this is again can not be Negative, at the most it can be zero.

Therefore, r cannot be greater than +1, at the most it can +1.

Hence,

$$-1 \leq r \leq 1$$

proved.

UNIT: 2

RANDOM VARIABLES AND PROBABILITY DISTRIBUTION

1. Random Variable.

A random variable is one whoes all values are known before-hand but exect predictions about the occurance of this values can not be made with cent percent assurance.

Forexample, in case of throwing a unbias dice, only 6 results are possible, i.e., 1,2,3,....6 but we can not say with certainty which value ~~of~~ out of this six we will get.

Note:

(I) Random variable takes values with their respective probabilities and the sum of all these probabilities is equal to unity. $(p+q=1)$.

(II) The value taken by a random variable are called: Spectrum.

Types: Random variable

$$\boxed{\text{Random variable}} \begin{cases} \rightarrow \text{Discrete Random Variable.} \\ \rightarrow \text{Continuous Random variable.} \end{cases}$$

(I) Discrete Random variable:

A Discrete random variable is one which takes finite or countable number of infinite values.

(II) Continuous Random variable:

A continuous random variable is one which takes values within some limit or range.

2. PMF (Probability mass function) or Discrete probability distribution.

If X is a discrete random variable then $p(x) = p(X=x)$ is said to be a probability mass function of X if the following two condition are satisfies:-

CONDITION(I): Condition of Non-Negativity, i.e., $p(x) \geqslant 0$

CONDITION(II): Condition of totality. i.e., $\Sigma p(x) = 1$

3. PDF (Probability Density Function) or continuous probability distribution.

If X is a continuous random random variable taking values within the range $-\alpha$ to $+\alpha$ then $f(x)$ will be called the PDF of X, if the following two condition are satisfied:-

CONDITION (I): Condition of Non-Negativity, $f(x) \geqslant 0$

CONDITION (II): Condition of totality, i.e., $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Exercise:

Q. Given the following probability Distribution, check wheather $P(x)$ is a PMF of X.

X	2	5	9	10
$P(x)$	$1/8$	$3/8$	$5/8$	$1/8$

Soln

To be the PMF of X, $P(x)$ needs to satisfy the following two condition.

(I) $P(x) \geqslant 0$

(II) $\Sigma P(x) = 1$

In the given problem, No value of $p(x)$ is less than 0, i.e., to say here, $P(x) \geqslant 0$

Again,

$$\Sigma p(x) = 1/8 + 3/8 + 5/8 + 1/8$$
$$= 10/8$$
$$= 1.25 > 1$$

$\because \Sigma P(x) > 1$, thus the second condition is not satisfied.

since, only one of the condition for PMF is satisfied here, can be conclude that $P(x)$ in this problem is not the PMF of X.

Q. Given the following probability distribution, check wheather $P(x)$ is a PMF of X.

$e = 2.718$

X	0	1	2	3
$P(x)$	e^{x}	e^{-x}	$1/e^{2x}$	0

Soln
To be the pmf of X. the $P(x)$ need to be satisfy two condition -

(i) $P(x) \geq 0$

(ii) $\sum P(x) = 1$

in given problem, no value of $P(x)$ is less than 0. other than e^{-x}, if x takes any value the condition will be satisfy, so, $P(x) \geq 0$.

Again,

$\sum P(x) = 1$

$\Rightarrow e^{-x} = e^{-1} = 0.37$

$\Rightarrow \frac{1}{e^{2x}} = \frac{1}{e^4} = \frac{1}{54.60} = 0.018$

$\therefore \sum P(x) = 1 + 0.37 + 0.018 + 0$

$= 1.38 > 1$

We know,

$\sum P(x) = 1$

Here, $\Rightarrow e^{x} + e^{-x} + \frac{1}{e^{2x}} + 0 =$

$\Rightarrow e^{0} + e^{-1} + \frac{1}{e^{2\times 2}}$

$\Rightarrow 1 + \frac{1}{e} + \frac{1}{e^4}$

$\Rightarrow 1 + \frac{1}{2.718} + \frac{1}{(2.718)^4}$

$\Rightarrow 1 + 0.368 + \frac{1}{54.576}$

$\Rightarrow 1 + 0.368 + 0.018 =$

$\Rightarrow 1.386 > 1$

Thus the condition of totality is not satisfied and therefore the given $P(x)$ is not the pmf of X.

4. Distribution Function.

Property:

First derivative / first differentiation of distribution function gives the value of probability density function (PDF).

i.e.,

$$\underbrace{\frac{d}{dx}\{F(x)\}}_{\text{Distribution function}} = \underbrace{f(x)}_{\text{pdf}}$$

$$\boxed{\text{Distribution}} \xrightarrow{\text{1st derivative}} \boxed{\text{PDF}}$$

Example:

For the following distribution function, find the values of their corresponding pdf.

(1) $F(x) = \log(1+x)^2 + x.e^{-x} + \frac{1}{x^4}$

Soln

$\frac{2}{(1+x)} + x.\frac{1}{e^x}$

- Distribution Function
  - Distribution function for Discrete random variable.
  - Distribution function for Continuous random variable

Distribution function for Discrete random variable

If X is a discrete random variable with pmf $P(X=x)$, then cumulative distribution function or simply distribution function of X is denoted by $F(X)$, is defined as–

Distribution Function for continuous Random Variable

If X is a continuous random variable taking the values within the interval $-\infty$ to $+\infty$ and $f(x)$ be its pdf then its distribution function is denoted by $F(X)$ and is define as –

$$F(X) = \int_{-\infty}^{x} f(t)\,dt$$

5. Mathematical expectation.

Mathematical expectation of a Discrete random Variable:

If X is a discreter random variable taking the values $x_1, x_2, \dots x_n$ with their respective probabilities $P(x_1), P(x_2), \dots P(x_n)$, then its mathematical expectation or simply expectation is denoted as- $E(X)$ and is defined as –

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + \dots\dots\dots + x_n P(x_n)$$

$$= \sum_{i=1}^{n} x_i \cdot P(x_i)$$

here, $p(x_i) = $ pmf of X.

Mathematical expectation of a Continuous Random variable:

If X is continuous random variable taking values within the interval $-\infty$ to $+\infty$, then its mathematical expectation is denoted as $E(X)$ and is defined as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

here, $f(x) = $ pdf of X.

Mathematical expectation of a function of a Random variable:

Discrete:

Let $g(x)$ be the function of a discrete random variable X with its pmf $p(x)$. Now mathematical expectation of $g(x)$ is denoted as $E\{g(x)\}$ and is defined as–

$$E\{g(x)\} = \sum_{x} g(x) . p(x)$$

Continuous:

Let $g(x)$ be the function of a continuous random variable X taking the values within the interval $-\infty$ to $+\infty$ with its pdf $f(x)$. Now mathematical expectation of $g(x)$ is denoted as $E\{g(x)\}$ and is defined as–

$$E\{g(x)\} = \int_{-\infty}^{\infty} g(x) . p(x)\, dx$$

Later * **Properties of mathematical expectation**

(1). $E(a) = a$, where a is any arbitary constant.

Proff:

$$\begin{aligned} L.H.S &= E(a) \\ &= \sum a . p(x) \\ &= a \sum p(x) \\ &= a . 1 \\ &= a = R.H.S. \text{ proved} \end{aligned}$$

6. Variance of a Random Variable.

If x is a random variable then its variance is denoted as $V(x)$ and is defined as -

$$V(x) = E\left[\left\{X - E(X)\right\}^2\right]$$

For solving the practical problem $V(X)$ is calculated by the following simplified formulla -

$$V(X) = E(X^2) - \left\{E(X)\right\}^2$$

7. Covariance of two random variables.

If X and Y are two random variables then covariance between them is denoted by $Cov(X, Y)$ and is defined as -

$$Cov(X, Y) = E\left[\left\{X - E(X)\right\}\left\{Y - E(Y)\right\}\right]$$

For solving the practical problem $cov(X, Y)$ is calculated by the following simplified formulla -

$$cov(X, Y) = E(XY) - E(X) \cdot E(Y)$$

$\boxed{\text{Note:}}$

(I) $covariance(X, Y) = E(XY) - E(X) \cdot E(Y)$

$$= E(X) \cdot E(Y) - E(X) \cdot E(Y)$$

(II) $E(X) = \overline{X}$

8. Show that average is the most expected value, i.e., $E(X) = \bar{X}$.

Prof:

For a frequiency distribution f_i/x_i, average is defined as,

$$\bar{x} = \frac{1}{n} \Sigma f_i x_i \quad \text{------------------ (I)}$$

By difinition of mathemetical expectation,

$$E(X) = \sum_{x} X \cdot P(x) \quad \text{------------------ (II)}$$

By the Classical difinition of probability,

$$P(x) = \frac{m}{n} \quad \text{------------------ (III)}$$

here, n = total number of cases
m = favourable number of cases for occurance of an event.

Again by definition of relation frequency, we have,

$$RF = \frac{f_i}{N} \quad \text{------------------ (IV)}$$

Equation (III) shows its favourable number of cases for its occurance of an event out of total number of cases, where eq^n (IV) shows the frequency of i th case in relation to total frequency. Hence, corresponding and logically of eq^n (III) and (IV) we can write that –

$$P(x) = RF = \frac{f_i}{N} \quad \text{------------------ (V)}$$

Let us, re-write equation(1),

$$\overline{x} = \frac{1}{N}\sum f_i x_i$$

$$= \sum \frac{f_i}{N} \cdot x_i$$

$$= \sum x_i \frac{f_i}{N}$$

$$= \sum x_i \cdot P(x) \quad [\text{from eq}^n \text{ (v)}]$$

$$= E(x)$$

~~prov.~~

$\therefore E(x) = \overline{x}$ proved.

9. Properties of Mathematical expectation.

(I) $E(a) = a$, where a is any arbitary constant

Proof:

$$\begin{aligned} L.H.S &= E(a) \\ &= \sum a \cdot p(x) \\ &= a \cdot \sum p(x) \\ &= a \cdot 1 \\ &= a = R.H.S \end{aligned}$$

proved.

(II) $E(ax) = a\,E(X)$

Proof:

$$\begin{aligned} L.H.S &= E(ax) \\ &= \sum a x \cdot p(x) \\ &= a \sum x \cdot p x \\ &= a\,E(X) = \end{aligned}$$

Proved

(III) $E(ax+b) = a\,E(x) + b$

Proof:

$$\begin{aligned} L.H.S &= E(ax+b) \\ &= \sum (ax+b) \cdot p(x) \\ &= \sum \{ax \cdot p(x) + b \cdot p(x)\} \\ &= \sum ax \cdot p(x) + b \sum p(x) \\ &= a \sum x \cdot p(x) + b \quad \left[\because \sum p(x) = 1\right] \\ &= a\,E(X) + b. \\ &= R.H.S \end{aligned}$$

proved.

(IV) $E(\bar{x}) = \bar{x}$

Proof:

$$
\begin{aligned}
L.H.S &= E(\bar{x}) \\
&= \Sigma\, \bar{x} \cdot p(x) \\
&= \bar{x}\, \Sigma\, p(x) \\
&= \bar{x} \cdot 1 \\
&= \bar{x} \\
&= R.H.S \text{ proved.}
\end{aligned}
$$

(V) Additive property of Mathematical expectation:

This property stats that if X and Y are two random variable then expectation of the sum of these two variable is equal to the sum of their individual expectations. i.e.,

$$E(X+Y) = E(X) + E(Y)$$

This property can continue till 'n' number of variable.

(VI) Multiplicative property of Mathematical expectation:

This property stats that if X and Y are two independent random variable then mathematical expectation of the product of these two variable is equall to the product of their individual mathematical expectation. i.e.,

$$E(X,Y) = E(X). E(Y)$$

This can be continue till 'n' number of variable.

10. Properties of varience.

(I) $v(a)=0$

proof:

$$\begin{aligned} L.H.S &= v(a) \\ &= E(a^2) - \{E(a)\}^2 \\ &= a^2 - a^2 \\ &= 0 \\ &= R.H.S \text{ proved.} \end{aligned}$$

(II) $v(ax) = a^2 \cdot v(x)$

proof:

$$\begin{aligned} L.H.S &= v(ax) \\ &= E(a.x)^2 - \{E(ax)\}^2 \\ &= E(a^2 . x^2) - \{a\,E(x)\}^2 \\ &= a^2 E(x^2) - a^2\{E(x)\}^2 \\ &= a^2\left[E(x^2) - \{E(x)\}^2\right] \\ &= a^2 . v(x). \\ &= R.H.S \text{ proved.} \end{aligned}$$

(III) $v(\bar{x})=0$

proof:

$$\begin{aligned} L.H.S &= v(\bar{x}) \\ &= E(\bar{x})^2 - \{E(\bar{x})\}^2 \\ &= \bar{x}^2 - \bar{x}^2 \\ &= 0 \\ &= R.H.S \text{ proved.} \end{aligned}$$

(iv) $v(ax+b) = a^2 v(X)$

proof:

$$
\begin{aligned}
L.H.S &= v(aX+b) \\
&= E\{(ax+b)^2\} - [E(aX+b)]^2 \\
&= E\{(aX)^2 + 2.aX.b + b^2\} - [a\,E(X)+b]^2 \\
&= E\{a^2.X^2 + 2aXb + b^2\} - [\{aE(X)\}^2 + 2.aE(X).b + b^2] \\
&= a^2 E(X^2) + 2ab\,E(X) + b^2 - a^2\{E(X)\}^2 - 2ab\,E(X) - b^2 \\
&= a^2 E(X^2) - a^2\{E(X)\}^2 \\
&= a^2[E(X^2) - \{E(X)\}^2] \\
&= a^2 v(X) \\
&= R.H.S \text{ proved.}
\end{aligned}
$$

(v) $v(X+Y) = v(X) + v(Y)$, if X and Y are indipendent random variable.

proof:

$$
\begin{aligned}
L.H.S &= v(X+Y) \\
&= E[(X+Y)^2 - \{E(X+Y)\}^2] \\
&= E(X^2+2XY+Y^2) - \{E(X)+E(Y)\}^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - [\{E(X)\}^2 + 2.E(X).E(Y) + \{E(Y)\}^2] \\
&= E(X^2) + 2.E(X).E(Y) + E(Y^2) - \{E(X)\}^2 - 2E(X).E(Y) - \{E(Y)\}^2 \\
&= [E(X^2) - \{E(X)\}^2] + [E(Y^2) - \{E(Y)\}^2] \\
&= v(X) + v(Y)
\end{aligned}
$$

(VI) $v(X+Y) = v(X) + v(Y) + 2\,cov(X,Y)$, if X and Y are not independent random variables.

Proof:

$$
\begin{aligned}
L.H.S &= v(X+Y) \\
&= E(X+Y)^2 - \{E(X+Y)\}^2 \\
&= E(X^2+2XY+Y^2) - \{E(X) + E(Y)\}^2 \\
&= E(X)^2 + 2E(XY) + E(Y)^2 - \left[\{E(X)\}^2 + 2.E(X).E(Y) + \{E(Y)\}^2\right] \\
&= E(X)^2 - \{E(X)\}^2 + E(Y)^2 - \{E(Y)\}^2 + 2E(XY) - 2E(X).E(Y) \\
&= v(X) + v(Y) + 2\{E(XY) - E(X).E(Y)\} \\
&= v(X) + v(Y) + 2\,Cov(X,Y).
\end{aligned}
$$

(VII) $v(X-Y) = v(X) + v(Y)$ if X and Y are independent random variable.

Proof.

$$
\begin{aligned}
R.H.S &= v(X-Y) \\
&= E\left[(X-Y)^2 - \{E(X-Y)\}^2\right] \\
&= E(X^2 - 2XY + Y^2) - \{E(X) - E(Y)\}^2 \\
&= E(X)^2 - 2E(XY) + E(Y)^2 - \left[\{E(X)\}^2 - 2E(X).E(Y) + \{E(Y)\}^2\right] \\
&= E(X)^2 - \cancel{2E(X)E(Y)} + E(Y)^2 - \{E(X)\}^2 + \cancel{2E(X).E(Y)} - \{E(Y)\}^2 \quad [\text{Since } X \text{ and } Y \text{ are independent} \\
&= \left[E(X)^2 - \{E(X)\}^2\right] + \left[E(Y)^2 - \{E(Y)\}^2\right] \\
&= v(X) + v(Y) \\
&= R.H.S \text{ proved.}
\end{aligned}
$$

(VIII) $v(X-Y) = v(X) + v(Y) - 2\,Cov(X,Y)$, if X and Y are not independent random variable.

Proof:

$$
\begin{aligned}
L.H.S &= v(X-Y) \\
&= E(X-Y)^2 - \{E(X-Y)\}^2 \\
&= E(X^2 - 2XY + Y^2) - \{E(X) - E(Y)\}^2 \\
&= E(X^2) - 2E(XY) + E(Y^2) - \left[\{E(X)\}^2 - 2E(X)E(Y) + \{E(Y)\}^2\right] \\
&= E(X^2) - 2E(XY) + E(Y^2) - \{E(X)\}^2 + 2E(X)E(Y) - \{E(Y)\}^2 \\
&= E(X^2) - \{E(X)\}^2 + E(Y^2) - \{E(Y)\}^2 - 2E(XY) + 2E(X)E(Y) \\
&= v(X) + v(Y) - 2\{E(XY) + E(X), E(Y)\} \quad [\because X, Y \text{ are independent random variable}] \\
&= v(X) + v(Y) - 2\,Cov(X,Y) \\
&= R.H.S
\end{aligned}
$$

proved.

Exercise

Q. Let x is a discrete random variable with following pdf.

x	0	1	2	3	4	5
$P(x)$	k	$2k$	$3k$	k^2	$6k^2$	0

Find the following values –

(i) Find the value of k

(ii) Find the value of E(X) and V(X)

(iii) Find the value of $V(2x+5)$

Solⁿ

(i) since total probability = 1

$\therefore \Sigma P(x) = 1$

$\Rightarrow k+2k+3k+k^2+6k^2+0 = 1$

$\Rightarrow 7k^2+6k-1 = 0$

$\Rightarrow 7k^2+7k-k-1 = 0$

$\Rightarrow 7k(k+1)-1(k+1) = 0$

$\Rightarrow (7k-1)(k+1)$

$\therefore$ Either,

$7k-1 = 0$ or $k+1 = 0$

$\Rightarrow k = 1/7$ $\qquad \Rightarrow k = -1$

Since, the probability of anything cannot be negative,

$\therefore$ the value of $k = 1/7$.

(ii) By definition,

$$E(x) = \sum x \cdot p(x)$$
$$= 0.k + 1.2k + 2.3k + 3.k^2 + 4.6k^2 + 5.0$$
$$= 2k + 6k + 3k^2 + 24k^2$$
$$= 8k + 27k^2$$
$$= 8\times\left(\frac{1}{7}\right) + 27\left(\frac{1}{7}\right)$$
$$= \cancel{\frac{8}{7}}$$
$$= 8/7 + 27/49$$
$$= 0.55 + 1.14$$
$$= 1.69$$

Again,

By definition

$$V(x) = E(x^2) - \{E(x)\}^2$$
$$= E(x^2) - (1.69)^2$$
$$= E(x^2) - 2.86 \quad \text{------------ (1)}$$

Now, $E(x^2) = \sum x^2 \cdot p(x)$

$$= 0^2.k + 1^2.2k + 2^2.3k + 3^2.k^2 + 4^2.6k^2 + 5^2 0$$
$$= 2k + 12k + 9k^2 + 96k^2$$
$$= 105k^2 + 14k$$
$$= 105\left(\frac{1}{49}\right) + 14\left(1/7\right)$$
$$= 2.14 + 2$$
$$= 4.14$$

Now, inserting the value of $E(x^2)$ in equation (1)

we get, $E(x^2) - 2.86$

$$= 4.14 - 2.86$$
$$= 1.29$$

(iii) $v(2x+5) = a^2 \cdot v(x)$ [property of variance]

$= 2^2 \times 1.29$

$= 4 \times 1.29$

$= 5.16$

Q. Examine wheather the following is a pdf of x or not

$f(x) = 2x^2+3, \quad 0 \leq x \leq 1$

$= 0$, other wise.

Sol^n

For $f(x)$ to be the pdf of x, the following two condition must be satisfy –

(i) $f(x) \geq 0$

(ii) $\int_0^1 f(x)\,dx = 1$

For the extreme values of x, the value of $f(x)$ in this problem are,

when $x=0$, $f(x) = 3$ and $(2.0^2+3 = 3)$

when $x=1$, $f(x) = 5$ $(2.1^2+3 = 5)$

Moreover for any value of x between 0 and 1 the value of $f(x)$ is non-negative.

For example,

if $x = 1/2$, then $f(x) = 3.5$

This means that $f(x) \geq 0$, for $0 \leq x \leq 1$

Hence the first condition of non-negativity is satisfied.

Now, let us check the second condition,

$$\int_0^1 f(x)\,dx = \int_0^1 (2x^2+3)\,dx$$

$$= \left[2 \cdot \frac{x^3}{3}\right]_0^1 + \left[3x\right]_0^1$$

$$= \left[\left(2 \cdot \frac{1}{3}\right) - (2.0)\right] + \left[(3.1) - (3.0)\right]$$

$$= \frac{2}{3} + 3$$

$$= .67 + 3$$

$$= 3.67 > 1$$

Thus the condition of totality is not satisfied and therefore the given $f(x)$ is not a pdf of X.

Q. For the following pdf,

$$f(x) = cx^2(1-x) \;;\; 0 \leq x \leq 1$$

Find the values of the constant c, mean and variance.

Soln

Since the total probability = 1

we have,

$$\int_0^1 f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 cx^2(1-x)\,dx = 1$$

$$\Rightarrow \int_0^1 (cx^2 - cx^3)\,dx = 1$$

$$\Rightarrow c\int_0^1 x^2 dx - c\int_0^1 x^3\,dx = 1$$

$$\Rightarrow c\left[\frac{x^3}{3}\right]_0^1 - c\left[\frac{x^4}{4}\right]_0^1 = 1$$

$$\Rightarrow c\left(\frac{1^3}{3} - \frac{0^3}{3}\right) - c\left(\frac{1^4}{4} - \frac{0^4}{4}\right) = 1$$

$$\Rightarrow c\left(\frac{1}{3}\right) - c\left(\frac{1}{4}\right) = 1$$

$$\Rightarrow \frac{c}{3} - \frac{c}{4} = 1$$

$$\Rightarrow \frac{4c - 3c}{12} = 1$$

$$\Rightarrow \frac{c}{12} = 1$$

$$\Rightarrow c = 12$$

2ND Part:

By definition,

$$E(X) = \int_0^1 x.f(x)\,dx$$

$$= \int_0^1 x . 12x^2(1-x)\,dx$$

$$= 12\int_0^1 x^3(1-x)\,dx$$

$$= 12\int_0^1 (x^3 - x^4)\, dx$$

$$= 12\int_0^1 x^3\, dx - 12\int_0^1 x^4\, dx$$

$$= 12\left[\frac{x^4}{4}\right]_0^1 - 12\left[\frac{x^5}{5}\right]_0^1$$

$$= 12\left[\frac{1^4}{4} - \frac{0^4}{4}\right] - 12\left[\frac{1^5}{5} - \frac{0^5}{5}\right]$$

$$= 12 \times \frac{1}{4} - 12 \times \frac{1}{5}$$

$$= 3 - 2.4$$

$$= 0.6$$

Again,

$$V(X) = E(x^2) - \{E(x)\}^2$$

$$= E(x^2) - (0.6)^2$$

$$= E(x^2) - 0.36 \text{ ----------------- } (1)$$

We know that,

$$E(x^2) = \int_0^1 x^2 f(x)\, dx$$

$$= \int_0^1 x^2 \cdot c x^2 (1 - x)\, dx$$

$$= 12\int_0^1 x^2 (x^2 - x^3)\, dx$$

$$= 12\int_0^1 (x^4 - x^5)\, dx$$

$$= 12\left[\frac{x^5}{5} - \frac{x^6}{6}\right]_0^1$$

$$= \frac{12}{5} - \frac{12}{6}$$

$$= 2.4 - 2$$

$$= 0.4$$

Now putting the value of $E(x) = 0.4$ in eqⁿ (1), we have,

$$E(x^2) - 0.36$$

$$\Rightarrow 0.4 - 0.36$$

$$\Rightarrow 0.04 \;//$$

9 Q. If X is a continuous random variable with following pdf

$$f(x) = k.(x-x^2) \; ; \; 0 \leq x \leq 1$$

Find the value of k, mean and variance.

Soln

Since the total probability = 1

$$\therefore \int_0^1 f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 k.(x-x^2)\,dx = 1$$

$$\Rightarrow \int_0^1 (kx - kx^2)\,dx = 1$$

$$\Rightarrow k\int_0^1 x\,dx - k\int_0^1 x^2\,dx = 1$$

$$\Rightarrow k\left[\frac{x^2}{2}\right]_0^1 - k\left[\frac{x^3}{3}\right]_0^1 = 1$$

$$\Rightarrow k\left(\frac{1^2}{2} - \frac{0^2}{2}\right) - k\left(\frac{1^3}{3} - \frac{0^3}{3}\right) = 1$$

$$\Rightarrow k \times \frac{1}{2} - k \times \frac{1}{3} = 1$$

$$\Rightarrow \frac{k}{2} - \frac{k}{3} = 1$$

$$\Rightarrow \frac{3k-2k}{6} = 1$$

$$\Rightarrow \frac{k}{6} = 1$$

$$\Rightarrow k = 6$$

Now by definition,

mean, $E(X) \Rightarrow \int_0^1 x.f(x)\,dx$

$$\Rightarrow \int_0^1 x.k(x-x^2)\,dx$$

$$\Rightarrow \int_0^1 x.6(x-x^2)\,dx$$

$$\Rightarrow 6\int_0^1 x.(x-x^2)\,dx$$

$$\Rightarrow 6\int_0^1 (x^2-x^3)\,dx$$

$$\Rightarrow 6\left[\frac{x^3}{3}-\frac{x^4}{4}\right]_0^1$$

$$\Rightarrow 6\left(\frac{1^3}{3}\right) - 6\left(\frac{1^4}{4}\right)$$

$$\Rightarrow \frac{2\cancel{6}}{\cancel{3}} - \frac{\cancel{6}^{3}}{\cancel{4}_2}$$

$$\Rightarrow 2-1.5$$

$$\Rightarrow 0.5$$

Again,

$$V(X) = E(x^2) - \{E(x)\}^2$$

$$= E(x^2) - \int_0^1 x^2 . f(x)\,dx^2,$$

$$= E(x^2) - \int_0^1 6(x-x^2)\,dx$$

$$= E(x^2) - (0.5)^2 \quad \text{---(1)}$$

$$= E(x^2) - 0.25$$

Here, $E(x^2) = \int_0^1 x^2 . f(x)\,dx$

$$= \int_0^1 x^2 . 6(x-x^2)\,dx$$

$$= 6\int_0^1 x^2(x-x^2)\,dx$$

$$= 6\int_0^1 (x^3-x^4)\,dx$$

$$= 6\left[\frac{x^4}{4}-\frac{x^5}{5}\right]_0^1$$

$= 6\left(\frac{1^4}{4}\right) - 6\left(\frac{1^5}{5}\right)$

$= \frac{6}{4} - \frac{6}{5}$

$= 1.5 - 1.2$

$= 0.3$

Now putting the value of $E(x^2) = 0.3$ in equation (1)

we get,

$E(x^2) - 0.25$

$\Rightarrow 0.3 - 0.25$

$= 0.05$ //

Q. Let x be a continuous random variable with pdf given by,

$$f(x) = ax \; ; \; 0 \le x \le 1$$
$$= a \; ; \; 1 \le x \le 2$$
$$= -ax + 3a \; ; \; 2 \le x \le 3$$
$$= 0, \text{ elsewhere.}$$

Find the value of a, mean and varience.

Soln

Since the total probability = 1

$$\therefore \int_0^1 f(x)dx + \int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 (ax)dx + \int_1^2 a\,dx + \int_2^3 (-ax+3a)\,dx = 1$$

$$\Rightarrow a\int_0^1 x\,dx + a\int_1^2 1\,dx - a\int_2^3 x\,dx + 3a\int 1\,dx = 1$$

$$\cancel{\Rightarrow a\int_0^1 x\,dx + a\int_1^2 1\,dx - a\int_2^3 x\,dx} +$$

$$\Rightarrow a\left[\frac{x^2}{2}\right]_0^1 + a\left[x\right]_1^2 - a\left[\frac{x^2}{2}\right]_2^3 + 3a\left[x\right]_2^3 = 1$$

$$\Rightarrow a\cdot\frac{1}{2} + a\left[2-1\right] - a\left[\frac{3^2}{2} - \frac{2^2}{2}\right] + 3a\left[3-2\right] = 1$$

$$\Rightarrow \frac{a}{2} + a - a\left[\frac{5}{2}\right] + 3a\left[1\right] = 1$$

$$\Rightarrow \frac{a}{2} + a - \frac{5a}{2} + 3a = 1$$

$$\Rightarrow \frac{a+2a-5a+6a}{2} = 1$$

$$\Rightarrow \frac{\cancel{4}^{2}a}{\cancel{2}} = 1$$

$$\Rightarrow 2a = 1$$

$$\Rightarrow a = \frac{1}{2}$$

Again, by the definition of mean,

$$E(X) = \int_0^1 x \cdot f(x)dx + \int_1^2 x \cdot f(x)dx) + \int_2^3 x \cdot f(x)\,dx$$

$$= \int_0^1 x \cdot ax\,dx + \int_1^2 x \cdot a\,dx + \int_2^3 x \cdot (-ax+3a)\,dx$$

$$= \frac{1}{2}\int_0^1 x^2\,dx + \frac{1}{2}\int_1^2 x\,dx - \frac{1}{2}\int_2^3 x^2 + \frac{3}{2}\int_2^3 x\,dx$$

$$= \frac{1}{2}\left[\frac{x^3}{3}\right]_0^1 + \frac{1}{2}\left[\frac{x^2}{2}\right]_1^2 - \frac{1}{2}\left[\frac{x^3}{3}\right]_2^3 + \frac{3}{2}\left[\frac{x^2}{2}\right]_2^3$$

$$= \frac{1}{2}\cdot\frac{1}{3} + \frac{1}{2}\left[\frac{4}{2} - \frac{1}{2}\right] - \frac{1}{2}\left[\frac{27}{3} - \frac{8}{3}\right] + \frac{3}{2}\left[\frac{9}{2} - \frac{4}{2}\right]$$

$$= \frac{1}{6} + \frac{3}{4} - \frac{19}{6} + \frac{15}{4}$$

$$= \frac{2+9-38+45}{12}$$

$$= 1.5$$

Again,
By definition of variance,

$$V(X) = E(X^2) - \{E(X)\}^2$$

$$= E(X^2) - (1.5)^2$$

$$= E(X^2) - 2.25 \qquad (1)$$

Now,

$$E(X^2) = \int_0^1 x^2 ax\,dx + \int_1^2 x^2 \cdot a\,dx + \int_2^3 x^2(-ax+3a)\,dx$$

$$= \frac{1}{2}\int_0^1 x^3 dx + \frac{1}{2}\int_1^2 x^2\,dx - \frac{1}{2}\int_2^3 x^3\,dx + \frac{3}{2}\int_2^3 x^2 dx$$

$$= \frac{1}{2}\left[\frac{x^4}{4}\right]_0^1 + \frac{1}{2}\left[\frac{x^3}{3}\right]_1^2 - \frac{1}{2}\left[\frac{x^4}{4}\right]_2^3 + \frac{3}{2}\left[\frac{x^3}{3}\right]_2^3$$

$$= \frac{1}{2}\cdot\frac{1}{4} + \frac{1}{2}\left[\frac{8}{3} - \frac{1}{3}\right] - \frac{1}{2}\left[\frac{81}{4} - \frac{16}{4}\right] + \frac{3}{2}\left[\frac{27}{3} - \frac{9}{3}\right]$$

$$= \frac{1}{8} + \frac{1}{2}\left[\frac{7}{3}\right] - \frac{1}{2}\left[\frac{65}{4}\right] + \frac{3}{2}\left[\frac{18}{3}\right]$$

$$= \frac{1}{8} + \frac{7}{6} - \frac{65}{8} + \frac{57}{6}$$

$$= 0.125 + 1.17 - 8.125 + 9.5$$

$$= 2.67$$

Now putting the value of value of $E(x^2) = 2.67$ in eq^n (1) $\Rightarrow$

$$v(X) = 2.67 - 2.25$$

$$= 0.42$$

11. Moment generating Function.

Discrete:

If X is a discrete random variable with pmf $p(x)$ then moment generating function of x is denoted by $M_X(t)$ and is defined as –

$$M_X(t) = E\left(e^{tX}\right)$$
$$= \sum e^{tx} . p(x)$$

Continuous:

If X is a continuous random variable with pdf $f(x)$ then moment generating function of x is denoted by $M_X(t)$ and is defined by –

$$M_X(t) = E\left(e^{tX}\right)$$
$$= \int_{-\infty}^{\infty} e^{tx} . f(x)\, dx .$$

Property:

(1) Moment generating function of the sum of independent random variables is equall to the product of their individual moment generating functions.

Symbolically, if $x_1, x_2, \ldots x_n$ are 'n' independent random variables, then according to the property of moment generating function,

$$M_{(X_1+X_2+\ldots+X_n)}(t) = M_{X_1}(t) . M_{X_2}(t) . \ldots . M_{X_n}(t)$$

Proof:

$$L.H.S = M_{(x_1+x_2+\cdots x_n)}(t)$$

$$= E\left\{e^{t(x_1+x_2+\cdots x_n)}\ t\right\}$$

$$= E\left\{e^{x_1t+x_2t+\cdots x_nt}\right\}$$

$$= E\left\{e^{x_1t}.\ e^{x_2t}.\cdots e^{x_nt}\right\}$$

$$= E\left(e^{x_1t}\right).\ E\left(e^{x_2t}\right).\cdots E\left(e^{x_nt}\right)$$

Since, $x_1, x_2 \cdots x_n$ are independent random variable

$$\therefore M_{x_1}(t).\ M_{x_2}(t).\cdots.\ M_{x_n}(t)$$

12. Theoritical Probability Distribution.

Probability Distribution → Discrete Probability Distribution.

Probability Distribution → Continuous Probability Distribution.

DISCRETE PROBABILITY DISTRIBUTION

12.1. Binomial Distribution.

Bernoulli Trials:

Trials having two and only two results are called Bernoulli trials. The concept of Bernoulli was developed by prof. James Bernoulli.

Binomial Distribution:

A discrete random variable x is said to follow a binomial distribution if its pmf is given by,

$$P(X=x) = {}^{n}C_{x} \cdot p^{x} \cdot q \quad , \quad x = 0, 1, 2, \ldots, n$$

where, x = no. of success.

n = total no. of independent Bernoulli trials.

$n-x$ = No. of failure

p = probability of success.

q = probability of failure.

n and p are called the parameter of Binomial Distr

Note:
If x is following distribution with parameter n then it is symbolically represent as: $X \sim B(n,p)$

Condition/Assumption of Binomial Distribution:

(I) The trials must have two and only two results. i.e. the trails must be Bernoulli trials.

(II) The total number of Bernoulli trials is finite, say '

(III) All the Bernoulli trials are independent of each other.

(IV) The probability of success in each trial is constant.

Properties of Binomial Distribution:

(I) Binomial Distribution has two parameter namely: n an

(II) Mean and varience of Binomial Distribution $E(X)= np$ and $V(X)=$

(III) In case of Binomial Distribution mean is always greater than va

(IV) The moment generating function of a binomial distrib is,

$$M_X(t) = \{q + p.e^t\}^n$$

Moment Generating function of Binomial Distribution:

Let $X \sim B(n,p)$ then pmf of X is, $P(X=x) = {}^nC_x . p^x . q^{n-x}, x = 0,1,2,\ldots$

By definition,

$$M_X(t) = E(e^{tX})$$
$$= \sum_x e^{tx} . P(x)$$
$$= \sum e^{tx} . {}^nC_x . p^x . q^{n-x} \text{ (using equating [1])}$$
$$= (q^n + e^t . {}^nC_1 . p . q^{n-1} + e^{2t} . {}^nC_2 p^2 . q^{n-2} + \cdots + e^{nt} . p^n . 1)$$
$$= \{q^n + {}^nC_1 (p.e^t) . q^{n-1} + {}^nC_2 . (p.e^t)^2 . q^{n-2} + \cdots + {}^nC_n (p.e^t)^n$$
$$= \{q + p.e^t\}^n,$$ required moment G. function of binomial distribution.

binomial series:
$(a+b)^n = a^n + {}^nC_1 . b . a^{n-1} + {}^nC_2 . b^2 . a^{n-2} + \cdots + {}^nC_n . b^n . a^{n-n}$

Q. If x is following Binomial Distribution with $n=6$ and is also having a situation $9P(X=4) = P(X=2)$, then find the value of p.

Soln
Given that,

$$9P(X=4) = P(X=2)$$

we know that,

$$P(X=x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}$$

$$\Rightarrow 9P(X=4) = P(X=2)$$

$$\Rightarrow 9\left\{{}^{6}C_{4} \cdot p^{4} \cdot q^{6-4}\right\} = {}^{6}C_{2} \cdot p^{2} \cdot q^{6-2}$$

$$\Rightarrow 9\left\{\frac{\lfloor 6}{\lfloor 6-4 \cdot \lfloor 4} \cdot p^{4} \cdot q^{2}\right\} = \frac{\lfloor 6}{\lfloor 6-2 \cdot \lfloor 2} \cdot p^{2} \cdot q^{4}$$

$$\Rightarrow 9\, p^{4} q^{2} = p^{2} \cdot q^{4}$$

$$\Rightarrow 9\, \frac{p^{4} q^{2}}{p^{2} q^{4}} = 1$$

$$\Rightarrow 9\, \frac{p^{2}}{q^{2}} = 1$$

$$\Rightarrow 9p^{2} = q^{2}$$

$$\Rightarrow 9p^{2} = (1-p)^{2}$$

$$\Rightarrow 9p^{2} = 1 - 2p + p^{2}$$

$$\Rightarrow 8p^{2} + 2p - 1 = 0$$

$$\Rightarrow 8p^{2} + 4p - 2p - 1 = 0$$

$$\Rightarrow 4p(2p+1) - 1(2p+1) = 0$$

$$\Rightarrow (4p-1)(2p+1) = 0$$

$$\Rightarrow 4p - 1 = 0 \quad \text{or} \quad 2p + 1 = 0$$

$$\Rightarrow p = 1/4 \qquad \Rightarrow p = -1/2 \ (-ve)$$

$$\therefore p = 1/4$$

Q. If X is a discrete random variable following binomial distribution with mean 2.4 and variance 1.44. then find the following:

(i) Find the value of n

(ii) Find $P(x \geq 3)$.

Soln

Given that,

$$\text{mean}(np) = 2.4$$

$$\text{varience}(npq) = 1.44$$

(i) ~~Now~~,

$X \sim B(n,p)$, so,

$$\text{Mean} \Rightarrow E(X) = np = 2.4 \quad \text{——(1)}$$

and $\text{varian} \Rightarrow V(X) = npq = 1.44$

$$\therefore np.q = 1.44$$

$$\Rightarrow (2.4).q = 1.44$$

$$\Rightarrow q = \frac{1.44}{2.4}$$

$$= 0.6$$

Again we know that,

$$p = 1-q$$

$$= 1-0.6$$

$$= 0.4$$

Now putting the value of $p=0.4$, in equation (1), we get,

$$np = 2.4$$

$$\Rightarrow n(.4) = 2.4$$

$$\Rightarrow n = \frac{2.4}{0.4}$$

$$= 6$$

(ii) $P(X \geq 3) = P(X=3) + P(X=4) + P(X=5) + P(X=6)$

$n = 6$
$p = 0.4$
$q = .6$

$= {}^6C_3 p^3 q^3 + {}^6C_4 p^4 q^2 + {}^6C_5 p^5 q^1 + {}^6C_6 p^6 q^0$

$= \frac{\lfloor 6}{\lfloor 6-3 \lfloor 3} p^3 q^3 + \frac{\lfloor 6}{\lfloor 6-4 \lfloor 4} p^4 q^2 + \frac{\lfloor 6}{\lfloor 6-5 \lfloor 5} p^5 q + \frac{\lfloor 6}{\lfloor 6-6 \lfloor 6} p^6 q^0$

$= \frac{\lfloor 6}{\lfloor 3 \lfloor 3} p^3 q^3 + \frac{\lfloor 6}{\lfloor 2 . \lfloor 4} p^4 q^2 + \frac{\lfloor 6}{\lfloor 1 \lfloor 5} p^5 q + \frac{\lfloor 6}{\lfloor 0 \lfloor 6} p^6 . 1$

$= \frac{6\times5\times4\times3\times2\times1}{3\times2\times1 . 3\times2\times1} p^3 q^3 + \frac{6\times5\times4\times3\times2\times1}{2\times1\times4\times3\times2\times1} p^4 q^4 + \frac{6\times5\times4\times3\times2\times1}{1\times5\times4\times3\times2\times1} p^5 q + \frac{6\times5\times4\times3\times2\times1}{1\times6\times5\times4\times3\times2\times1} p^6$

$= 20 p^3 q^3 + 15 p^4 q^2 + 6 p^5 q + p^6$

$= 20(.4^3)(.6^3) + 15(.4^4)(.6^2) + 6(.4^5)(.6) + (.4^6)$

$= 20(0.064)(0.216) + 15(0.0256)(0.36) + 6(0.01024)(0.6) + 0.004096$

$= 0.27648 + 0.13824 + 0.036864 + 0.004096$

$= 0.45568$ //

$\boxed{\text{OR}}$

$P(X \geq 3) = 1 - P(X=0) - P(X=1) - P(X=2)$

$= 1 - {}^6C_0 p^0 q^6 - {}^6C_1 p^1 q^5 - {}^6C_2 p^2 q^4$

$= 1 - \frac{\lfloor 6}{\lfloor 6} q^6 - \frac{\lfloor 6}{\lfloor 5} p q^5 - \frac{\lfloor 6}{\lfloor 4 \lfloor 2} p^2 q^4$

$= 1 - q^6 - 6 p q^5 - 15 p^2 q^4$

$= 1 - (.6)^6 - 6(.4)(.6)^5 - 15(.4)^2(.6)^4$

$= 1 - 0.046656 - 6(.4)(0.07776) - 15(0.16)(0.1296)$

$= 1 - 0.046656 - 0.186624 - 0.31103$

$= 0.45568$ //

. The probability of a man heating a target is $\frac{1}{4}$. If he fires 7 times what is the probability of his hitting the target atleast 3.

Soln

Given that,

man hitting a target $= \frac{1}{4}$

$$\therefore p = \frac{1}{4} = 0.25$$

$$\therefore q = 1 - p$$
$$= 1 - \frac{1}{4}$$
$$= \frac{3}{4}$$
$$= 0.75$$

$$n = 7$$

Now,

The man fires 7 times, so,

$$n = 7$$

$$\cancel{\therefore p(x \geq 3) = p(x \geq 4) = 1 - p(x \leq 3)}$$

$$\cancel{\Rightarrow p(x = x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{n-x}}$$

$$\Rightarrow \cancel{1 - p(0) + p(1) + p(2) + p(3)}$$

$$= \cancel{1}\ {}^{7}C$$

$$\therefore p(x \geq 3) = p(x=3) + p(x=4) + p(x=5) + p(x=6) + p(x=7)$$

$$= {}^{7}C_{3} \cdot p^{3} \cdot q^{4} + {}^{7}C_{4} p^{4} q^{3} + {}^{7}C_{5} \cdot p^{5} \cdot q^{2} + {}^{7}C_{6} \cdot p^{6} \cdot q^{1} + {}^{7}C$$

$$= \frac{\lfloor 7}{\lfloor 7-3 \cdot \lfloor 3} p^{3} q^{4} + \frac{\lfloor 7}{\lfloor 7-4 \cdot \lfloor 4} \cdot p^{4} q^{3} + \frac{\lfloor 7}{\lfloor 7-5 \cdot \lfloor 5} p^{5} \cdot q^{2} + \frac{\lfloor 7}{\lfloor 7-6 \cdot \lfloor 6} p^{6} \cdot q^{1} +$$

$$= \frac{\lfloor 7}{\lfloor 4 \cdot \lfloor 3} \cdot p^{3} \cdot q^{4} + \frac{\lfloor 7}{\lfloor 3 \cdot \lfloor 4} p^{4} q^{3} + \frac{\lfloor 7}{\lfloor 2 \cdot \lfloor 5} p^{5} \cdot q^{2} + \frac{\lfloor 7}{\lfloor 1 \cdot \lfloor 6} p^{6} \cdot q^{1} + \frac{\lfloor}{\lfloor}$$

$$= \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \times 3 \times 2 \times 1} p^{3} \cdot q^{4} + \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 4 \times 3 \times 2 \times 1} p^{4} \cdot q$$

$$+\frac{7\times\overset{3}{6}\times5\times4\times3\times2\times1}{2\times1\times5\times4\times3\times2\times1}\cdot p^5q^2+\frac{7\times6\times5\times4\times3\times2\times1}{1\times6\times5\times4\times3\times2\times1}\cdot p^6\cdot q^1+\frac{7\times6\times5\times4\times3\times2\times1}{1\times7\times6\times5\times4\times3\times2\times1}p^7\cdot1$$

$$=35\cdot p^3q^4+35p^4q^3+21p^5q^2+7p^6q+p^7$$

$$=35(\cdot25^3)(\cdot75^4)+35(\cdot25^4)(\cdot75^3)+21(\cdot25^5)(\cdot75^2)+7(\cdot25^6)(\cdot75)+(\cdot25^7)$$

$$=35(0.016)(\cdot316)+35(\cdot0039)(\cdot422)+21(\cdot00098)(\cdot563)+7(\cdot00024)(\cdot75)+(\cdot000061)$$

$$=0.177+0.058+0.012+0.00126+0.000061$$

$$=0.248321$$

OR

$$P(X\geq3)=1-P(X=0)-P(X=1)-P(X=2)$$

$$=1-{}^7C_0p^0q^7-{}^7C_1p^1q^6-{}^7C_2p^2q^5$$

$$=1-\frac{\lfloor7}{\lfloor7}q^7-\frac{\lfloor7}{\lfloor6\cdot1}pq^6-\frac{\lfloor7}{\lfloor5\lfloor2}p^2q^5$$

$$=1-q^7-7pq^6-21p^2q^5$$

$$=1-(\cdot75)^7-7(\cdot25)(\cdot75)^6-21(\cdot25)^2(\cdot75)^5$$

$$=1-0\cdot133484-0\cdot311462-21(0.0625)(0.2373)$$

$$=1-0\cdot133484-0\cdot311462-0\cdot31146$$

$$=0\cdot242238$$

Poisson Distribution

Popularly Known as : Distribution of RARE EVENTS

A discrete random variable x is said to follow poisson distribution if its pmf. is given by:-

$$p(x=x) = \frac{e^{-\lambda}\lambda^{x}}{\lfloor x}, \quad x = 0, 1, 2, \ldots, \infty.$$

n≥50 consid

where,

λ = The parameter of poisson distribution.

Condition/Assumption of Poisson Distribution:

(i) 'n' the number of independent Bernoulli trials must b sufficiently large. i.e., 'n' $\to \infty$

(ii) 'p' the probability of success for each trials must be very small, i.e., 'p' $\to 0$

(iii) The changes of occurance of the events under poisso distribution are very rare.

(iv) Mean and varience must be equal each other,

$$E(x) = var(x).$$

~~Moment Generating function of~~

properties of poisson Distribution :

(I) Poisson distribution is a single parametric distribution with the parameter λ.

(II) In poisson distribution mean and varience are equall and denoted by λ.

(III) Number of trails (n) in poisson distribution is infinite.

(IV) Probability of success in case of poisson distribution is very low

(V) Poisson distribution is used for rare events.

(VI) The moment generating function of poisson distribution is, $M_X(t) = e^{\lambda(e^t - 1)}$.

Moment Generating Function of POISSON DISTRIBUTION

Let, $X \sim P(\lambda)$, then pmf of X is,

$$P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{\lfloor x}, \quad x = 0,1,2,\dots\infty \qquad \text{——(1)}$$

Now,
The definition of moment generating function is,-

$$M_X(t) = E(e^{tx})$$

$$= \sum_{n=1}^{\infty} e^{tx} \cdot P(x)$$

$$= \sum_{\infty} e^{tx} \cdot \frac{e^{-\lambda} \cdot \lambda^x}{\lfloor x} \qquad \left[\text{from eq}^n (1)\right]$$

$$= e^{-\lambda} \sum \frac{e^{tx} \cdot \lambda^x}{\lfloor x}$$

$$= e^{-\lambda}\left\{\frac{\lambda^0 e^{t(0)}}{\lfloor 0} + \frac{\lambda^1 e^{t(1)}}{\lfloor 1} + \frac{\lambda^2 e^{t(2)}}{\lfloor 2} + \frac{\lambda^3 e^{t(3)}}{\lfloor 3} + \cdots\right\}$$

$$= e^{-\lambda}\left\{\frac{1}{\lfloor 0} + \frac{\lambda e^{t}}{\lfloor 1} + \frac{\lambda^2 e^{2t}}{\lfloor 2} + \frac{\lambda^3 e^{3t}}{\lfloor 3} + \cdots\right\}$$

$$= e^{-\lambda}\left[1 + \frac{\lambda . e^{t}}{\lfloor 1} + \frac{\lambda^2 . e^{2t}}{\lfloor 2} + \frac{\lambda^3 . e^{3t}}{\lfloor 3} + \cdots\right\}$$

$$= e^{-\lambda}\left\{1 + \frac{\lambda . e^{t}}{\lfloor 1} + \frac{(\lambda e^{t})^2}{\lfloor 2} + \frac{(\lambda . e^{t})^3}{\lfloor 3} + \cdots\right\}$$

$$= e^{-\lambda} . e^{\lambda . e^{t}}$$

$$= e^{-\lambda + \lambda . e^{t}}$$

$$= e^{\lambda(-1+e^{t})}$$

$$= e^{\lambda(e^{t}-1)}$$, which is the required Moment generating function of Poisson Distribution.

$\boxed{\text{Note}}$

Exponential Series :

$$e^{x} = 1 + \frac{x}{\lfloor 1} + \frac{x^2}{\lfloor 2} + \frac{x^3}{\lfloor 3} + \cdots$$

$$e^{-x} = 1 - \frac{x}{\lfloor 1} + \frac{x^2}{\lfloor 2} - \frac{x^3}{\lfloor 3} + \cdots$$

Q. If x follows Poisson Distribution with $n=40$ and $p=1/10$, then find the value of standard deviation as poisson distribution and also find the $p(1 \leq x \leq 2)$.

Soln

Given that,

$$n = 40$$
$$p = 1/10$$

Now,

variance of poisson distribution $= \lambda$

$\therefore$ standard deviation of poisson distribution $= \sqrt{\lambda}$

since,

poisson distribution is derive as a limiting form of Binomial Distribution, so value of λ can be determine as,

$$\lambda = n.p$$
$$= 40 \cdot 1/10$$
$$= 4$$

Now required standard deviation of poisson distribution is, $\sqrt{\lambda} = \sqrt{4} = 2$

Again, $p(1 \leq x \leq 2) = p(x=1) + p(x=2)$

$$= \frac{e^{-\lambda}.\lambda^{x}}{\lfloor x} + \frac{e^{-\lambda}.\lambda^{x}}{\lfloor x}$$

$$= \frac{e^{-4}.4^{1}}{\lfloor 1} + \frac{e^{-4}.4^{2}}{\lfloor 2}$$

$$= e^{-4}\left[\frac{4}{1} + \frac{16}{2}\right]$$

$$= \frac{1}{2.718^{4}}\left[4+8\right]$$

$$= \frac{12}{54.5755}$$

$$= 0.2199.$$

8. An automatic machine makes paper clips from coils of wire; on the average 1 in 400 paper clips is defective. If the paper clips are packed in boxes of 100 then determine the probability that any given box of clips will contain –

(i) No defective clip
(ii) 1 or more defective clip
(iii) Less than 2 defective clips.

Soln

Given that,

$n = 100$

$p = \frac{1}{400}$

$\boxed{e = 2.718 \quad e^{.25} = 0.779}$

$\therefore$ mean, $\lambda = n.p$

$= 100 \times 1/400$

$= 0.25$

(i) No defective clip :

$$P(x) = \frac{e^{-\lambda} . \lambda^{x}}{\lfloor x}$$

$$\Rightarrow P(0) = \frac{e^{-.25} . 0.25^{0}}{\lfloor 0}$$

$$= \frac{e^{-0.25} . 1}{1}$$

$$= e^{-0.25}$$

$$= \frac{1}{e^{.25}}$$

$$= \frac{1}{(2.718)^{.25}} \quad [\because e = 2.\cancel{178}718 \text{ approxemetely}]$$

$$= \frac{1}{1.284}$$

$$= 0.779$$

* $e = 2.718$
* $e^{.25} = 1.284$

(ii) <u>one or more defective Clips :</u>

$$p(x \geq 1) = 1 - p(x=0)$$

$$= 1 - \frac{e^{-\lambda} \cdot \lambda^{x}}{\lfloor x}$$

$$= 1 - \frac{e^{-.25} \cdot .25^{0}}{\lfloor 0}$$

$$= 1 - \frac{e^{-.25} \cdot 1}{1}$$

$$= 1 - \frac{1}{e^{.25}}$$

$$= 1 - 0.779$$

$$= 0.221$$

(iii) <u>Less than 2 defective Clips:</u>

$$p(x<2) = p(0) + p(1)$$

$$= \left[\frac{e^{-\lambda} \cdot \lambda^{x}}{\lfloor x}\right] + \left[\frac{e^{-\lambda} \cdot \lambda^{x}}{\lfloor x}\right]$$

$$= \left[\frac{e^{-.25} \cdot .25^{0}}{\lfloor 0}\right] + \left[\frac{e^{-.25} \cdot .25^{1}}{\lfloor 1}\right]$$

$$\cancel{= 0.779 + \frac{\quad}{1 \times}}$$

$$= e^{-.25}\left[\frac{1}{1} + \frac{.25}{1}\right]$$

$$= e^{-.25}(1 + .25)$$

$$= \frac{1}{2.718^{.25}} \times 1.25$$

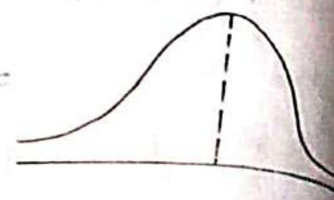
$$= \frac{1.25}{1.284}$$

$$= 0.9735$$

CONCEPT

* symmetrical Distribution :

A distribution is said to be symmetrical if its mean, median and mode are equal.



Mean = Median = Mode .

* Assymmetrical Distribution .

A distribution is said to be asymmetrical if its mean, median and mode are not equal.

TYPE : → (I) Positive asymmetrical distribution,
→ (II) Negative asymmetrical Distribution,

positive asymmetrical Distribution:

A distribution is said to be positive asymmetrical or positively skewed, if:

Mean > Median > Mode .

Negative asymmetric distribution:

A distribution is said to be negative asymmetrical, if

Mean < Median < Mode .

NORMAL DISTRIBUTION

OR

GAUSSIAN DISTRIBUTION

Father of Normal Distribution

$\Rightarrow$ De. Moivre (1971)

A continuous random variable x taking the value within the interval $(-\infty, +\infty)$ is said to follow Normal Distribution with parameter μ (mean) and σ^2 (variance) if its pdf is given by,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty \leq x \leq \infty$$

$$\mu, \sigma^2 \geq 0$$

Note:

If x is following a Normal distribution then is symbolically represented as $X \sim N(\mu, \sigma^2)$

Book- Statistical Method N.G. DAS (V

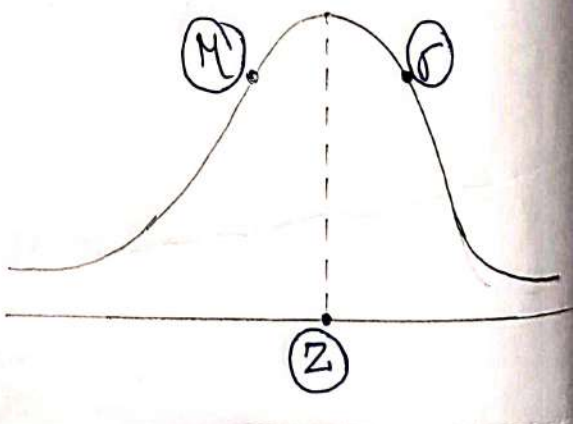
Standard Normal Distribution

If $X \sim N(\mu, \sigma^2)$, then a variable Z which is define as, $Z = \frac{x - \mu}{\sigma}$ is said to follow standard normal distribution with parameter 0 and 1, if its pdf is given by,

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}z^2} \quad , 0 \leq Z \leq 1$$

where, Z = Standard normal variate / variable.

Point of inflation.



Two inflati point —

$\mu \pm \sigma$

Q/.

let $\mu = 2$

$\sigma^2 = 25$

∴ point of inflation,

$\mu \pm \sigma$

$\Rightarrow 2 \pm 5$

$\Rightarrow 2 + 5$ and $2 - 5$

$\Rightarrow 7$ and -3

∴ At 7 → ~~in~~ shows increasing ~~rate~~ curve with increasing rate

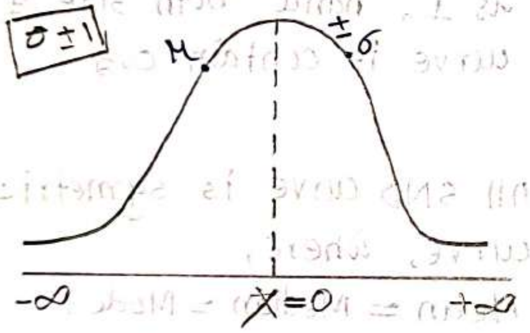
At -3 → shows ~~d~~ increase curve at diminishing rate.

* Normal curve	Standard Normal Distribution curve
PROPERTY	PROPERTY
(I) All Normal curve are bell-shaped with points of inflection at μ and $\pm\sigma$.	(I) All SND curve are bell-shaped with point of inflection at μ and $\pm\sigma$.
(II) All normall curve symmetric about the mean.	(II) All SND curve are symmetric about the mean.
(III) The area under an entire Normal curve is 1.	(III) The area under an entire SND curve is 1.
(IV) Though the entire NC is 1, hence, both side of the curve contain 0.5	(IV) Though the entire SND curve is 1, hance both side of the curve is contain 0.5.
(V) Normal curve is symetric curve, where, Mean = Median = Mode	(V) All SND curve is symetric curve, where, Mean = Median = Mode.
(VI) Normal curve satisfies asympotic property, i.e, the curve doesnot touch x-axis.	(VI) SND curve satisfies asympotic property, i.e., the curve does not touch X-axis.
(VII) In x axis of NC contains different values of $\boxed{X}$, and Y-axis contains different values of $f(x)$.	(VII) In x axis of SND curve contains different values of $\boxed{Z}$. And Y-axis contains different values of $f(Z)$.

Normal curve

property

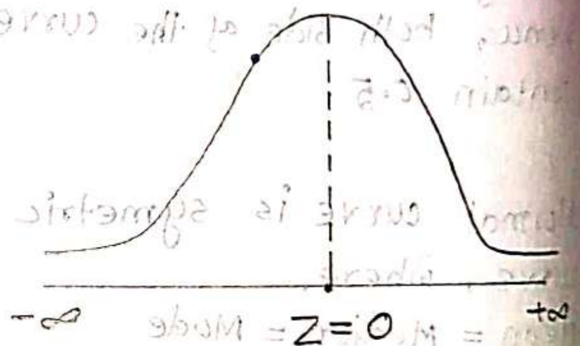
(VIII) Since $f(x)$ represents the probability of different values considered by ~~NC~~ Normal variable, NO portion of the NC will lie below the x- axis, because ~~f(x)~~ probability lies between 0 to 1, no negative.



SND curve

property

(VIII) Since $f(z)$ represents the probability of different values considered by ~~A~~ to SND variable, No portion of the SND curve will lie below the x-axis, because probability lies between 0 and 1, cannot be negative.



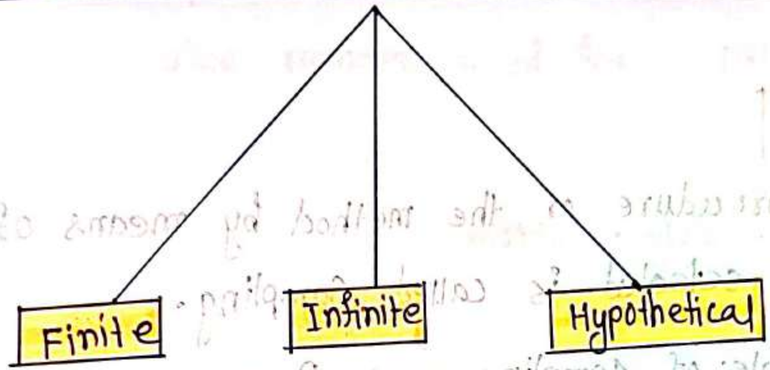
UNIT:3

SAMPLING AND ESTIMATION

1. Statistical population / universe.

A statistical population / universe is one which deal with the study of objects, finite or infinite in aggregate form.

CLASSIFICATION OF STATISTICAL POPULATION:



2. Human population are statistical population but the reverse is not truth.

Statistical population is a broader concept which include the counting of human population, different object population, any type of ~~count~~ statistical counting. Therefore human population is a part of statistical population.

But the human population indicates only the human, people counting. Rather than all other statistical population.

——— Therefore,
Human population are statistical population but the reverse is not truth.

3. Some concept.

SAMPLE:

Sample refers to a part of population which is selected and analyse for giving conclusion about the entire statistical population.

SAMPLING

The procedure or the method by means of which sample is selected is called Sampling.

Examples of sampling are- Purposive sampling, Samp Random Sampling, Statified Sampling, Systemetic samplin etc.

SAMPLING UNIT

It is the ultimate unit from which we gether necessary information.

SAMPLING FRAME

It refers to the construction of a list of questions that are supposed to be ask to the respondence by the investigeters.

Suppose $N=5$ (population) $= 10, 20, 30, 40, 50$

$n=3$ (sample) $= 10, 20, 30$

$\therefore$ mean, $\bar{x} = 60/3 = 20$ ($\bar{x} \rightarrow$ Sample mean)

- Population is not varies because it take entire population

POPULATION PARAMETER

Population parameter is a term which is use to denote various population characterstics, i.g., population mean (μ), population variance, population median, etc.

population parameter are usuall define as statistical constant whose value remain fixed for a particular population or structure.

In Practice, population parameter are unknown and they are estimated by statistics.

STATISTIC

Any statistical measure which is calculated on the basis of sample value is called statistic. Thus, a statistic is actually a function of sample value.

Example of statistic are, sample mean ($\bar{x}$), sample variance (s^2), etc. ~~vatu~~

Values of the statistic vary from sample to sample, and hance - it is called a variable.

4. Difference between : Parameter and Statistic.

Sl. No.	Parameter	Statistic
1.	Any function of the population values is called the parameter.	Any function of the sample observation is called statistics.
2.	The Parameter is an unknown constant.	The statistics does not contain unknown constant.
3.	Parameters are not used to estimate population characterstics.	statistics are used to estimate population characterstics.
4.	Parameters are free from sampling and other Errors.	statistics are subject to sampling and Non-Sampling error.
5.	There is no distribution of parameter.	statistics have distribution, called sampling distribution.

5. Why statistic is called variable.

(1) Statistic is called a variable because the value may vary between data units in a population, and may change in value over time. The value of statistic vary from sample to sample. (2)

or Statistic is called a variable because the value of statistic vary from sample to sample. The value of statistic may change over time.

7. Estimator vs Estimate.

An Estimator refl. to formula with help such estimate the value of population parameter and the value so obtain by using this formula is called estimate.

8. Unbiased estimator.

An estimator is said to be unbiased if its expected value is equal to the relevent population parameter value. Forexample, if θ denots a population parameter and $\hat{\theta}$ denotes its estimator then $\hat{\theta}$ will be called the unbiased estimator of θ if the following condition is satisfied:-

$$E(\hat{\theta}) = \theta$$

More specifically, sample mean $(\bar{x})$ will be called an unbiased estimator of population ~~parameter~~ mean (μ) if the following condition is satisfied,

$$E(\bar{x}) = \mu .$$

$$\boxed{\begin{array}{ll} * \text{ Bias :} & E(\hat{\theta}) \neq \theta \\ & E(\bar{x}) \neq \mu \end{array}}$$

FORMULA

1. Population variance $= \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

2. Sample variance $= s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

3. Population Mean square $= S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$

4. Sample Mean Square $= s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
   ↳ (called unbias)

---

5. Sample mean square (s'^2) is considered as an unbias estimator of population variance,

   i.e., $E(s'^2) = \sigma^2$

   This is true for small as well as large samples.

6. However, for large sample, i.e, for ($\forall$) $n \to \infty$, sample variance can be consider as an unbias estimator of population variance.

7. Thus, for small samples, sample variance is an Bias estimator of population variance.

Q. Show that, sample variance is an unbias estimator of population variance, if sample size is large.

Soln

By definition,

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

$$\Rightarrow ns^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ ——— (1)}$$

Again,

Sample mean square $= s'^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

$$\Rightarrow (n-1)s'^2 = \sum (x_i - \bar{x})^2 \text{ ——— (2)}$$

From eqn (1) and (2) $\Rightarrow ns^2 = (n-1)s'^2$

$$\Rightarrow s^2 = \left(\frac{n-1}{n}\right) s'^2$$

$$= \left(1 - \frac{1}{n}\right) s'^2 \text{ ——— (3)}$$

Since,
according to the question, sample size is large,
i.e., $n \to \infty$ and hence $\frac{1}{n} \to 0$
using above condition in eqn (3),

$$s^2 \simeq s'^2 \text{ ——— (4)}$$

We have,

$$E(s'^2) = \sigma^2$$

$$\Rightarrow E(s^2) = \sigma^2 \quad [\text{using eq}^n (4)]$$

Thus sample variance is an unbias estimator of population variance.

9. Method of Estimation

Point Estimation

Interval Estimation

Different Method of Point Estimation :—

(i) Method of moments

(ii) Least square method

(iii) Maximum likelyhood estimation method.

(iv) Generalised method of ~~Esti~~ Moments.

properties/Characterstics of good estimator / ideal estimator.

(1) Unbiasness Property : $E(\hat{\theta}) = \theta$ (Small sample property)

(ii) consistency property : $\lim_{n \to \infty} E(\hat{\theta}) = \theta$ → unbias

(Large sample property.)

NB: * A consistent estimator does not always goes in unbias.

(iii) Efficiency property: According to efficiency property an estimator is said to be efficient if the value of its variant is lower in relation to that of other estimator.

Forexample, $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ is three estimator.

here, $\hat{\theta}_1 < \hat{\theta}_2 < \hat{\theta}_3$, here $\hat{\theta}_1$ is Efficient.

(iv) Property of sufficient:

An estimator is sufficient if it uses all the sample information.

Mean, median and Mode.

10. Sampling Error v/s Non- Sampling Error.

SAMPLING ERROR: ⇒ (Root in sample survey)

Sampling Error/ Biase stands for the Errors arising due to the sampling procedure followed by the statistians. To be more elaborate, actually population parameters are unknown and in practice they are estimated by the statistics. Since, under sampling theory only a part of the population is studied, the value of the statistic is never going to be exactly equal to the population parameter values and thus some differences between them are bound to be there. This difference in sampling theory represents Sampling Error or Sampling Biase.

For example, if $\bar{x}$ denots the sample mean, which is a statistic and μ denots population mean, which is a parameter, then the absolute difference $|\bar{x}-\mu|$ represents Sampling Error.

* $|Z| = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
* Sampling Error: $\bar{x}-\mu$
* Standard Error: $\dfrac{\sigma}{\sqrt{n}}$

$\therefore$ Test Error $= \dfrac{\text{Sampling Error}}{\text{Standard Error}}$

$= \dfrac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

NON-SAMPLING ERROR: → Both in sample survey & census
→ Human Error

Non- Sampling Error are likely to present both census and Sample Survey. These error mainly arise due to the mistakes committed by investigators and statistician at different stages of the survays like- at the time of Data Collection, Coding of variables, entering the data, etc.

11. Sampling Distribution of a Statistic.

The value of a statistic are likely to change from sample to sample. since they are based on sample values. These different values of a statistic can be arranged with the form of a probability distribution showing values of the statistic along with their respective probabilities. Thus, sampling distribution simply show different values of the statistics along with their respective probabilities.

■ Parameter v/s population mean.

parameter doesn't have sampling Distribution. But Population mean have Sampling Distribution.

12. Standard Error of Sampling Distribution of the Statistic.

The standard deviation of the sampling distribution of statistic is called standard Error. It shows the extend of variability of different values of the statistic from their population mean/unbias estimator of population mean.

In case of simple random sampling with replacement (SRSWR), standard Error of ~~Stan~~ Statistic say, standard sample mean $(\bar{x})$ is given by, –

SRSWR

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$, where σ is the population SD.

In case of simple random sampling without replacemen (SRSWOR), standard error of sample mean is given by,

SRSWOR

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} - \frac{N-n}{N-1}$$

where, σ is the population standard deviation.

n is the sample size.

N is the population size.

$\frac{N-n}{N-1}$ is the Sampling Fraction.

13. Are Standard deviation and Standard Error are different ? Justify your answer.

Yes.

The SD reflects variability within a sample, while the Standard Error (SE) estimates the variability accross samples of a population.

$\boxed{\text{Similarity:}}$ SD and SE are both measures of variability.

14. ~~Def~~ Usefulness of Standard Error in Statistics.

Standard Error is used to estimate the efficiency, accuracy and consistency of a sample. In other word, SE measures how precisely a sampling distribution represents a population. It can be applied in statistics and economics.

15. Various sampling methods of collecting sample.

- Probability / Random Sampling
  Eg. Statified Random sampling
- Non-Probability / Non-Random Sampling
- Mixed Sampling

FORMULA

1. Samples of Size 'n' from a population of size 'N' in case of SRSWR can be drawn in $\boxed{N^n}$ ways. In case of SRSWOR can be drawn in $\boxed{{}^{N}C_{N}}$ ways

$$* \text{SRSWR} = N^n$$
$$* \text{SRSWOR} = {}^{N}C_n$$

2. Sample Mean $= \bar{x} = \frac{1}{n} \sum x_i$, $i = 1, 2, \ldots n$

3. Population Mean $= \mu = \frac{1}{N} \sum X_i$, $i = 1, 2, \ldots N$

4. Population variance $= \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2$

5. sample variance $= s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

6. Population Mean Square $= S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \mu)^2$

7. Sample mean Square $= s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

8. Mean of the sampling distribution of sample mean

$$= E(\bar{x}_i) = \frac{1}{N^n} \sum_{i=1}^{n} \bar{x}_i \quad \Big| \text{ in case of SRSWR}$$

OR

$$\frac{1}{{}^{N}C_n} \sum_{i=1}^{n} \bar{x}_i \quad \Big| \text{ in case of SRSWOR}$$

9. Variance of the sampling distribution of sample mean

$$= V(\bar{x}) = \frac{1}{N^n} \sum_{i=1}^{n} \left\{ \bar{x} - E(\bar{x}) \right\}^2 \quad \Big| \text{ SRSWR}$$

OR

$$= \frac{1}{{}^{N}C_n} \sum_{i=1}^{n} \left\{ \bar{x} - E(\bar{x}) \right\}^2 \quad \Big| \text{ SRSWOR}$$

14. **Property:** Show that in case of simple random sampling, Mean is unbias estimator of population mean.

Proof:

Let us consider a population of size N where X_i $(i=1,2,\dots N)$ denots verious population units. Let us also consider a sample of size n which would be drawn from the population by simple Random sampling and $x_i(i=1,2,\dots n)$ denots verious sample units.

Since, x_i would be drawn from the population population of X_i units by SRS, it actually indicates that x_i can take only one of X_i values, each with an equal probability.

Thus, the probability distribution of x_i will be as follows –

x_i	X_1	X_2	-------	X_N
$P(x_i)$	$1/N$	$1/N$	----	$1/N$

Now
by definition,

$$E(x_i) = \sum_{i=1}^{n} x_i \, P(x_i)$$

$$= X_1 \frac{1}{N} + X_2 \frac{1}{N} + \dots X_n \frac{1}{N}$$

$$= \frac{1}{N}(X_1 + X_2 + \dots X_N)$$

$$= \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$= \mu \quad \text{———} \quad (1)$$

Now,

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

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \frac{1}{\cancel{n}} \cdot \cancel{n}\mu$$

$$= \mu$$

proved

Q/ From a population of 4 units namely 2, 4, 6 and 8 draw all possible samples of size 2 in case 'SRSWOR' ~~draw~~ and calculate their sample mean,

(i). Find the mean of the sampling distribution of sample mean.

(ii). And verify that sample mean is an unbias estimator of population mean.

(iii). Also calculate the value of standard error.

(iv). Also constract the sampling distribution of sample mean.

Sol^n

Here,

$N = 4$

$n = 2$

From the population of 4 units, samples of size 2 can be drawn in ${}^4C_2 = 6$.

Now, all these possible samples along with their sample mean are represented with the following table:

Sl. NO	Sample values	Sample mean
1	$x_1 = (2,4)$	$\bar{x}_1 = 3$
2	$x_2 = (2,6)$	$\bar{x}_2 = 4$
3	$x_3 = (2,8)$	$\bar{x}_3 = 5$
4	$x_4 = (4,6)$	$\bar{x}_4 = 5$
5	$x_5 = (4,8)$	$\bar{x}_5 = 6$
6	$x_6 = (6,8)$	$\bar{x}_6 = 7$

$$\sum_{i=1}^{6} \bar{x}_i = 30$$

Now,

(I) By definition,

$$E(\bar{x}) = \frac{1}{^NC_n} \sum_{i=1}^{6} \bar{x}_i$$

$$= \frac{1}{6} \times 30$$

$$= 5$$

(II) Population mean $= \mu = \frac{1}{N} \sum_{i=1}^{N} X_i$

$$= \frac{1}{4}(2+4+6+8)$$

$$= \frac{1}{4} \times 20$$

$$= 5$$

Thus the $E(\bar{x}) = \mu = 5$, and hance sample mean $(\bar{x})$ is an unbias estimator of population mean.

proved.

(iii) By the definition of standard error of sample mean,

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{N-n}{N-1} \quad \text{.............} (1)$$

Here,

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$= \frac{1}{4} \left\{ (x_1-\mu)^2 + (x_2-\mu)^2 + (x_3-\mu)^2 + (x_4-\mu)^2 \right\}$$

$$= \frac{1}{4} \left\{ (2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 \right\}$$

$$= \frac{1}{4} \left\{ 9 + 1 + 1 + 9 \right\}$$

$$= \frac{1}{4} \times 20$$

$$= 5$$

$$\therefore \sigma = \sqrt{5}$$

$$= 2.24$$

Now from equation (1) $\Rightarrow$ $SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \frac{N-n}{N-1}$

$$= \frac{2.24}{\sqrt{2}} \cdot \frac{4-2}{4-1}$$

$$= \frac{2.24}{1.41} \cdot \frac{2}{3}$$

$$= \frac{4.48}{4.23}$$

$$= 1.059$$

1. What is Analysis of variance (ANOVA)

The term analysis of variance popularly known as ANOVA was introduced by prof. Ronald. A. Fischer. It is defined as the process by means of which we are in a position to split the total amount of variations in a data set into two categories:

(I) Variation outside.

(II) Variation inside

It involve the investigation of the effects of one treatment variable on an interval-scaled dependent variable.

Assumption:

(I) All the sample observation must be independent.

(II) All the effect under ANOVA must be additive in nature.

(III) The error term is independently and normally distributed with zero mean (M) and variane (σ^2).

Use/ Application:

(I) ANOVA is used to test the significance of the difference between several population mean.

(II) It is use to test the overall significance of a regression model.

2) Is ANOVA superior to Students' test? / is ANOVA better than t-test.

Yes,
ANOVA is superior to Student's t test.

Because, the Student's t test is use to compare the mean between two group, whereas ANOVA is used to compare the mean among three or more groups.

3) Difference. ANOVA and Analysis of covariance (ANCOVA)

ANOVA	ANCOVA
1) use both linear and Non-linear model.	use only linear model.
2) It include only categorical variable	include both categorical and interval variable
3) It ignore covariate.	It consider covariate
4) It is the analysis of variane	It is the analysis of covariance

• What is parametric test ?

Parametric is a test in which parameters are assumed and the population distribution is always known

• Non parametric test ?

Non parametric test are experiments that donot require the underlying population for assumption.

• Advantage of Non Parametric test over parametric test.

(1) Do not depend on any distribution

(2) They are useful in analyzing data.

(3) They are relatively simple to calculate.

(4) When sample size is small then it is the only method for calculation.