2nd SEMESTER B.Sc.PSYCHOLOGY

CALICUT UNIVERSITY

REGRESSION ANALYSIS AND PROBABILITY THEORY

Prepared by

Suhaila.N.P (Assistant professor)

Shafeekha chungath (Assistant professor)

PG DEPARTMENT OF COMMERCE

CPA College of Global of Studies, Puthanathani

SYLLABUS

STA 2C 02- REGRESSION ANALYSIS AND PROBABILITY THEORY

Time: 6 Hours per week Credits: 4
Internal 20: External 80

Module 1: *Bivariate data*- relationship of variables, correlation analysis, methods of studying correlation, Scatter Diagram, Karl Pearson's Coefficient of Correlation, Calculation of Correlation from a 2-way table, Interpretation of Correlation Coefficient, Rank Correlation

11 Hours

Module 2: *Regression analysis*- linear regression, Regression Equation, Identifying the Regression Lines properties of regression coefficients, numerical problems

9 Hours

Module 3: *Partial and Multiple Correlation Coefficients*- Multiple Regression Equation, Interpretation of Multiple Regression Coefficients (three variable cases only)

16 Hours

Module 4: *Basic probability*- Sets, Union, Intersection, Complement of Sets, Sample Space, Events, Classical, Frequency and Axiomatic Approaches to Probability, Addition and Multiplication Theorems, Independence of Events (Up-to three events)

20 Hours

Module 5: *Random Variables and their probability distributions*- Discrete and Continuous Random Variables, Probability Mass Function, Distribution Function of a Discrete Random Variable

16 Hours

References

1. Gupta, S.P. *Statistical Methods*. Sultan Chand and Sons: New Delhi.
2. Gupta, S.C., &Kapoor, V.K. *Fundamentals of Applied Statistics*. New Delhi: Sultan Chand and Sons.
3. Garret, H.E., &Woodworth, R.S. *Statistics in Psychology and Education*. Bombay: Vakila, Feffex and Simens Ltd.
4. Mood, A.M., Graybill, F.A and Boes, D.C. *Introduction to Theory of Statistics*. 3rd Edition Paperback – International Edition.
5. Mukhopadhyay, P. *Mathematical Statistics*. New central Book Agency (P) Ltd: Calcutta.

Module-1

CORRELEATION ANALYSIS

Introduction:

- In practice, we may come across with lot of situations which need statistical analysis of either one or more variables. The data concerned with one variable only is called univariate data. For Example: Price, income, demand, production, weight, height marks etc are concerned with one variable only. The analysis of such data is called univariate analysis.
- The data concerned with two variables are called bivariate data. For example: rainfall and agriculture; income and consumption; price and demand; height and weight etc. The analysis of these two sets of data is called bivariate analysis.
- The date concerned with three or more variables are called multivariate date. For example: agricultural production is influenced by rainfall, quality of soil, fertilizer etc.

Definition:

Two or more variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

According to Simpson and Kafka, -Correlation analysis deals with the association between two or more variables‖.

Lun chou defines, -Correlation analysis attempts to determine the degree of relationship between variables‖.

Boddington states that -Whenever some definite connection exists between two or more groups or classes of series of data, there is said to be correlation.‖

Correlation Coefficient:

Correlation analysis is actually an attempt to find a numerical value to express the extent of relationship exists between two or more variables. The numerical measurement showing the degree of correlation between two or more variables is called correlation coefficient. Correlation coefficient ranges between -1

and +1.

SIGNIFICANCE OF CORRELATION ANALYSIS

Correlation analysis is of immense use in practical life because of the following reasons:

1. Correlation analysis helps us to find a single figure to measure the degree of relationshipexists between the variables.
2. Correlation analysis helps to understand the economic behavior
3. Correlation analysis enables the business executives to estimate cost, price and other variables.
4. Correlation analysis can be used as a basis for the study of regression. Once we know that two variables are closely related, we can estimate the value of one variable if the value of other is known.
5. Correlation analysis helps to reduce the range of uncertainty associated with decision making. The prediction based on correlation analysis is always near to reality.
6. It helps to know whether the correlation is significant or not. This is possible by comparing the correlation co-efficient with 6PE. It _r' is more than 6 PE, the correlation is significant.

Classification of Correlation

Correlation can be classified in different ways. The following are the most important classifications

1. Positive and Negative correlation
2. Simple, partial and multiple correlation
3. Linear and Non-linear correlation

Positive and Negative correlation

When the variables are varying in the same direction, it is called positive correlation. In other words, if an increase in the value of one variable is accompanied by an increase in the value of other variable or if a decrease in the

value of one variable is accompanied by a decree se in the value of other variable, it is called positive correlation.

Eg: 1)	A:	10	20	30	40	50
	B:	80	100	150	170	200

When the variables are moving in opposite direction, it is called negative correlation. In other words, if an increase in the value of one variable is accompanied by a decrease in the value of other variable or if a decrease in the value of one variable is accompanied by an increase in the value of other variable, it is called negative correlation.

Simple, Partial and Multiple correlation

Simple Correlation

In a correlation analysis, if only two variables are studied it is called simple correlation. Eg. the study of the relationship between price & demand, of a product or price and supply of a product is a problem of simple correlation.

Multiple correlation

In a correlation analysis, if three or more variables are studied simultaneously, it is called multiple correlations. For example, when we study the relationship between the yield of rice with both rainfall and fertilizer together, it is a problem of multiple correlation.

Partial correlation

In a correlation analysis, we recognize more than two variable, but consider one dependent variable and one independent variable and keeping the other Independent variables as constant. For example yield of rice is influenced b the amount of rainfall and the amount of fertilizer used. But if we study the correlation between yield of rice and the amount of rainfall by keeping the amount of fertilizers used as constant, it is a problem of partial correlation.

Linear and Non-linear correlation

Linear Correlation

In a correlation analysis, if the ratio of change between the two sets of variables is same, then itis called linear correlation.

For example when 10% increase in one variable is accompanied by 10% increase in the othervariable, it is the problem of linear correlation.

X:	10	15	30	60
Y:	50	75	150	300

Here the ratio of change between X and Y is the same. When we plot the data in graph paper, allthe plotted points would fall on a straight line.

Non-linear correlation

In a correlation analysis if the amount of change in one variable does not bring the same ratio ofchange in the other variable, it is called nonlinear correlation.

X:	2	4	6	10	15
Y:	8	10	18	22	26

Here the change in the value of X does not being the same proportionate change in the value of Y.

.

Degrees of correlation:

Correlation exists in various degrees

1. **Perfect positive correlation**

If an increase in the value of one variable is followed by the same proportion of increase in other related variable or if a decrease in the value of one variable is followed by the same proportion of decrease in other related variable, it is perfect positive correlation. eg: if 10% rise inprice of a commodity results in 10% rise in its supply, the correlation is perfectly positive. Similarly, if 5% full in price

results in 5% fall in supply, the correlation is perfectly positive.

2. **Perfect Negative correlation**

If an increase in the value of one variable is followed by the same proportion of decrease in other related variable or if a decrease in the value of one variable is followed by the same proportion of increase in other related variably it is Perfect Negative Correlation. For example if 10% rise in price results in 10% fall in its demand the correlation is perfectly negative. Similarly if 5% fall in price results in 5% increase in demand, the correlation is perfectly negative.

3. **Limited Degree of Positive correlation**:

When an increase in the value of one variable is followed by a non-proportional increase in other related variable, or when a decrease in the value of one variable is followed by a non- proportional decrease in other related variable, it is called limited degree of positive correlation.

For example, if 10% rise in price of a commodity results in 5% rise in its supply, it is limited degree of positive correlation. Similarly if 10% fall in price of a commodity results in 5% fall in its supply, it is limited degree of positive correlation.

4. **Limited degree of Negative correlation**

When an increase in the value of one variable is followed by a non-proportional decrease in other related variable, or when a decrease in the value of one variable is followed by a non- proportional increase in other related variable, it is called limited degree of negative correlation.

For example, if 10% rise in price results in 5% fall in its demand, it is limited degree of negative correlation. Similarly, if 5% fall in price results in 10% increase in demand, it is limited degree of negative correlation.

5. **Zero Correlation (Zero Degree correlation)**

If there is no correlation between variables it is called zero correlation. In other words, if the values of one variable cannot be associated with the values of the other variable, it is zero correlation.

Methods of measuring correlation

Correlation between 2 variables can be measured by graphic methods and algebraic methods.

I Graphic Methods

1) Scatter Diagram

2) Correlation graph

II Algebraic methods (Mathematical methods or statistical methods or Co-efficient of correlationmethods):

1) Karl Pearson‘s Co-efficient of correlation

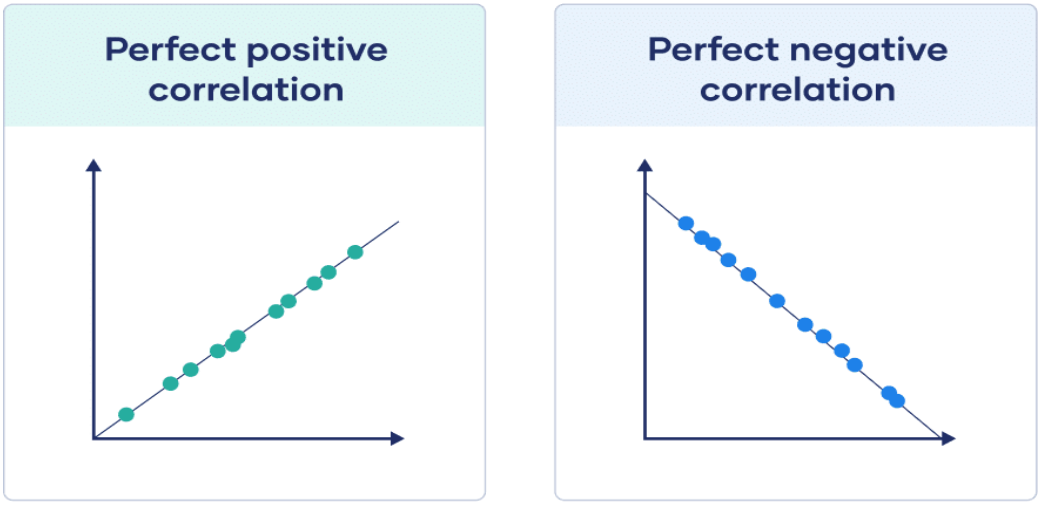
2) Spear man‘s Rank correlation method

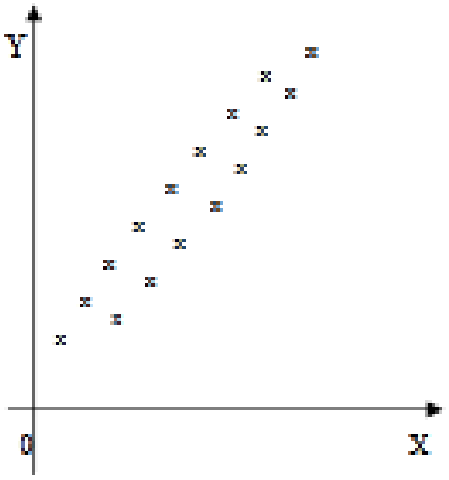
3) Concurrent deviation method

Scatter Diagram

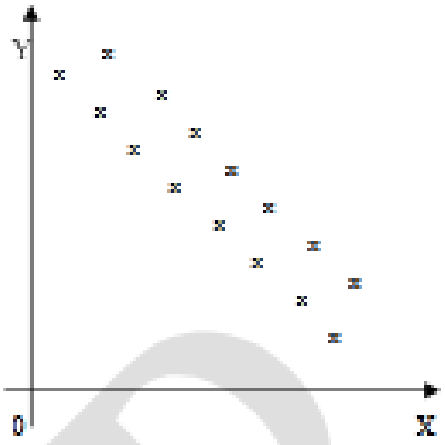
This is the simplest method for ascertaining the correlation between variables. Under this method all the values of the two variable are plotted in a chart in the form of dots. Therefore, it is also known as dot chart. By observing the scatter of the various dots, we can form an idea that whether the variables are related or not.

A scatter diagram indicates the direction of correlation and tells us how closely the two variables under study are related. The greater the scatter of the dots, the lower is the relationship.

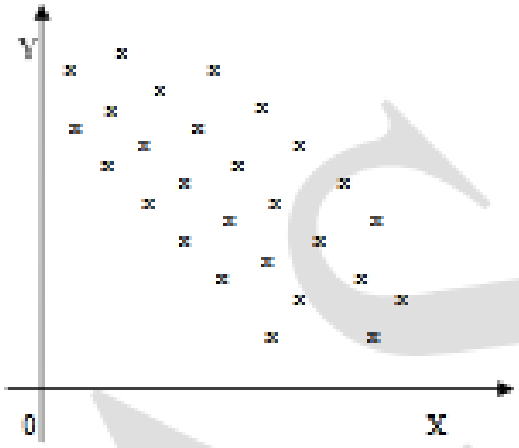




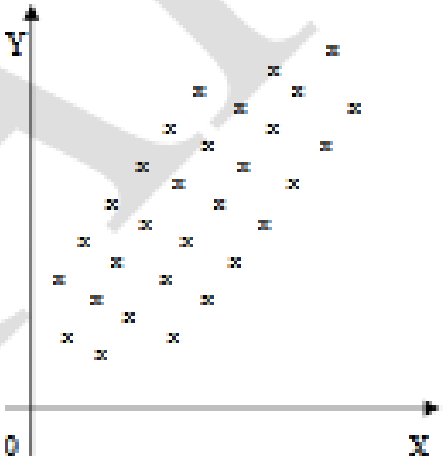
High Degree of Positive Correlation



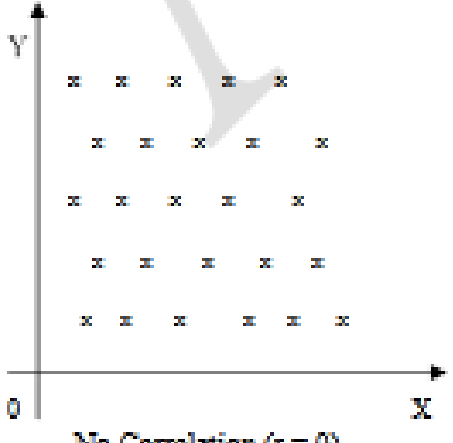
High Degree of Negative Correlation



Low Degree of Positive Correlation



Low Degree of Negative Correlation



No Correlation ($r = 0$)

Merits of Scatter Diagram method

1. It is a simple method of studying correlation between variables.
2. It is a non-mathematical method of studying correlation between the variables. It does notrequire any mathematical calculations.
3. It is very easy to understand. It gives an idea about the correlation between variables evento a layman.
4. It is not influenced by the size of extreme items.
5. Making a scatter diagram is, usually, the first step in investigating the relationshipbetween two variables.

Demerits of Scatter diagram method

1. It gives only a rough idea about the correlation between variables.
2. The numerical measurement of correlation co-efficient cannot be calculated underthis method.
3. It is not possible to establish the exact degree of relationship between the variables.

Correlation graph Method

Under correlation graph method the individual values of the two variables are plotted on a graph paper. Then dots relating to these variables are joined separately so as to get two curves. By examining the direction and closeness of the two curves, we can infer whether the variables are related or not. If both the curves are moving in the same direction(either upward or downward) correlation is said to be positive. If the curves are moving in the opposite directions, correlation is said to be negative.

Merits of Correlation Graph Method

1. This is a simple method of studying relationship between the variable
2. This does not require mathematical calculations.
3. This method is very easy to understand

Demerits of correlation graph method:

1. A numerical value of correlation cannot be calculated.
2. It is only a pictorial presentation of the relationship between variables.
3. It is not possible to establish the exact degree of relationship between the variables.

Karl Pearson's Co-efficient of Correlation

Karl Pearson's Coefficient of Correlation is the most popular method among the algebraic methods for measuring correlation. This method was developed by Prof. Karl Pearson in 1896. It is also called product moment correlation coefficient.

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

Interpretation of Co-efficient of Correlation

Pearson's Co-efficient of correlation always lies between +1 and -1. The following general ruleswill help to interpret the Co-efficient of correlation:

1. When r - +1, It means there is perfect positive relationship between variables.
2. When r = -1, it means there is perfect negative relationship between variables.

3. When r = 0, it means there is no relationship between the variables.
4. When ‗r' is closer to +1, it means there is high degree of positive correlation betweenvariables.
5. When ‗r' is closer to – 1, it means there is high degree of negative correlation betweenvariables.
6. When ‗r' is closer to ‗O', it means there is less relationship between variables.

Properties of Pearson's Co-efficient of Correlation

1. If there is correlation between variables, the Co-efficient of correlation lies between

   +1 and -1.
2. If there is no correlation, the coefficient of correlation is denoted by zero (ie r=0)
3. It measures the degree and direction of change
4. If simply measures the correlation and does not help to predict cansation.
5. It is the geometric mean of two regressions co-efficient.

Probable Error and Coefficient of Correlation

Probable error (PE) of the Co-efficient of correlation is a statistical device which measuresthe reliability and dependability of the value of co-efficient of correlation.

$$P.E.r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

$\overline{n}$

If the value of coefficient of correlation (r) is less than the PE, then there is no evidence ofcorrelation.

If the value of _r' is more than 6 times of PE, the correlation is certain and significant.

By adding and submitting PE from coefficient of correlation, we can find out the upperand lower limits within which the population coefficient of correlation may be expected to lie.**Uses of PE**:

1) PE is used to determine the limits within which the population coefficient ofcorrelation may be expected to lie.
2) It can be used to test whether the value of correlation coefficient of a sample issignificant with that of the population

If r = 0.6 and N = 64, find out the PE and SE of the correlation coefficient. Also determinethe limits of population correlation coefficient.

Merits of Pearson's Coefficient of Correlation:-

1. This is the most widely used algebraic method to measure coefficient of correlation.
2. It gives a numerical value to express the relationship between variables
3. It gives both direction and degree of relationship between variables
4. It can be used for further algebraic treatment such as coefficient of determinationcoefficient of non-determination etc.
5. It gives a single figure to explain the accurate degree of correlation between twovariables

Demerits of Pearson's Coefficient of correlation

1. It is very difficult to compute the value of coefficient of correlation.
2. It is very difficult to understand

Spearman's Rank Correlation Method

Pearson's coefficient of correlation method is applicable when variables are measured in quantitative form. But there were many cases where measurement is not possible because of the qualitative nature of the variable. For example, we cannot measure the beauty, morality, intelligence, honesty etc in quantitative terms. However it is possible to rank these qualitative characteristics in some order.

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)}$$

Merits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values arenot used for calculations
2. It is very simple to understand the method.
3. It can be applied to any type of data, ie quantitative and qualitative
4. It is the only way of studying correlation between qualitative data such as honesty, beautyetc.
5. As the sum of rank differences of the two qualitative data is always equal to zero,this method facilitates a cross check on the calculation.

Demerits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values are notused for calculations.
2. It is not convenient when number of pairs (ie. N) is large
3. Further algebraic treatment is not possible.
4. Combined correlation coefficient of different series cannot be obtained as in the case of mean and standard deviation. In case of mean and standard deviation, it is possible to compute combine arithmetic mean standard deviation.

Concurrent Deviation Method:

Concurrent deviation method is a very simple method of measuring correlation. Under this method, we consider only the directions of deviations. The magnitudes of the values are completely ignored. Therefore, this method is useful when we are interested in studying correlation between two variables in a casual manner and not interested in degree (or precision).

Under this method, the nature of correlation is known from the direction of deviation in the values of variables. If deviations of 2 variables are concurrent, then they move in the same direction, otherwise in the opposite direction.

The formula for computing coefficient of concurrent deviation is:

$$r_c = \pm\sqrt{\pm (2c - n) / n}$$

where c = number of concurrent deviations

n = number of pairs of signs (not the pairs of observations)

C = No. of concurrent deviations (ie, No. of + signs in _dx dy‘ column)

Steps:

1. Every value of _X‘ series is compared with its proceeding value. Increase is shown by
   _+‘ symbol and decrease is shown by _-_
2. The above step is repeated for _Y‘ series and we get _dy‘
3. Multiply _dx‘ by _dy‘ and the product is shown in the next column. The columnheading is _dxdy‘.
4. Take the total number of _+‘ signs in _dxdy‘ column. _+‘ signs in _dxdy‘ columndenotes the concurrent deviations, and it is indicated by _C‘.
5. Apply the formula:

MODULE-2

Regression analysis

Correlation analysis analyses whether two variables are correlated or not. After having established the fact that two variables are closely related, we may be interested in estimating the value of one variable, given the value of another. Hence, regression analysis means to analysesthe average relationship between two variables and thereby provides a mechanism for estimation or predication or forecasting.

The term ‗Regression‖ was firstly used by Sir Francis Galton in 1877. The dictionary meaning of the term ‗regression‖ is ―stepping back‖ to the average.

Definition:

–Regression is the measure of the average relationship between two or more variables in terms of the original units of the date‖.

–Regression analysis is an attempt to establish the nature of the relationship between variables-that is to study the functional relationship between the variables and thereby provides a mechanism for prediction or forecasting‖.

Types of Regression:-

There are two types of regression. They are linear regression and multiple regressions.

Linear Regression:

It is a type of regression which uses one independent variable to explain and/or predict thedependent variable.

Multiple Regression:

It is a type of regression which uses two or more independent variable to explain and/orpredict the dependent variable.

Regression Lines:

Regression line is a graphic technique to show the functional relationship between the two variables X and Y. It is a line which shows the average relationship between two variables X and Y.

If there is perfect positive correlation between 2 variables, then the two regression lines are winding each other and to give one line. There would be two regression lines when there is no perfect correlation between two variables. The nearer the two regression lines to each other, the higher is the degree of correlation and the farther the regression lines from each other, the lesser isthe degree of correlation.

Properties of Regression lines:-

1. The two regression lines cut each other at the point of average of X and average of Y (i.e $\bar{X}$ and $\bar{Y}$)
2. When r = 1, the two regression lines coincide each other and give one line.
3. When r = 0, the two regression lines are mutually perpendicular.

Regression Equations (Estimating Equations)

Regression equations are algebraic expressions of the regression lines. Since there are tworegression lines, therefore two regression equations. They are :-

1. Regression Equation of X on Y:- This is used to describe the variations in the valuesof X for given changes in Y.
2. Regression Equation of Y on X :- This is used to describe the variations in the value ofY for given changes in X.

Regression Equation of Y on X:-

$$Y = a + bx$$

The normal equations to compute _a' and _b' are: -

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Regression Equation of X on Y:-

$$X = a + by$$

The normal equations to compute _a' and _b' are:-

$$\Sigma x = Na + n\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Properties of Regression Coefficient:

1. There are two regression coefficients. They are b_{xy} and b_{yx}
2. Both the regression coefficients must have the same signs. If one is +ve, the otherwill also be a +ve value.
3. Thegeometricmean of regressioncoefficientswill be thecoefficient ofcorrelation. $r = \sqrt{b_{sy} \cdot b_{ys}}$.
4. If x $\tilde{x}$ and $\tilde{y}$ are the same, then the regression coefficient and correlation coefficient will be the same.

Computation of Regression Co-efficient

Regression co-efficient can be calculated in 3 different ways:

1. Actual mean method

- Regression coefficient x on y (b_{xy}) $= \frac{\Sigma sy}{\Sigma y^2}$
- Regression coefficient y on x ($\overline{b}_{yx}$) $= \frac{\Sigma sy}{\Sigma s^2}$

Correlation	**Regression**
It studies degree of relationshipbetween variables	It studies the nature of relationship between variables
It is not used for prediction purposes	It is basically used for prediction purposes
It is basically used as a tool for determining the degree of relationship	It is basically used as a tool for studying causeand effect relationship
There may be nonsense correlationbetween two variables	There is no such nonsense regression
There is no question of dependent andindependent variables	There must be dependent and independent variables

MODULE -3

Partial and multiple correlation coefficient

Correlation

Correlation is a term that is a measure of the strength of a linear relationship between two quantitative variables. Thus two variables are correlated if the change in one variable results in corresponding changes in the other variable. Eg: when a price of a commodity changes, the demand of the commodity also changes.

Partial correlation

Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables. In partial correlation we consider only two variables and other variables are treated as normal or having no effect and so ignored. Eg; consider three variables: yield, rainfall and temperature. Here the correlation between yield and rainfall treating temperature as normal is partial correlation.

Partial correlation coefficient

Partial correlation measures the strength of a relationship between two variables, while controlling for the effect of one or more other variables.

Let $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{x_3}$ be three variables then $\mathbf{r_{12.3}}$ is the partial correlation coefficient between x_1 and x_2 treating x_3 as constant. Similarly we have $\mathbf{r_{13.2}}$ and $\mathbf{r_{23.1}}$. Formulas are follows;

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

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\ \sqrt{1-r_{31}^2}}$$

$$r_{31.2} = \frac{r_{31} - r_{32}r_{12}}{\sqrt{1-r_{32}^2}\ \sqrt{1-r_{12}^2}}$$

note: r_{12} and r_{21} are same. r_{13} and r_{31} are same. r_{23} and r_{32} are same.

Ex.1. if r_{12} =0.7, r_{13} =0.61 , r_{23} =0.4. Find $\mathbf{r_{12.3}}$, $\mathbf{r_{13.2}}$ and $\mathbf{r_{23.1}}$.

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

$$\mathbf{r_{12.3}} = \frac{0.7-(0.61\times0.4)}{\sqrt{1}-0.61^2\ \sqrt{1}-0.4}^2$$

$$= 0.7\text{-}0.244/0.79\times0.92$$

$$= 0.63$$

b) $$r_{23.1}= \frac{r_{23}-r_{21}r_{31}}{\sqrt{1}-r_{21}^2\ \sqrt{1}-r_{31}^2}$$

$$r_{23.1}= \frac{0.4-(0.7\times0.61)}{\sqrt{1}-0.7^2\ \sqrt{1}-0.61^2}$$

$$= \frac{0.4-0.427}{0.71\times0.79}$$

$$= \frac{-0.027}{0.5609} = -0.048$$

c) $$r_{31.2}= \frac{r_{31}-r_{32}r_{12}}{\sqrt{1}-r_{32}^2\ \sqrt{1}-r_{12}^2}$$

$$= \frac{0.61-(0.7\times0.4)}{\sqrt{1}-0.7^2\ \sqrt{1}-0.4^2}$$

$$= \frac{0.61-0.28}{0.71\times0.92}$$

$$= \frac{0.33}{0.65} = 0.50$$

MULTIPLE CORRELATION

When there are more than two variables and we study the relationship between one variable and all the other variables taken together, then it is the case of multiple correlation. Suppose there are three variables, namely x, y and z. The correlation between x and (y & z) taken together is multiple correlation. Similarly, the relation between y and (x & z) taken together is multiple correlation. Again, the relation between z and (x & y) taken together is multiple correlation. In all these cases, the correlation coefficient obtained will be termed as coefficient of multiple correlation.

Suppose there are 3 variables namely x1, x2 and x3.Here, we can find three multiple correlation coefficients. They are:

1. Multiple Correlation Coefficient between x1 on one side and x2 and x3 together on the other side. This is denoted by R1.23

2.Multiple Correlation Coefficient between x2 on one side and x1 and x3 togetherontheotherside.ThisisdenotedbyR2.13

3. Multiple Correlation Coefficient between x3 on one side and x1 and x2 together on the other side. This is denoted by R3.12

The formulae for computing the above multiple correlation coefficients are:

$$R_{1.23} = \sqrt{[r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{23}^2]}$$

$$R_{2.13} = \sqrt{[r_{12}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{13}^2]}$$

$$R_{3.12} = \sqrt{[r_{13}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{12}^2]}$$

Qn: If $r_{12} = 0.6$, $r_{23} = r_{13} = 0.8$, find $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

Sol:

$$R_{1.23} = \sqrt{[r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{23}^2]}$$

$$= \sqrt{[0.6^2 + 0.8^2 - 2 \times 0.6 \times 0.8 \times 0.8] \div [1 - 0.8^2]}$$

$$= \sqrt{[0.36 + 0.64 - 0.768] \div [1 - 0.64]}$$

$$= \sqrt{0.232 / 0.36} = \sqrt{0.6444} = \underline{0.8028}$$

$$R_{2.13} = \sqrt{[r_{12}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{13}^2]}$$

$$= \sqrt{[0.6^2 + 0.8^2 - 2 \times 0.6 \times 0.8 \times 0.8] \div [1 - 0.8^2]}$$

$$= \sqrt{[0.36 + 0.64 - 0.768] \div [1 - 0.64]}$$

$$= \sqrt{0.232 / 0.36} = \sqrt{0.6444} = \underline{0.8028}$$

$$R_{3.12} = \sqrt{[r_{13}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{12}^2]}$$

MULTIPLE REGRESSION

In multiple regression there are more than two variables. Here, we examine the effect of two or more x3 independent variables on one dependent variable. Suppose there are three variables, namely, x1, x2 and x3. Here we may find three regression equations. They are:

1. Regression equation of x1 on x2 and x3

2. Regression equation of x2 on x1 and x3

3. Regression equation of x3 on x1 and x2

Equations of regression lines are generally termed as equations of planes of regression. Following are the formulae for computing the above 3 regression plane equations:

1. Regression equation of x_1 on x_2 and x_3:

$$(x_1 - \bar{x}_1) = b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3)$$

2. Regression equation of x_2 on x_1 and x_3:

$$(x_2 - \bar{x}_2) = b_{21.3}(x_1 - \bar{x}_1) + b_{23.1}(x_3 - \bar{x}_3)$$

3. Regression equation of x_3 on x_1 and x_2:

$$(x_3 - \bar{x}_3) = b_{31.2}(x_1 - \bar{x}_1) + b_{32.1}(x_2 - \bar{x}_2)$$

where $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ are actual means of x_1, x_2 and x_3 respectively.

Yule's Notation

Yule suggested that, the above equations may be simplified by taking $(x3 - \bar{x}\ 3) = X1$, $(x3 - \bar{x}\ 3) = X2$ and $(x3 - \bar{x}\ 3) = X3$. Then the equations are:

1. Regression equation of x_1 on x_2 and x_3:

   $$\mathbf{X_1 = b_{12.3}\, X_2 + b_{13.2}\, X_3}$$

2. Regression equation of x_2 on x_1 and x_3:

   $$\mathbf{X_2 = b_{21.3}\, X_1 + b_{23.1}\, X_3}$$

3. Regression equation of x_3 on x_1 and x_2:

   $$\mathbf{X_3 = b_{31.2}\, X_1 + b_{32.1}\, X_2}$$

In the above three equations, we used six regression coefficients. Following are the formulae for computing regression coefficients:

$$b_{12.3} = (\sigma_1/\sigma_2)\,[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$$

$$b_{13.2} = (\sigma_1/\sigma_3)\,[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$$

$$b_{21.3} = (\sigma_2/\sigma_1)\,[(r_{12} - r_{23}r_{13})/(1 - r_{13}^2)]$$

$$b_{23.1} = (\sigma_2/\sigma_3)\,[(r_{23} - r_{12}r_{13})/(1 - r_{13}^2)]$$

$$b_{31.2} = (\sigma_3/\sigma_1)\,[(r_{13} - r_{23}r_{12})/(1 - r_{12}^2)]$$

$$b_{32.1} = (\sigma_3/\sigma_2)\,[(r_{23} - r_{13}r_{12})/(1 - r_{12}^2)]$$

Qn: If $r_{12} = 0.7$, $r_{31} = r_{23} = 0.5$, $\sigma_1 = 2$, $\sigma_2 = 3$ and $\sigma_3 = 3$, find the equation of plane of regression x_1 on x_2 and x_3.

Sol:

Here, means of the variables are not given, and therefore, it is convenient to write the equations of planes of regression using Yule's notation.

Equation of plane of regression x_1 on x_2 and x_3 is :

$X_1 = b_{12.3}X_2 + b_{13.2}X_3$

$b_{12.3} = (\sigma_1/\sigma_2)\,[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$

$= (2/3)\,[(0.7 - 0.5 \times 0.5)/(1 - 0.5^2)]$

$= (2/3)\,[(0.7 - 0.25)/(1 - 0.25)]$

$= (2/3)\,[0.45/0.75] = (2/3)\,(0.6) = \underline{0.4}$

$b_{13.2} = (\sigma_1/\sigma_3)\,[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$

$= (2/3)\,[(0.5 - 0.7 \times 0.5)/(1 - 0.5^2)]$

$= (2/3)\,[(0.5 - 0.35)/(1 - 0.25)]$

$= (2/3)\,[0.15/0.75] = (2/3)\,(0.2) = \underline{0.133}$

$\therefore X_1 = \underline{0.4X_2 + 0.133X_3}$

Qn: In a trivariate distribution, $\bar{x}_1 = 53$, $\bar{x}_2 = 52$, $\bar{x}_3 = 51$, $\sigma_1 = 3.88$, $\sigma_2 = 2.97$, $\sigma_3 = 2.86$, $r_{23} = 0.8$, $r_{31} = 0.81$ and $r_{12} = 0.78$. Find the linear regression equation of x_1 on x_2 and x_3.

MODULE-4

Basic probability

Sets

Sets in mathematics, are simply a collection of distinct objects forming a group. A set can have any group of items, be it a collection of numbers, days of a week, types of vehicles, and so on. Every item in the set is called an element of the set. Curly brackets are used while writing a set. A very simple example of a set would be like this. Set A = {1,2,3,4,5}. There are various notations to represent elements of a set. Sets are usually represented using a roster form or a set builder form. Let us discuss each of these terms in detail.

Types of Sets

Sets are classified into different types. Some of these are singleton, finite, infinite, empty, etc.

Singleton Sets

A set that has only one element is called a singleton set or also called a unit set. Example, Set A = { k | k is an integer between 3 and 5} which is A = {4}.

Finite Sets

As the name implies, a set with a finite or countable number of elements is called a finite set. Example, Set B = {k | k is a prime number less than 20}, which is B = {2,3,5,7,11,13,17,19}

Infinite Sets

A set with an infinite number of elements is called an infinite set. Example: Set C = {Multiples of 3}.

Empty or Null Sets

A set that does not contain any element is called an empty set or a null set. An empty set is denoted using the symbol '∅'. It is read as '**phi**'. Example: Set X = {}.

Equal Sets

If two sets have the same elements in them, then they are called equal sets. Example: A = {1,2,3} and B = {1,2,3}. Here, set A and set B are equal sets. This can be represented as A = B.

Unequal Sets

If two sets have at least one element that is different, then they are unequal sets.Example: A = {1,2,3} and B = {2,3,4}. Here, set A and set B are unequal sets. This can be represented as A ≠ B.

Equivalent Sets

Two sets are said to be equivalent sets when they have the same number of elements, though the elements are different. Example: A = {1,2,3,4} and B = {a,b,c,d}. Here, set A and set B are equivalent sets since n(A) = n(B)

Overlapping Sets

Two sets are said to be overlapping if at least one element from set A is present in set B. Example: A = {2,4,6} B = {4,8,10}. Here, element 4 is present in set A as well as in set B. Therefore, A and B are overlapping sets.

Disjoint Sets

Two sets are disjoint sets if there are no common elements in both sets. Example: A = {1,2,3,4} B = {5,6,7,8}. Here, set A and set B are disjoint sets.

Subset and Superset

For two sets A and B, if every element in set A is present in set B, then set A is a subset of set B(A ⊆ B) and B is the superset of set A(B ⊇ A).
Example: A = {1,2,3} B = {1,2,3,4,5,6}
A ⊆ B, since all the elements in set A are present in set B.
B ⊇ A denotes that set B is the superset of set A.

Universal Set

A universal set is the collection of all the elements in regard to a particular subject. The universal set is denoted by the letter 'U'. Example: Let U = {The list of all road transport vehicles}. Here, a set of cars is a subset for this universal set, the set of cycles, trains are all subsets of this universal set.

Union of Sets

Union of sets, which is denoted as A U B, lists the elements in set A and set B or the elements in both set A and set B. For example, {1, 3} ∪ {1, 4} = {1, 3, 4}

Intersection of Sets

The intersection of sets which is denoted by A ∩ B lists the elements that are common to both set A and set B. For example, {1, 2} ∩ {2, 4} = {2}

Set Difference

Set difference which is denoted by A - B, lists the elements in set A that are not present in set B. For example, A = {2, 3, 4} and B = {4, 5, 6}. A - B = {2, 3}.

Set Complement

Set complement which is denoted by A', is the set of all elements in the universal set that are not present in set A. In other words, A' is denoted as U - A, which is the difference in the elements of the universal set and set A.

- **Example 1: Find the elements of the sets represented as follows and write the cardinal number of each set. a) Set A is the first 8 multiples of 7 b) Set B = {a,e,i,o,u} c) Set C = {x | x are even numbers between 20 and 40}**

  Solution:

  a) Set A = {7,14,21,28,35,42,49,56}. These are the first 8 multiples of 7.

  Since there are 8 elements in the set, cardinal number n (A) = 8

  b) Set B = {a,e,i,o,u}. There are five elements in the set,

  Therefore, the cardinal number of set B, n(B) = 5.

c) Set C = {22,24,26,28,30,32,34,36,38}. These are the even numbers between 20 and 40, which make up the elements of the set C.

Therefore, the cardinal number of set C, n(C) = 9.

- **Example 2: If Set A = {a,b,c}, Set B = {a,b,c,p,q,r}, U = {a,b,c,d,p,q,r,s}, find the following using sets formulas, a) A U B b) A ∩ B c) A' d) Is A ⊆ B? (Here 'U' is the universal set).**

  Solution: a) A U B = {a,b,c,p,q,r} b) A ∩ B = {a,b,c} c) A' = {d,p,q,r,s}

  d) A ⊆ B, (Set A is a subset of set B) since all the elements in set A are present in set B.

- **Example 3: Express the given set in set-builder form: A = {2, 4, 6, 8, 10, 12, 14}**

  Solution: Given: A = {2, 4, 6, 8, 10, 12, 14}

  Using sets notations, we can represent the given set A in set-builder form as,

  A = {x | x is an even natural number less than 15}

A sample space

A sample space is **a collection or a set of possible outcomes of a random experiment**. The sample space is represented using the symbol, “S”. The subset of possible outcomes of an experiment is called events. A sample space may contain a number of outcomes that depends on the experiment.

Events

An **event** is something that happens, especially when it is unusual or important. You can use **events** to describe all the things that are happening in a particular situation.

Sure Event (Certain Event)

An event whose occurrence is inevitable is called sure even.

Eg:- Getting a white ball from a box containing all while balls.
Impossible Events

An event whose occurrence is impossible, is called impossible event. Eg:- Getting a white ball from a box containing all red balls.

Uncertain Events

An event whose occurrence is neither sure nor impossible is called uncertain event. Eg:- Getting a white ball from a box containing white balls and blackballs.

Equally likely Events

Two events are said to be equally likely if anyone of them cannot be expected to occur in preference to other. For example, getting herd and getting tail when a coin is tossed are equally likely events.

Mutually exclusive events

A set of events are said to be mutually exclusive of the occurrence of one of them excludes the possibility of the occurrence of the others.

Exhaustive Events:

A group of events is said to be exhaustive when it includes all possible outcomes of the random experiment under consideration.

Dependent Events:

Two or more events are said to be dependent if the happening of one of them affects the happening of the other.

1. Classical or Priori Approach

If out of 'n' exhaustive, mutually exclusive and equally likely outcomes of an experiment; 'm' are favourable to the occurrence of an event 'A', then the probability of 'A' is defined as to be $\frac{m}{n}$

$$P(A) = \frac{m}{n}$$

According to Laplace, a French Mathematician, "the probability is the ratios of the number of favourable cases to the total number of equally likely cases."

$$P(A) = \frac{number\ of\ favourable\ cases}{total\ number\ of\ equally\ likely\ cases}$$

Question

What is the chance of getting a head when a coin is tossed?

Total number of cases = 2

No. of favorable cases = 1

Probability of getting head = $\frac{1}{2}$

Question

A die is thrown. Find the probability of getting.

a '4'

an even number

'3' or '5'

less than '3'

Solution

Sample space is (1,2, 3, 4, 5, 6)

Probability (getting '4) = $\frac{1}{6}$

Probability (getting an even number) = $\frac{3}{6} = \frac{1}{6}$

Probability (getting 3 or 5) = $\frac{2}{6} = \frac{1}{3}$

Probability (getting less than '3') = $\frac{2}{6} = \frac{1}{3}$

Addition Theorem

Here, there are 2 situations.

a. Events are mutually exclusive

b. Events are not mutually exclusive

(a) Addition theorem (Mutually Exclusive Events)

If two events, ‘A’ and ‘B’, are mutually exclusive the probability of the occurrence of either ‘A’ or ‘B’ is the sum of the individual probability of A and B.

$P(A \text{ or } B) = P(A) + P(B)$

i.e., $P(A\ B) = P(A) + P(B)$

(b)Addition theorem (Not mutually exclusive events)

If two events, A and B are not mutually exclusive the probability of the occurrence of either A or B is the sum of their individual probability minus probability for both to happen.

$P(A \text{ or } B) = P(A) + P(B) – P(A \text{ and } B)$

i.e., $P(A\ B) = P(A) + P(B) – P(A \cap B)$

Question
What is the probability of picking a card that was red or black?

Solution

Here the events are mutually exclusive

P(picking a red card)= 26/52

P(pickingablackcard)=26/52
P(picking a red or black card)=26/52+26/52

=1

(b)Multiplication theorem (dependent Events):-

If two events, A and B are dependent, the probability of occurring 2nd event will be affected by the outcome of the first.

$P(A \cap B) = P(A).P(B/A)$

Question
A bag contains 5 white balls and 8 black balls. One ball is drawn at random from the bag. Again, another one is drawn without replacing the first ball. Find the probability that both the balls drawn are white.

Solution
P (drawing a white ball in Ist draw) = 5/13

P(drawingawhiteballinIInddraw)=4/12
=2/156

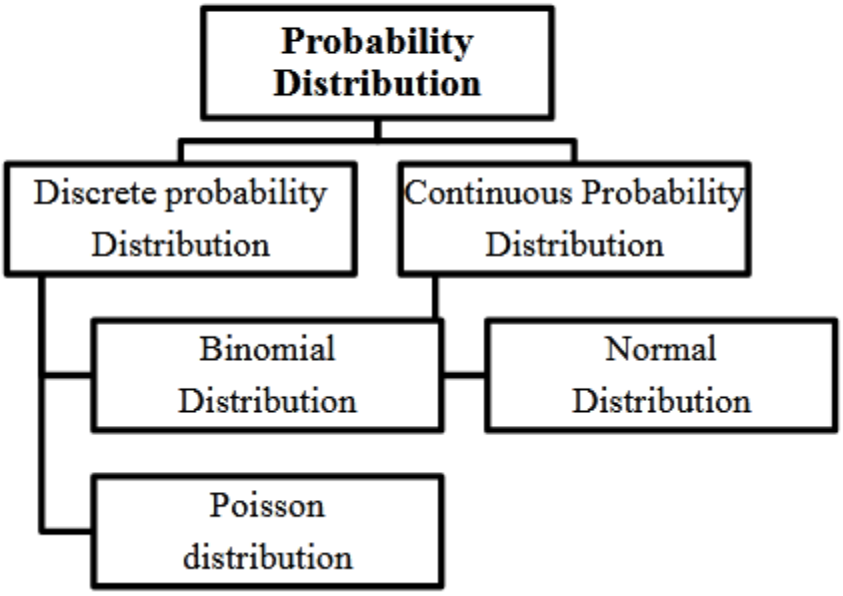
Module 5

Random Variables and their probability distributions

Random variable

Random variable is a variable who value is determined by the outcome of a random experiment. Random variable is also called chance variable or stochastic variable. For example, suppose we toss a coin. Obtaining of head in this random experiment is a random variable. Here the random variable of "obtaining heads" can take the numerical values. Now, we can prepare a table showing the values of the random variable and corresponding probabilities. This is called probability distributions or theoretical distribution.

Classification of Probability Distribution



Discrete Probability Distribution

If the random variable of a probability distribution assumes specific values only, it is called discrete probability distributions. Binomial distribution and poisson distribution are discrete probability distributions.

Continuous Probability Distributions

If the random variable of a probability distribution assumes any value in a given interval, then it is called continuous probability distributions. Normal distributions is a continuous probability distribution.

BionomialDistribution

Binomial Distribution is associated with James Bernoulli, a Swiss Mathematician. Therefore, it is also called Bernoulli distribution. Binomial distribution is the probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure.

In other words, it is used to determine the probability of success in experiments on which there are only two mutually exclusive outcomes. Binomial distribution is discrete probability distribution. Binomial Distribution can be defined as follows: "A random variable r is said to follow Binomial Distribution with parameters n and p if its probability function is;

$$P(r) = {}^{n}C_{r}p^{r}q^{n-r}$$

Where, P = probability of success in a single trial

q = 1 – p

n = number of trials

r = number of success in 'n' trials.

Assumption of Binomial Didstribution

Binomial distribution can be applied when:-

1. The random experiment has two outcomes i.e., success and failure.
2. The probability of success in a single trial remains constant from trial to trial of the experiment.
3. The experiment is repeated for finite number of times.
4. The trials are independent.

Properties (features) of Binomial Distribution

1. It is a discrete probability distribution.
2. The shape and location of Binomial distribution changes as 'p' changes for a given 'n'.
3. The mode of the Binomial distribution is equal to the value of 'r' which has the largest probability.
4. Mean of the Binomial distribution increases as 'n' increases with 'p' remaining constant.
5. The mean of Binomial distribution is np.
6. The Standard deviation of Binomial distribution is $\sqrt{npq}$
7. If 'n' is large and if neither 'p' nor 'q' is too close zero, Binomial distribution may be approximated to Normal Distribution.
8. If two independent random variables follow Binomial distribution, their sum also follows Binomial distribution.

Qn: Six coins are tossed simultaneously. What is the probability of obtaining 4 heads?

Sol: P(r) = nC r prqn-r

r = 4

n = 6

p = ½

q = 1 – p = 1 – ½ = ½

∴ p(r = 4) = 6C4 (½)4 (½)$^{6-4}$

$$= \frac{6!}{(6-4)!4!}! \text{ x } (½)^{4+2}$$

$$= \frac{6!}{2!4!}! \text{ x } (½)^{6}$$

$$= \frac{6\times5}{2\times1} \text{ x } \frac{1}{64}$$

$$= \frac{30}{128}$$

$$= \underline{0.234}$$

PoissonDistribution

Poisson Distribution is a limiting form of Binomial Distribution. In Binomial Distribution, the total numbers of trials are known previously. But in certain real life situations, it may be impossible to count the total number of times a particular event occurs or does not occur. In such cases Poisson Distribution is more suitable. Poison Distribution is a discrete probability distribution. It was originated by Simeon Denis Poisson. The Poisson Distribution is defined as:-

$$p(r) = \frac{e^{-m} m^r}{r!}$$

Where r = random variable (i.e., number of success in '**n**' trials.

e = 2.7183

m = mean of poisson distribution

Properties of Poisson Distribution

1. Poisson Distribution is a discrete probability distribution.

2. Poisson Distribution has a single parameter 'm'. When 'm' is known all the terms can be found out.

3. It is a positively skewed distribution.

4.Mean and Varriance of Poisson Distribution are equal to 'm'.

5. In Poisson Distribution, the number of success is relatively small.

6. The standard deviation of Poisson Distribution is √m.

Practical situations where Poisson Distribution can be used

1. To count the number of telephone calls arising at a telephone switch board in a unit of time.

2. To count the number of customers arising at the super market in a unit of time.

3. To count the number of defects in Statistical Quality

Control.

4. To count the number of bacterias per unit.
5. To count the number of defectives in a park of manufactured goods.
6. To count the number of persons dying due to heart attack in a year.

7. To count the number of accidents taking place in a day on a busy road.

Qn: It is known from the past experience that in a certain plant, there are on an average four industrial accidents per year. Find the probability that in a given year there will be less than four accidents. Assume poisson distribution.

Sol: $p(r<4) = p(r = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3)$

$= p(r = 0) + p(r = 1) + p(r = 2) + p(r = 3)$

$P(r) = \frac{e^{-m} m^r}{r!}$

$m = 4$

$\therefore p(r = 0) = \frac{e^{-4} \cdot 4^0}{0!} = \frac{0.01832 \times 1}{1} = 0.01832$

$p(r = 1) = \frac{e^{-4} \cdot 4^1}{1!} = \frac{0.01832 \times 4}{1} = 0.07328$

$p(r = 2) = \frac{e^{-4} \cdot 4^2}{2!} = \frac{0.01832 \times 16}{2 \times 1} = 0.14656$

$p(r = 3) = \frac{e^{-4} \cdot 4^3}{3!} = \frac{0.01832 \times 64}{3 \times 2 \times 1} = 0.19541$

$\therefore p(r < 4) = 0.01832 + 0.07328 + 0.14656 + 0.19541$

= 0.43357