MATHEMATICAL AND STATISTICAL FOUNDATIONS

UNIT-2

Simple Linear Regression and Correlation and random variables and Probability distributions

2.0 OBJECTIVES

- To Understand the importance of having Simplified expression in Linear Regression
- To understand the concept of correlation
- To learn the ideas of Probability and axioms of Probability
- To understand the concept of random variables, exhaustive events.
- To classify the discrete and continuous random variables.
- To learn the method to find Probability distributions.

2.1 Introduction

Linear regression is a very Powerful statistical technique. Many people have some familiarity with regression just from reading the news where graphs with straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is linear relationship between two numerical variables. The representation is a linear

equation that combines a specific set of input values (x). As such, both the input values (x) and the output value are numeric. Linear Regression refers to a group of techniques for fitting and studying the straight line relationship between two variables. Linear regression estimates the regression coefficients β_0 and β_1 in the equation $Y_j = \beta_0 + \beta_1 X_j + j \in (0,1)$ where X is the independent variable, Y is the dependent variable, β_0 is the Y intercept, β_1 is the slope and $\in$ is the error.

Once the intercept and slope have been estimated using least squares, various indices are studied to determine the reliability of these estimates. One of the most popular of these reliability indices is the correlation coefficient. The correlation coefficient, or simply the correlation, is an index that ranges

from -1 to 1. when the value is near zero, there is no linear relationship. As the correlation gets closer to plus or minus one, the relationship is stronger. A value of one (or negative one) indicaty a perfect linear relationship between two variables. Actully, the strict interpretation of the correlation is different from that given in the last paragraph. The correlation is a parameter of the bivariate normal distribution. This distribution is used to describe the association between two variables. This association does not include a cause and effect statement. That is, the variables are not labeled as dependent and independent. one does not depend on the other. Rather, they are considered as two random variables that seem to vary together. The important point is that in linear

regression, Y is assumed to be a random variable and X is assumed to be a fixed variable. In correlation analysis, both Y and X are assumed to be random variable.

2.2 Simple Linear Regression

Simple Linear regression, which involves two variables. one is denoted by Y, which is the variable of primary Concern. Its behaviour is uncontrolled and uncertain. That's why, Y is a random variable. The mean of Y and the value assumed by Y depends on some other variable X. The relationship between the mean of Y and x is linear.

we know that the behaviour of Y depends on x, so Y is called the dependent variable (or) the response variable. Use observed data to develop a linear equation, express the mean of Y interms of x. The values of x used in developing this equation is controlled. so, x is not considered as a random variable.

Select the values of x to use in our experiment and then observe the values which we assumed by the random variable Y at these points. Since the values of x don't depend on Y, then x is called the "independent variable" (or) predictor variable.

Example: A physician wants to predict the concentration of a particular drug in the blood stream (Y) based on the length of time (X) since the drug was administered to the patient.

Example: An economist wants to develop an equation by which the price of Rice in Nizamabad (Y) can be predicted from the amount of received during the growing season in the midwest farm belt (X).

From the above discussion Y is a random variable whose distribution depends on x. Mostly we are interested mainly in the relation between x and the mean of the corresponding distribution of the Y's.

2.3 The Method of Least squares:

The Method of Least squares is most systematic method to fit a unique curve through the given data points and is widely used in Practical computations.

Let the observed value at $x=x_i$ is y_i and corresponding value on the curve is $f(x_i)$

Let e_i is the error of approximation at $x=x_i$. Then we have

$$e_i = y_i - f(x_i)$$

Consider,

$$S = \left[(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_n - f(x_n))^2\right]$$

$$S = e_1^2 + e_2^2 + \cdots + e_n^2$$

The Method of least squares consists of minimizing S.

2.3.1 Fitting a straight line:

Let x_i, y_i (where $i = 1, 2, 3, \ldots n$) be the data of an unknown function $y = f(x)$. Let $y = a + bx$ is a straight line to be fitted to the given data.

Now, minimizing S

i.e, $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$

We have,

$$S = [(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \cdots + (y_n - f(x_n))^2]$$

$$S = [(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2]$$

Now, $\frac{\partial S}{\partial a} = 0$

$$\Rightarrow 2[y_1 - a - bx_1](-1) + 2[y_2 - a - bx_2](-1) + \cdots + 2[y_n - a - bx_n](-1) = 0$$

$$\Rightarrow (y_1 + y_2 + y_3 + \cdots + y_n) - na - b(x_1 + x_2 + \cdots + x_n) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i}$$

Now, $\frac{\partial S}{\partial b} = 0$

$$\Rightarrow 2[y_1 - a - bx_1)(-x_1) + 2[y_2 - a - bx_2](-x_2) + \cdots + 2[y_n - a - bx_n](-x_n) = 0$$

$$\Rightarrow (y_1x_1 - ax_1 - bx_1^2) + (y_2x_2 - ax_2 - bx_2^2) + \cdots + (y_nx_n - ax_n - bx_n^2) = 0$$

$$\Rightarrow (y_1x_1 + y_2x_2 + \cdots + y_nx_n) - a(x_1 + x_2 + \cdots + x_n) - b(x_1^2 + x_2^2 + \cdots x_n^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2}$$

∴ The two normal equations to fit a straight line are

$$\boxed{\begin{aligned} \Sigma y &= na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}}$$

Example: Fit a straight line to the following data by the method of least squares.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution: Let $y = a + bx$ is the required straight line to be fitted.

The two normal equations of a straight line are

$$\Sigma y = na + b\Sigma x \quad \text{——} (1)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{——} (2)$$

Here, n = no. of observations = 5

12

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x = 10$	$\Sigma y = 16.9$	$\Sigma xy = 47.1$	$\Sigma x^2 = 30$

Substitute all the above values in equation ① & ②

$$16.9 = 5a + 10b \quad \text{—} \quad ③$$

$$47.1 = 10a + 30b \quad \text{—} \quad ④$$

By solving eqn ③ & ④,

we get $a = 0.72$

$b = 1.33$

$\therefore y = (0.72) + (1.33)x$ is the required straight line.

Example: The temperature T (in °c) and length (in mm) of a heated rod are given below, if $L = a_0 + a_1 T$. Find the best values of a_0 and a_1?

T	20	30	40	50	60	70
L	800.3	800.4	800.6	800.7	800.9	801.0

Solution: Let $L = a_0 + a_1 T$ —(1) is the required straight line to be fitted.

The two normal eqns of a straight line are

$$\Sigma L = n a_0 + a_1 \Sigma T \quad -(2)$$

$$\Sigma LT = a_0 \Sigma T + a_1 \Sigma T^2 \quad -(3)$$

T	L	TL	T^2
20	800.3	16006	400
30	800.4	24012	900
40	800.6	32024	1600
50	800.7	40035	2500
60	800.9	48054	3600
70	801.0	56070	4900
$\Sigma T = 270$	$\Sigma L = 48039$	$\Sigma TL = 216201$	$\Sigma T^2 = 13900$

Substitute all the above values in eqn ② & ③

$$48039 = 6a_0 + 270a_1 \quad \text{— ④}$$

$$216201 = 270a_0 + 13900a_1 \text{ — ⑤}$$

By solving eqn ④ & ⑤

we get,

$$a_0 = 799.9943$$

$$a_1 = 0.0146$$

2.3.2 Fitting a Parabola (or) Second degree Polynomial:

Let $y = a+bx+cx^2$ — ① is a parabola to be fitted to the given data.

WKT $S = [y_1 - a - bx_1 - cx_1^2]^2 + [y_2 - a - bx_2 - cx_2^2]^2$
$+ \cdots\cdots + [y_n - a - bx_n - cx_n^2]^2$

To minimize S, we have $\frac{\partial S}{\partial a} = 0$,

$\frac{\partial S}{\partial b} = 0$, $\frac{\partial S}{\partial c} = 0$.

Now, $\frac{\partial S}{\partial a} = 0$

$2[(y_1 - a - bx_1 - cx_1^2)](-1) + 2[y_2 - a - bx_2 - cx_2^2](-1) +$
$\cdots\cdots\cdots + 2[y_n - a - by_n - cy_n^2](-1) = 0$

$(-2)[y_1 + y_2 + \cdots + y_n - na - b(x_1 + x_2 + \cdots + x_n) -$
$c(x_1^2 + x_2^2 + \cdots + x_n^2)] = 0$

$$\Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2} \quad - (2)$$

Now $\frac{\partial S}{\partial b} = 0$

$\Rightarrow 2[y_1 - a - bx_1 - cx_1^2](-x_1) + 2[y_2 - a - bx_2 - cx_2^2](-x_2)$

$+ \cdots + 2[y_n - a - bx_n - cx_n^2](-x_n) = 0$

$\Rightarrow -2[y_1x_1 - ax_1 - bx_1^2 - cx_1^3 + y_2x_2 - ax_2 - bx_2^2 - cx_2^3 +$

$\cdots + y_nx_n - ax_n - bx_n^2 - cx_n^3] = 0$

$\Rightarrow (x_1y_1 + x_2y_2 + \cdots + x_ny_n) - a(x_1 + x_2 + \cdots + x_n)$

$-b(x_1^2 + x_2^2 + \cdots + x_n^2) - c(x_1^3 + x_2^3 + \cdots + x_n^3) = 0$

$\Rightarrow \sum_{i=1}^{n} x_iy_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 - c\sum_{i=1}^{n} x_i^3 = 0$

$\Rightarrow \boxed{\sum_{i=1}^{n} x_iy_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3} \quad - (3)$

Now $\frac{\partial S}{\partial c} = 0$

$\Rightarrow 2[y_1 - a - bx_1 - cx_1^2](-x_1^2) + 2[y_2 - a - bx_2 - cx_2^2](-x_2^2)$

$+ \cdots + 2[y_n - a - bx_n - cx_n^2](-x_n^2) = 0$

$\Rightarrow -2[y_1x_1^2 - ax_1^2 - bx_1^3 - cx_1^4 + y_2x_2^2 - ax_2^2 - bx_2^3 - cx_2^4$

$+ \cdots + y_nx_n^2 - ax_n^2 - bx_n^3 - cx_n^4] = 0$

$\Rightarrow (y_1x_1^2 + y_2x_2^2 + \cdots + y_nx_n^2) - a(x_1^2 + x_2^2 + \cdots + x_n^2)$

$-b(x_1^3 + x_2^3 + \cdots + x_n^3) - c(x_1^4 + x_2^4 + \cdots + x_n^4) = 0$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^2 - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i^3 - c\sum_{i=1}^{n} x_i^4 = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} y_i x_i^2 = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4} \quad -(4)$$

∴ The three normal equations to fit a Parabola are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Example: Fit a Parabola of the form $y = a + bx + cx^2$ to the following data

x	1	2	3	4	5	6	7
y	2.3	5.2	9.7	16.5	29.4	35.5	54.4

By the method of least squares.

Solution: Let $y = a + bx + cx^2$ —① is the required Parabola to be fitted. The three normal equations to fit a Parabola

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \quad -(2)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad -(3)$$
$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad -(4)$$

x	y	xy	x^2	x^2y	x^3	x^4
1	2.3	2.3	1	2.3	1	1
2	5.2	10.4	4	20.8	8	16
3	9.7	29.1	9	87.3	27	81
4	16.5	66	16	264	64	256
5	29.4	147	25	735	125	625
6	35.5	213	36	1278	216	1296
7	54.4	380.8	49	2665.6	343	2401
$\Sigma x = 28$	$\Sigma y = 153$	$\Sigma xy = 848.6$	$\Sigma x^2 = 140$	$\Sigma x^2y = 5053$	$\Sigma x^3 = 784$	$\Sigma x^4 = 4676$

Substitute the above values in eqn (2) (3) & (4)

we get,

$$153 = 7a + 28b + 140c$$
$$848.6 = 28a + 140b + 784c$$
$$5053 = 140a + 784b + 4676c$$

By solving the above equations

we get,

$$a = 2.3714$$

$$b = -1.0928$$

$$c = 1.1928$$

$\therefore$ $y = 2.3714 - (1.0928)x + (1.1928)x^2$ is the required Parabola.

Example: Fit a Parabola of the form $y = a + bx + cx^2$ to the following data

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

By method of least squares.

Solution: Let $y = a + bx + cx^2$ ——(1) is the required Parabola to be fitted. The Three normal equations to fit a Parabola

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \text{ ——(2)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \text{ ——(3)}$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \text{ ——(4)}$$

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	6.3	25.2	16	100.8	64	256
$\Sigma x = 10$	$\Sigma y = 12.9$	$\Sigma xy = 37.1$	$\Sigma x^2 = 30$	$\Sigma x^2y = 130.3$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$

Substitute the above values in equations ②, ③ & ④.

$$12.9 = 5a + 10b + 30c \quad -⑤$$

$$37.1 = 10a + 30b + 100c \quad -⑥$$

$$130.3 = 30a + 100b + 354c \quad -⑦$$

By solving equations ⑤, ⑥ & ⑦

we get $a = 1.42$, $b = -1.07$, $c = 0.55$

$\therefore y = 1.42 - 1.07x + 0.55x^2$ is the required parabola.

2.3.3 Fitting a exponential curve

Let $y = ae^{bx}$ — ① is the required curve to be fitted to the given set of points i.e, (x_i, y_i)

Taking $\log_e$ on both sides

$$\log_e y = \log_e (ae^{bx})$$

$$\log_e y = \log_e a + \log_e e^{bx}$$

$$\log_e y = \log_e a + bx \quad \text{— ②}$$

Let $Y = \log_e y$

$A = \log_e a$

$B = b$

$X = x$

Substituting all the values in eqn ②

$Y = A + BX$ which is a straight line

The Normal equations to fit a straight line are

$$\Sigma Y = NA + B\Sigma X \quad \text{— ③}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \text{— ④}$$

By solving ③ & ④, we get A & B values.

But, $A = \log_e a$

$\Rightarrow \boxed{a = e^A}$

and $B = b \Rightarrow \boxed{B = b}$

Finally, Substitute a, b values in eqn (1)

i.e, $\boxed{y = ae^{bx}}$

Example: Determine the constants a, b by the method of least squares such that $y = ae^{bx}$ fit the following table.

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

Solution: Let $y = ae^{bx}$ — (1). is the required curve to be fitted to the given set of points i.e, (x_i, y_i)

Taking $\log_e$ on both sides

$\log_e y = \log_e (ae^{bx})$

$\log_e y = \log_e a + \log_e e^{bx}$

$$\log_e y = \log_e a + bx \quad -(2)$$

Let $Y = \log_e y$

$A = \log_e a$

$X = x$

$B = b$

Substituting all the values in eqn (2)

$Y = A + BX$ which is a straight line.

The Normal equations to fit a Straight line

$$\Sigma Y = NA + B\Sigma X \quad -(3)$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad -(4)$$

$x = X$	y	$Y = \log_e y$	XY	X^2
2	4.077	1.4054	2.8108	4
4	11.084	2.4055	9.6220	16
6	30.128	3.4055	20.4330	36
8	81.897	4.4055	35.244	64
10	222.62	5.4055.	54.055	100
$\Sigma X = 30$		$\Sigma Y = 17.0274$	$\Sigma XY = 122.1648$	$\Sigma X^2 = 220$

Substitute all the values in eqn ③ & ④

$$17.0274 = 5A + 30B \quad -⑤$$

$$122.1648 = 30A + 220B \quad -⑥$$

By solving ⑤ & ⑥ we get

$$A = 0.4054 \,, \quad B = 0.5$$

But $\log_e a = A \Rightarrow a = e^A$

$$\Rightarrow a = e^{0.4054}$$

$$\Rightarrow a = 1.49$$

$$B = b \Rightarrow b = B$$

$$\Rightarrow b = 0.5$$

Finally substitute a & b values in eqn ①

$$\text{i.e, } y = ae^{bx}$$

$$\Rightarrow y = (1.49)e^{(0.5)x}.$$

Example: Fit a curve of the form $y = ae^{bx}$

x	0	1	2	3
y	1.05	2.10	3.85	8.30

Solution: Let $y = ae^{bx}$ is the required curve — (1)
to be fitted to the given set of
points i.e, (x_i, y_i)

Taking $\log_e$ on both sides

$$\log_e y = \log_e a + \log_e e^{bx} \quad \text{— (2)}$$

Let $Y = \log_e y$

$A = \log_e a$

$B = b$

$X = x$

Substituting all the values in eqn (2)

$Y = A + BX$ which is a straight line

∴ The Normal equations to fit a straight line are

$$\Sigma Y = NA + B\Sigma X \quad \text{— (3)}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \text{— (4)}$$

$x = X$	y	$Y = \log_e y$	XY	X^2
0	1.05	0.048	0	0
1	2.10	0.741	0.741	1
2	3.85	1.348	2.696	4
3	8.30	2.116	6.348	9
$\Sigma X = 6$		$\Sigma Y = 4.253$	$\Sigma XY = 9.785$	$\Sigma X^2 = 14$

Substitute all the above values in eqn. ③ & ④

$$4.253 = 4A + 6B \quad -⑤$$

$$9.785 = 6A + 14B \quad -⑥$$

By solving eqns ⑤ and ⑥

we get,

$$A = 0.042$$

$$B = 0.6809$$

But, $a = e^A \Rightarrow a = e^{0.042} = 1.04$

and $b = B \Rightarrow b = 0.6809$

Substitute a & b values in eqn ①

i.e, $y = (1.04)e^{(0.6809)x}$

2.3.4 Fitting a Power Curve

Let $y = ab^x$ — ① is the required curve to be fitted to the given set of points (x_i, y_i).

Taking $\log_{10}$ on both sides

$$\log_{10} y = \log_{10} a + x \log_{10} b \quad — ②$$

Let $\log_{10} y = Y$, $\log_{10} a = A$, $x = X$,

$\log_{10} b = B$.

Substitute all the values in eqn ②

$$Y = A + BX \quad — ③$$

which is a straight line

$$\Sigma Y = NA + B\Sigma X \quad — ④$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad — ⑤$$

By solving eqn ④ & ⑤, we get A & B values.

But $\log_{10} a = A \Rightarrow a = 10^A$ (anti log)

and $\log_{10} b = B \Rightarrow b = 10^B$ (anti log)

Substituting a & b values in eqn ①

i.e, in $y = ab^x$

Example:

Fit a curve of the form $y = ab^x$ to the data given below:

x	1	2	3	4	5	6
y	151	100	61	50	20	8

Solution: Let $y = ab^x$ — (1)

Take $\log_{10}$ on both sides

$$\log_{10} y = \log_{10}(ab^x)$$

$$\log_{10} y = \log_{10} a + \log_{10} b^x$$

$$\log_{10} y = \log_{10} a + x \log_{10} b \quad — (2)$$

Let $Y = \log_{10} y$

$$A = \log_{10} a$$

$$X = x$$

$$B = \log_{10} b$$

Substituting all the values in (2)

$$Y = A + BX \quad — (3)$$

which is a straight line

The normal equations are

$$\Sigma Y = nA + B\Sigma X \quad — (4)$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad — (5)$$

$X=x$	y	$Y=\log_{10} y$	XY	x^2
1	151	2.1790	2.1790	1
2	100	2	4	4
3	61	1.7853	5.3559	9
4	50	1.6990	6.7960	16
5	20	1.3010	6.5050	25
6	8	0.9031	5.4186	36
$\Sigma X=81$		$\Sigma Y=9.8674$	$\Sigma XY=30.2545$	$\Sigma x^2=91$

Here $n=6$.

Substitute all the values in eqn (4) & (5)

$$9.8674 = 6A + 21B \quad \text{— (6)}$$

$$30.2545 = 21A + 91B \quad \text{— (7)}$$

By solving eqn (6) & (7)

we get, $A = 2.5$, $B = -0.2427$

But, $A = \log_{10} a \Rightarrow a = 10^A = 10^{2.5}$

$$\boxed{a = 316.22}$$

and $B = \log_{10} b \Rightarrow b = 10^B = 10^{-0.2427}$

$$\boxed{b = 0.569}$$

Substitute a and b in eqn (1)

$\therefore y = (316.22)(0.569)^x$

2.4 Properties of the Least-Squares Estimators:

Under the assumptions of the Simple-regression Model, the sample least-squares coefficients A and B have several properties as estimators of the Population regression coefficients α and β

→ The least-squares intercept and slope are linear estimators, that is they are linear functions of the observations y_i.

For example, $$B = \sum_{i=1}^{n} m_i y_i$$

where, $$m_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

The result makes the distributions of the least-squares coefficients simple.

→ under the assumption of linearity, A and B are unbiased estimators of α and β.

$$E(A) = \alpha$$
$$E(B) = \beta$$

→ under the assumptions of linearity, constant variance, and independence of A and B have simple sampling variances.

$$V(A) = \frac{\sigma_\epsilon^2 \Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2}$$

$$V(B) = \frac{\sigma_\epsilon^2}{\Sigma(X_i - \bar{X})^2}$$

→ under normality, the least-squares estimators are most efficient among all unbiased estimators, not just among linear estimators. This is a much more compelling result.

→ under the assumptions, the least-squares coefficients A and B are the maximum-likelihood estimators of α and β.

→ under the assumption of normality, the least-squares coefficients are themselves normally distributed.

$$A \sim N\left[\alpha, \frac{\sigma_\varepsilon^2 \Sigma x_i^2}{n \Sigma (x_i - \bar{x})^2}\right]$$

$$B \sim N\left[\beta, \frac{\sigma_\varepsilon^2}{\Sigma (x_i - \bar{x})^2}\right]$$

Even if the errors are not normally distributed, the distribution of A and B are approximately normal, with the approximation improving as the sample size grows (The Central limit theorem).

2.5 : CORRELATION

Definition.

Correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship or independence of two sets of variables upon each other. One variable may be called the subject and the other relative (dependent).

Types of Correlation

Correlation is classified into many types.

1. Positive and negative.
2. Simple and multiple.
3. Partial and total
4. Linear and non-linear.

1. Positive And Negative Correlation.

Positive and Negative correlation depend upon the direction of change of the variables. If two variables tend to move together in the same direction. i.e an increase in the value of one variable is accompanied by an increase in the value of other variable.

or a decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called positive or direct correlation.

Simple and multiple correlation
When we study only two variables, the relationship is described as simple correlation; example quantity of money and price level, demand and price etc. But in a multiple correlation we study morethan two variables simultaneously.

Partial and Total correlation.
The study of two variables excluding some other variables is called partial correlation. for example, we study price and demand, eliminating the supply side. In total correlation, all the facts are taken into account.

Linear and Non-linear correlation.
If the ratio of change between two variable is uniform, then there will be linear correlation between them. Consider

A	2	7	12	17
B	3	9	15	21

We can see that the ratio of change between the variables is the same. If we plot these on the graph, we get a straight line.

Methods of Studying Correlation.

There are 2 different methods for finding out the relationship between variables. They are
(1) Graphic method (2) Mathematical method

(1). Graphic methods
(a) Scatter diagram or scattergram.
(b) Simple graph.

(2) Mathematical methods
(a) Karl Pearson's coefficient of correlati
(b) Spearman's Rank coefficient of correlati
(c) Coefficient of concurrent deviation.
(d) Method of least squares.

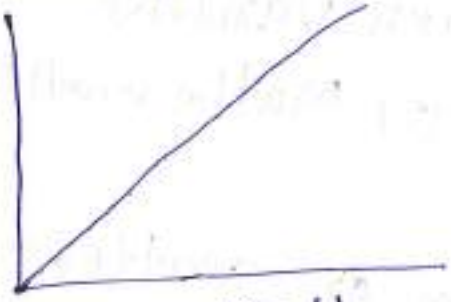
Scatter Diagram or Scattergram.

The scatter diagram is a chart obtained by plotting two variables to find out whether there is any relationship between them.

In this diagram are ploted on the horizontal axis and Y variables are ploted on the vertical axis. Thus we can know the scatter or concentration of various points.

Various scatter diagrams are briefly shown here.

Perfect positive correlation

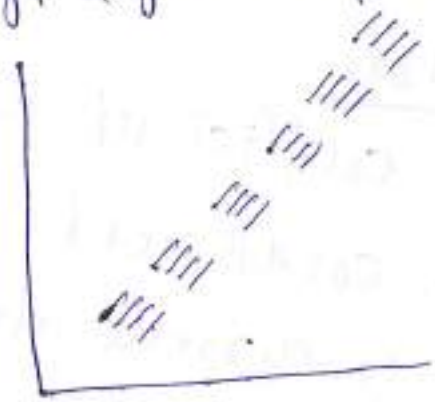


$r = +1$

Perfect Negative Correlati

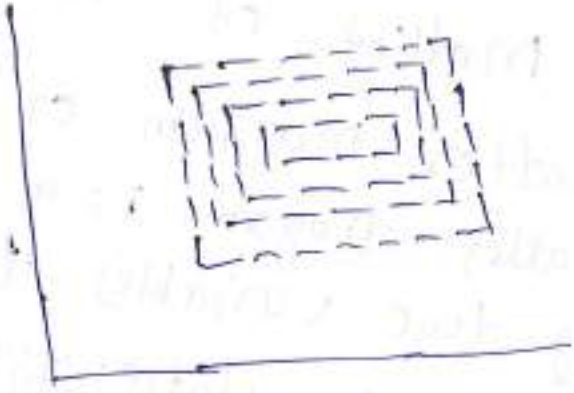
$r = -1$

High degree of positive Correlation



High degree of negative Correlation

No correlation



2.5.1 ADVANTAGES OF SCATTER DIAGRAM

1. Scatter diagram is a simple, attractive method to find out the nature of correlation.
2. It is easy to understand.
3. A rough idea is got at a glance whether it is positive or negative correlation.

SIMPLE GRAPH

The values of the two variables are plotted on a graph paper. We get two curves, one for X variables and another for Y variables. These two curves reveal the direction and closeness of the two variables and also reveal whether are not the variables are related If both the curves move in the same direction i.e parallel to each other, either upward or downward, correlation is said to negative.

Cofficient of correlation.

Correlation is a statistical technique used for analysing the behaviour of two or more variables. Its analysis deals with the association, between two or more variables Statistical measures of correlation relates

to covariation between series but not of function or casual relationship.

2.5.2. KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson is a British Biometrician and Statistician suggested a mathematical method for measuring the magnitude of linear relationship between two variables. This is known as Pearsonian Coefficient of correlation. It is denoted by r. This method is most widely used. It is also called product-moment correlation coefficient.

There are several formulae to calculate r.

They are (1) $r = \frac{\text{covariance of } xy}{\sigma x \times \sigma y}$

(2) $r = \frac{\Sigma xy}{N . \sigma x \sigma y}$

(3) $r = \frac{\Sigma XY}{\sqrt{\Sigma x^2 . \Sigma Y^2}}$

$X = (x - \bar{x})$, $Y = (y - \bar{y})$ where $\bar{x}$, $\bar{y}$ are mean of the series x and y.

σx = standard deviation of series x.

σy = standard deviation of series y.

2.53 Properties of Correlation Coefficient

1. Limits for correlation coefficient are $-1 \leq r \leq 1$ Hence Correlation coefficient can not exceed one numerically.
   If $r=1$ correlation is perfect and positive
   If $r=-1$ Correlation is perfect and negative
   If $r=0$ then there is no relationship between the variables.

2. Correlation coefficient is independent of change of origin and scale.
   If $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$ then $r(X,Y) = r(U$

3. If X, Y are random variables and a, b, c, d are any numbers suchthat $a \neq 0$, $c \neq 0$ then
   $$r(aX+b, cY+d) = \frac{ac}{|ac|} r(X,Y)$$

4. Two independent variables are uncorrelated. That is if X and Y are independent variables then $r(X,Y) = 0$.

EXAMPLES.

Example: Calculate the coefficient of correlation between age of cars and annual maintenance cost and comment:

42

Age of cars	2	4	6	7	8	10	12
Annual maintenance cost	1600	1500	1800	1900	1700	2100	2000

Solution: Let age of cars $= x$.
Annual maintenance $= y$.

Computation of coefficient of Corelation

x	y	$X = x - \bar{x}$, $\bar{x} = 7$	$Y = (y - \bar{y})/100$, $\bar{y} = 1800$	X^2	Y^2	XY
2	1600	−5	−2	25	4	10
4	1500	−3	3	9	9	9
6	1800	−1	0	1	0	0
7	1900	0	1	0	1	0
8	1700	1	−1	1	1	−1
10	2100	3	3	9	9	9
12	2000	5	2	25	4	10
$\Sigma x = 49$	$\Sigma y =$ 12600	$\Sigma X = 0$	$\Sigma Y = 0$	$\Sigma X^2 = 70$	$\Sigma Y^2 = 28$	$\Sigma XY = 37$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{37}{\sqrt{70 \times 28}} = 0.836.$$

We can observe that there is a high degree of positive correlation between age of cars and annual maintenance cost.

Example: Find Karl Pearson's coefficient of Correlation from the following data

Wages	100	101	102	102	100	99	97	98	96
Cost of living	98	99	99	97	95	92	95	94	90

Solution: Computation of Karl Pearson's coefficie of Correlation

Wages (x)	$X = x - \bar{x}$, $\bar{x} = 99$	X^2	Cost of living (y)	$Y = y - \bar{y}$, $\bar{y} = 95$	Y^2	X
100	+1	1	98	+3	9	+
101	+2	4	99	+4	16	+
102	+3	9	99	+4	16	+
102	+3	9	97	+2	4	+
100	+1	1	95	0	0	
99	0	0	92	−3	9	
97	−2	4	95	0	0	
98	−1	1	94	−1	1	
96	−3	9	90	−5	25	
95	−4	16	91	−4	16	+
$\Sigma x = 990$	$\Sigma X = 0$	$\Sigma X^2 = 54$	$\Sigma y = 950$	$\Sigma Y = 0$	$\Sigma Y^2 = 96$	=

$$\therefore r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}} = \frac{61}{\sqrt{54 \times 96}} = \frac{61}{\sqrt{5184}} = \frac{61}{72}$$

$$= 0.847$$

$$\therefore r = +0.847$$

Example: Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and Engineering ratio (E.R) calculate the coefficient of correlation.

student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

Solution: We construct the following table

Student	Intelligence ratio x	$x-\bar{x}=X$	Engineering ratio y	$y-\bar{y}=Y$	X^2	Y^2	XY
A	105	6	101	3	36	9	18
B	104	5	103	5	25	25	25
C	102	3	100	2	9	4	6
D	101	2	98	0	4	0	0
E	100	1	95	-3	1	9	-3
f	99	0	96	-2	0	4	0
G	98	-1	104	6	1	36	-6
H	96	-3	92	-6	9	36	18
I	93	-6	97	-1	36	1	6
J	92	-7	94	-4	49	16	28
Totals	990	0	980	0	170	140	125

$$= \frac{322}{\sqrt{196 \times 588}} = \frac{322}{359.48} = 0.95.$$

Example: find if there is any correlation between the heights and weighs given below.

Height in inches	57	59	62	63	64	65	55	58	5
Weight in lbs	113	117	126	126	130	129	111	116	1

Solution: Coefficient of correlation $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}}$

Computation of coefficient of correlation

Height in inches x	Deviation from Mean (60) $X = x - \bar{x}$	Squares of deviations X^2	Weight y	Deviations of Mean $Y = y - \bar{y}$	Squares deviations Y^2
57	−3	9	113	−7	49
59	−1	1	117	−3	9
62	2	4	126	6	36
63	3	9	126	6	36
64	4	16	130	10	100
65	5	25	129	9	81
55	−5	25	111	−9	81
58	−2	4	116	−4	16
57	−3	9	112	−8	64
540	0	102	1080	0	472

$$r = \frac{216}{\sqrt{102 \times 471}} = 0.98$$

Example: Calculate coefficient of Correlation from the following data

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution: In both series items are in small number. So there is no need to take deviations.

We use the formula $r = \frac{\text{Covariance of } xy}{\sigma(x) \cdot \sigma(y)}$

Computation of coefficient of Correlation:

X	Y	x^2	y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\Sigma x = 70$	$\Sigma y = 63$	$\Sigma x^2 = 728$	$\Sigma y^2 = 651$	$\Sigma xy = 676$

$$r = \frac{(\Sigma xy \times N) - (\Sigma x \times \Sigma y)}{\sqrt{(\Sigma x^2 \times N - (\Sigma x)^2) \cdot (\Sigma y^2 \times N - (\Sigma y)^2}}$$

Here $N = 7$

$$r = \frac{(676 \times 7) - (70 \times 63)}{\sqrt{728 \times 7 - (70)^2 \cdot (651 \times 7) - (63)^2}} = \frac{4732 - 4410}{\sqrt{5096 - 4900} \sqrt{4557 - 3969}}$$

From this table, mean of x, i.e $\bar{x} = \frac{990}{10} = 99$; mean of y i.e. $\bar{y} = \frac{980}{10} = 98$

$\Sigma X^2 = 170$, $\Sigma Y^2 = 140$ and $\Sigma XY = 92$

Substituting these values in $r = \frac{\Sigma XY}{\sqrt{(\Sigma X^2 \Sigma Y^2)}} =$

$$\frac{92}{\sqrt{170 \times 140}} = \frac{92}{154.3} = 0.59.$$

2.6 WHEN DEVIATIONS ARE TAKEN FROM AN ASSUMED MEAN.

When actual mean is not a whole number, but a fraction or when the series is large, the calculation by direct method will involve a lot of time. To avoid such tedious calculation, we can use the assumed mean method.

$$\therefore \text{formula} = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right)} \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

where

X = deviation of the items of x-series from a assumed mean. i.e $X = x - A$

Y = deviation of the items of y-series from an assumed mean. i.e $Y = (y - A)$.

N = Number of items

ΣXY = The total of the product of the deviations of x and y-series from their assumed mean.

Σx^2 = The total of the squares of the deviations of x-series from an assumed mean.

Σy^2 = The total of the squares of the deviations of y-series from an assumed mean.

$\Sigma X . \Sigma Y$ = Total of the deviations of x-series from assumed mean.

ΣY = Total of the deviations of y-series from assumed mean.

Example: Find the coefficient of correlation between X and Y

X	1	2	3	4	5	6	7	8	9
Y	12	11	13	15	14	17	16	19	18

Solution:

X	Y	X^2	Y^2	XY
1	12	1	144	12
2	11	4	121	22
3	13	9	169	39
4	15	16	225	60
5	14	25	196	70
6	17	36	289	102
7	16	49	256	112
8	19	64	361	132
9	18	81	324	162
45	135	285	2085	731

$$\bar{X} = \frac{45}{9} = 5 \; ; \quad \bar{Y} = \frac{135}{9} = 15$$

$\therefore$ Using the formula $r = \dfrac{Cov(X,Y)}{\sigma X \, \sigma Y}$

$$= \frac{\Sigma XY - N(\bar{X})(\bar{Y})}{\sqrt{(\Sigma X^2 - N(\bar{X})^2)(\Sigma Y^2 - N(\bar{Y})}}$$

$$= \frac{731 - 9(5)(15)}{\sqrt{(285 - 9\times 25)(2085 - 9\times 2}}$$

$$= 0.933$$

50

Example: Calculate Karl Pearson's Correlation coefficient for the following paired data.

X	28	41	40	38	35	33	40	32	36	3
Y	29	34	33	34	30	26	28	31	36	3

what inference would you draw from the estim

Solution: Computation of correlation coefficient

x	Deviation from assumed mean $X = x - 35$ $(X = x - \bar{x})$	Square of deviation X^2	y	Deviation from assumed mean $Y = y - 31$ $(Y = y - \bar{y})$	Square of deviation Y^2	product XY
28	−7	49	29	−8	64	56
41	+6	36	34	+3	9	18
40	+5	25	33	+2	4	10
38	+3	9	34	+3	9	9
35	0	0	30	−1	1	0
33	−2	4	26	−5	25	10
40	+5	25	28	−3	9	−1
32	−3	9	31	0	0	0
35	+1	1	36	+5	25	5
33	−2	4	38	+7	49	1
$\Sigma x = 355$	$\Sigma X = 6$	$\Sigma X^2 = 162$	$\Sigma y = 317$	$\Sigma Y = 3$	$\Sigma Y^2 = 195$	Σ =

$N = 10$, take $\bar{x} = 35$ and $\bar{y} = 31$

Applying to the above data, the formula

$$\therefore r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right] \times \left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

$$= \frac{79 - \frac{6 \times 3}{10}}{\sqrt{162 - \frac{6^2}{10}} \cdot \sqrt{195 - \frac{7^2}{10}}} = \frac{77.2}{\sqrt{158.4}\sqrt{194.9}}$$

$$= 0.45$$

Example: Calculate the correlation coefficient for the following heights of fathers (X) and their sons (Y):

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution: Computation of correlation coefficient

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4986
70	69	4900	4761	4830
72	71	5184	5041	5112
$\Sigma X = 538$	$\Sigma Y = 552$	$\Sigma X^2 = 37028$	$\Sigma Y^2 = 38172$	$\Sigma XY = 37560$

$$\therefore \bar{X} = \frac{1}{N}\Sigma x = \frac{544}{8} = 68, \quad \bar{Y} = \frac{1}{N}\Sigma y = \frac{552}{8} = 69$$

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{N}\Sigma xy - (\bar{x})(\bar{y})}{\sqrt{\left(\frac{1}{N}\Sigma x^2 - (\bar{x})^2\right)\left(\frac{1}{N}\Sigma y^2 - (\bar{y})^2\right)}}$$

$$= \frac{\frac{1}{8}\times 39560 - (68\times 69)}{\sqrt{\left(\frac{37028}{8} - (68)^2\right)\left(\frac{38132}{8} - (69)^2\right)}} = 0.803.$$

Example: find a suitable coefficient of Correlatio for the following data:

Fertiliser used	15	18	20	24	30	35	40	5
Productivity	85	93	95	105	120	130	150	1

Solution: Computation of coefficient of correlation

Fertiliser used (x)			Productivity (y)			
x	$X = x - \bar{x}$ $= x - 29$	X^2	y	$Y = y - \bar{y}$ $= y - 119$	Y^2	XY
15	−14	196	85	−34	1156	476
18	−11	121	93	−26	676	286
20	−9	81	95	−24	576	216
24	−5	25	105	−14	196	70
30	1	1	120	1	1	1
35	6	36	130	11	121	66
40	11	121	150	31	961	341
50	21	441	160	41	1681	861
$\Sigma x = 0$		$\Sigma X^2 = 1022$		$\Sigma Y = 14$	$\Sigma Y^2 = 5368$	$\Sigma XY = 2$

$$\therefore N=8, \quad r=\frac{\Sigma XY - N(\Sigma X)(\Sigma Y)}{\sqrt{(\Sigma X)^2 - N(\Sigma X)^2 \times (\Sigma Y)^2 - N(\Sigma}}$$

$$=\frac{2317}{\sqrt{1022 \times 5343.5}} = \frac{2317}{2336.89} = 0.99$$

Example: find out the coefficient of correlation in the following case.

Height of father	65	66	67	67	68	69	71	7
Height of son	67	68	64	68	72	70	69	7

Solution: Computation of coefficient of correlation

x	$X=x-\bar{x}$	x^2	y	$Y=y-\bar{y}$	y^2	XY
65	−2	4	67	−1	1	2
66	−1	1	68	0	0	0
67	0	0	64	−4	16	0
67	0	0	68	0	0	0
68	1	1	72	4	16	4
69	2	4	70	2	4	4
71	4	16	69	1	1	4
73	6	36	70	2	4	12
$\Sigma x=546$	$\Sigma X=10$	$\Sigma X^2=62$	$\Sigma y=548$	$\Sigma Y=4$	$\Sigma y^2=42$	Σ

N = no. of items = 8

$$\therefore \text{Coefficient of correlation } r = \frac{\Sigma XY - \frac{(\Sigma X)\times(\Sigma Y)}{N}}{\sqrt{\left\{\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right\}\left\{\Sigma Y^2 - \frac{(\Sigma Y)}{N}\right.}}$$

$$= \frac{(\Sigma XY)\times N - (\Sigma X)(\Sigma Y)}{\sqrt{\left((\Sigma X)^2 \times N - (\Sigma X)^2\right)\times\left((\Sigma Y)^2 \times N(\Sigma\right.}}$$

$$= \frac{(26\times 8) - (10\times 4)}{\sqrt{(62\times 8 - 10^2)\times(42\times 8 - 4^2)}}$$

$$= \frac{168}{355.98}$$

$$= 0.472$$

2.7 RANK CORRELATION COEFFICIENT

A British Psychologist Charles Adward Spearman found out the method of finding the coefficient of correlation by ranks. This method is based on rank and is useful in dealing with qualitative such as morality, character, intelligence and beauty. It can not be measured quantitatively as in the case of Pearson's coefficient of correlation. It is based on the ranks given to the obervatio Rank correlation is applicable only to the individu observations.

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}.$$

where $\rho \rightarrow$ Rank coefficient of Correlation

$D^2 \rightarrow$ Sum of the squares of the differences of two Ranks.

$N \rightarrow$ Number of paired observations.

2.7.1 PROPERTIES OF RANK CORRELATION COEFFICIENT

1. The value of ρ lies between +1 and −1.
2. If $\rho = 1$, there is complete agreement in the order if the ranks and the direction of the rank is same.
3. If $\rho = -1$ then there is complete disagreement in the order of the ranks and they are in opposite directions.

Procedure to solve problems:

1. when the ranks are given.

step-1. Compute the difference of two ranks and denote it by D.

step-2. Square D and get ΣD^2.

step-3. obtain ρ by substituting the figures in the formula.

2. when the ranks are not given, but actual

data are given, then we must give ranks, We can give ranks by taking the highest as 1 or the lowest value as 1, next to the highest as 2 and follow the same procedure for both the variables.

Example: Ten competitors in a musical test were ranked by the three judges A, B and C in the following order.

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearst approch to common likings in music.

Solution: Here $N = 10$.

Ranks by A X	Ranks by B Y	Ranks by C Z	$D_1 = X - Y$	$D_2 = X - Z$	$D_3 = Y - Z$	D_1^2	D_2^2	D_3^2
1	3	6	−2	−5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	−3	−4	−1	9	16	1
10	4	8	6	2	−4	36	4	1
3	7	1	−4	2	6	16	4	36
2	10	2	−8	0	8	64	0	0
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	1
7	6	5	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	1
Total			$\Sigma D_1 = 0$	$\Sigma D_2 = 0$	$\Sigma D_3 = 0$	$\Sigma D_1^2 = 200$	$\Sigma D_2^2 = 60$	

$$\rho_1(X,Y) = 1 - \frac{6\Sigma D_1^2}{N(N^2-1)} = 1 - \frac{6\times 200}{10\times 99} = 1 - \frac{40}{33} = \frac{-7}{33}$$

$$\rho_2(X,Z) = 1 - \frac{6\Sigma D_2^2}{N(N^2-1)} = 1 - \frac{6\times 60}{10\times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho_3(Y,Z) = 1 - \frac{6\Sigma D_3^2}{N(N^2-1)} = 1 - \frac{214}{165} = -\frac{49}{165}$$

Since $\rho_2(X,Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

Example: The ranks of 60 students in mathematics and statistics are as follows. (1,1), (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (14,12) (15,16) (16,13). Calculate the rank correlation coefficient for proficiencies of this group in mathematics and statistics.

Solution:

Ranks in maths (X)	1	2	3	4	5	6	7	8	9	10	11
	12	15	14	15	16						
Ranks in statistics (Y)	1	10	3	4	5	7	2	6	8	11	15
	9	14	12	16	13						
$D = X - Y$	0	-8	0	0	0	-1	5	2	1	-1	-4
	3	-1	2	-1	3		$\Sigma D = 0$				
D^2	0	64	0	0	0	1	25	4	1	1	16
	9	1	4	1	9		$\Sigma D^2 = 136.$				

∴ Rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6\times 16}{16\times 25} = 1 - \frac{1}{5}$$

$$= \frac{4}{5} = 0.8.$$

Example: A random sample of 5 college students is selected and their grades in mathematics and statistics are found to be

	1	2	3	4	5
mathematics	85	60	73	40	90
statistics	93	75	65	50	80

Calculate Pearman's rank correlation coefficient

Solution:

Marks in mathematics X	Ranks x	marks in statistics Y	Rank y	Rank difference x−y	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	−1	1
40	5	50	5	0	0
90	1	80	2	−1	1
					$\Sigma D^2 = 4$

Here $N = 5$, $\Sigma D^2 = 4$

∴ Pearman's Rank correlation

$$= 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 4}{5(5^2-1)} = 1 - \frac{24}{5 \times 24}$$

$$= 1 - \frac{1}{5} = 1 - 0.2 = 0.8.$$

Example: Following are the rank obtained by 10 students in two subjects, statistics and mathematics. To what extent the knowledge of the students in two subjects is related?

statistics	1	2	3	4	5	6	7	8	9	10
mathematics	2	4	1	5	3	9	7	10	6	8

Solution:

Rank in statistics (x)	Rank in maths (y)	D = (x − y)	D^2
1	2	−1	1
2	4	−2	4
3	1	+2	4
4	5	−1	1
5	3	+2	4
6	9	−3	9
7	7	0	0
8	10	−2	4
9	6	+3	9
10	8	+2	4

60

$$\therefore \rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 40}{10(10^2-1)}$$

$$= 1 - \frac{240}{10(100-1)}$$

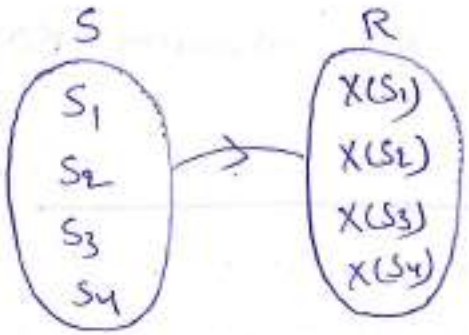
$$= 1 - \frac{240}{990}$$

$$= 1 - 0.24$$

$$= 0.76.$$

2.8 Random Variable:

Definition: A random variable X on a sample space S is a function $X: S \to R$ that assigns a real number $(X(S))$ to each sample point $S \in S$.



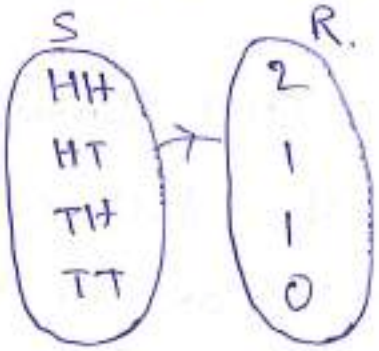
Note: Set of all possible values of a random variable X is called Range of X.

Example: In tossing two coins

Sample space $S = \{HH, HT, TH, TT\}$

Define a random variable $X: S \to R$ by

$X(S)$ = "No. of Heads"



Range of $X = \{2, 1, 0\}$ a set of real numbers.

$\therefore$ X is a random variable

There are Two Types of Random variables.

1) Discrete Random Variables
2) Continuous Random Variables.

1) Discrete Random Variables:-

A random variable X which can take only a finite number of discrete values in an interval of domain is called a discrete random variable.

(or)

If the random variable takes the values only on the set $\{0, 1, 2, \ldots, n\}$ is called a Discrete random variable.

Example: Tossing a coin, Throwing a die, The No. of defective items in a sample of electric bulbs, the No. of Printing mistakes in each page of a book,

The no. of telephone calls received by the telephone operator, no. of Accidents in a street etc.

2) Continuous Random Variable:

A random variable X which can take values continuously i.e, which takes all possible values in a given interval is called a continuous Random variable.

Example: The height, age and weight of individuals, temperature and time, life span of an electric bulb.

Depending on random variable we obtain the following Probability distributions.

Probability distribution function:

Let X be a random variable. Then the probability distribution function associated with X is defined as the Probability that the outcome of an

experiment will be one of the outcomes for which $X(S) \leq x$, $x \in R$

i.e, $F_X(x) = P(X \leq x) = P\{S : X(S) \leq x\}$, $-\infty < x < \infty$

is called the distribution function of X.

2.8.1 **Discrete Probability distribution (Probability Mass function):**

Let X be a Discrete random variable. Then the probability distribution given by X is known as Discrete probability distribution which is represented by the following Probability Mass function $p_i = P(X = x_i) = P(x_i)$ for $i = 1, 2, 3 \cdots$

with the property (i) $P(x_i) > 0, \forall i$

(ii) $\sum_{i=1}^{\infty} P(x_i) = 1$, $i = 1, 2, 3, \cdots$

2.8.2 **Cumulative Distribution Function of a Discrete Random Variable:**

Let X is a Discrete Random variable.

Then the Discrete distribution function (or) cumulative Distribution function $F(x)$ is defined by $F(x) = P(X \leq x)$

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

where, x is any integer.

Example: In Tossing two coins,

Then $S = \{HH, HT, TH, TT\}$

Define $X : S \to R$ by $X(S)$ = no. of heads.

Range of $X = \{0, 1, 2\}$

$X = x$	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

is called P.m.f (D.P.D)

$X = x$	0	1	2
$P(X \leq x)$	$\frac{1}{4}$	$\frac{3}{4}$	1

is called Distribution function (or) cumulative distribution function.

2.8.3 Probability Density function:

The Probability density function $f_X(x)$ is defined as the derivative of the Probability distribution function, $F_X(x)$ of the random variable X.

Thus, $f_X(x) = \frac{d}{dx}[F_X(x)]$

Properties:

1) If F is the distribution function of a random variable X and if $a<b$, then

(i) $P(a<X\leq b) = F(b) - F(a)$

(ii) $P(a\leq X\leq b) = P(X=a) + [F(b) - F(a)]$

(iii) $P(a<X<b) = [F(b) - F(a)] - P(X=b)$

(iv) $P(a\leq X<b) = [F(b) - F(a)] - P(X=b) + P(X=a)$

Note:

If $P(X=a) = P(X=b) = 0$ then

$P(a<X\leq b) = P(a\leq X\leq b) = P(a<X<b) =$

$P(a\leq X<b) = F(b) - F(a)$

2) All distribution functions are monotonically increasing and lie b/w '0' and '1'
i.e, if F is the distribution function of the random variable X, then

(i) $0 \leq F(x) \leq 1$

(ii) $F(x) < F(y)$ when $x < y$

3) (i) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$

(ii) $F(\infty) = \lim_{x \to \infty} F(x) = 1$.

2.8.4 Continuous Probability distribution:

Let X be a continuous random variable. Then the probability distribution given by X is known as continuous probability distribution which is represented by $f(x)$ and known as probability density function (P.d.f) with the properties

(i) $f(x) \geq 0$, $\forall x \in R$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

2.8.5 Cumulative Distribution Function of a Continuous random Variable:

The cumulative distribution (or) simply the distribution function of a continuous random variable X is denoted by F(x) and is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

Properties:

1) $0 \leq F(x) \leq 1$, $-\infty < x < \infty$
2) $F'(x) = f(x) \geq 0$, so that F(x) is a non-decreasing function
3) $F(-\infty) = 0$
4) $F(\infty) = 1$
5) F(x) is a continuous function of x on the right
6) The discontinuities of F(x) are countable
7) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$
8) Since, $F'(x) = f(x)$, we have $\frac{d}{dx}[F(x)] = f(x)$

$$\Rightarrow dF = f(x)\,dx$$

This is known as Probability differential of X.

2.9 General Properties of a Discrete Probability Distribution:

Mean:

The mean of a Discrete probability distribution is the mathematical expectation of a discrete random variable.

If X is a discrete random variable which takes the values $x_1, x_2, \ldots x_K$ and whose probabilities are $f(x_1), f(x_2), \ldots f(x_K)$, then it's mean (or) Mathematical expectation (or) expected value is $\sum_{i=1}^{K} x_i f(x_i)$. The mean is denoted by μ.

$$\therefore \mu = \sum_{i=1}^{K} x_i f(x_i)$$

Example: Let X be a discrete random variable taking the values $1, 2, \ldots 6$ with probability $f(x_i) = \frac{1}{6}$

Then $\mu = \sum_{i=1}^{6} x_i f(x_i)$

$= \frac{1}{6}(1+2+\cdots+6)$

$\mu = \frac{7}{2}$

Variance:

Let X be a discrete random variable. Then the variance of X, denoted by σ^2 is,

$$\sigma^2 = \Sigma (x-\mu)^2 f(x)$$

(or)

$$\sigma^2 = \Sigma x^2 f(x) - \mu^2$$

Example: Determine the variance of the probability distribution of the number of points rolled with a die.

Solution: Since $f(x) = \frac{1}{6}$

we get $\mu = \frac{7}{2}$

and $\Sigma x^2 f(x) = (1^2+2^2+3^2+4^2+5^2+6^2)\frac{1}{6} = \frac{91}{6}$

Hence $\sigma^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$

Standard Deviation:

Let X be a discrete random variable. The Standard Deviation of a discrete random variable X is denoted by σ, is the positive square root of σ^2

i.e, $\sigma = \sqrt{\Sigma (x-\mu)^2 f(x)}$

$\sigma = \sqrt{\Sigma x^2 f(x) - \mu^2}$

Example: If the variance $\sigma^2 = 2.0979$

Then Standard deviation $\sigma = \sqrt{2.0979}$

$= 1.448$

Results:

1) Variance of constant is zero
2) $V(KX) = K^2 V(X)$, where K-constant
3) If X is a random variable and K is a constant, then $V(X+K) = V(X)$
4) If X is a Discrete random Variable, then $V(aX+b) = a^2 V(X)$, where V(X) is a Variance of X and a, b are constants.

5) If X and Y are two independent random variables, then $V(X \pm Y) = V(X) \pm V(Y)$

Example: A random variable X has the following probability function:

X	0	1	2	3	4	5	6	7
P(x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

(i) Find K
(ii) Evaluate $P(X<6)$, $P(X \geq 6)$
(iii) Evaluate $P(0<X<5)$, $P(0 \leq X \leq 4)$
(iv) Find mean
(v) Find variance

Solution: (i) We know that $\sum_{x=0}^{7} P(x) = 1$

i.e, $0+K+2K+2K+3K+K^2+2K^2+7K^2+K=1$

i.e, $10K^2+9K-1=0$

i.e, $(10K-1)(K+1)=0$

$\therefore K = \frac{1}{10} = 0.1 \quad (\because K \neq -1)$

(ii) $P(X<6) = P(X=0) + P(X=1) + \ldots + P(X=5)$

$= 0 + K + 2K + 2K + 3K + K^2$

$= 8K + K^2$

$= 0.8 + 0.01$

$= 0.81$

$P(X \geq 6) = 1 - P(X<6)$

$= 1 - 0.81$

$= 0.19$

(iii) $P(0<X<5) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$

$= K + 2K + 2K + 3K$

$= 8K = \frac{8}{10} = 0.8$

$P(0 \leq X \leq 4) = P(X=0) + P(X=1) + P(X=2) +$
$P(X=3) + P(X=4)$

$= 0 + K + 2K + 2K + 3K$

$= 8K = 8\left(\frac{1}{10}\right) = 0.8$

(iv) Mean $(\mu) = \sum_{i=0}^{7} f_i x_i$

$= 0(0) + 1(K) + 2(2K) + 3(2K) + 4(3K)$
$+ 5(K^2) + 6(2K^2) + 7(7K^2 + K)$

$$= 66K^2 + 30K$$

$$= \frac{66}{100} + \frac{30}{10} = 0.66 + 3 = 3.66 \quad \left[\because K = \frac{1}{10}\right]$$

v) Variance $= \sum_{i=0}^{7} P_i f_i^2 - \mu^2$

$$= K + 8K + 18K + 48K + 25K^2 + 72K^2 + 343K^2 + 49K - (3.66)^2$$

$$= 440K^2 + 124K - (3.66)^2$$

$$= \frac{440}{100} + \frac{124}{10} - (3.66)^2$$

$$= 4.4 + 12.4 - 13.3956$$

$$= 3.4044.$$

Example: Two dice are thrown. Let X assign to each point (a,b) in S the maximum of its numbers i.e, $X(a,b) = \max(a,b)$. Find the probability distribution. X is a random variable with $X(S) = \{1, 2, 3, 4, 5, 6\}$. Also find the mean and variance of the distribution.

Solution: The total no. of cases are = 36

The maximum number could be 1,2,3,4,5,6

i.e, $X(S) = X(a,b) = \max(a,b)$.

$P(1) = P(X=1) = P(1,1) = \frac{1}{36}$

$P(2) = P(X=2) = \frac{3}{36}$ (∵ for maximum 2)

$P(3) = P(X=3) = \frac{5}{36}$ (∵ for max 3)

$P(4) = P(X=4) = \frac{7}{36}$ (∵ for max 4)

$P(5) = P(X=5) = \frac{9}{36}$ (∵ for max 5)

$P(6) = P(X=6) = \frac{11}{36}$ (∵ for max 6)

∴ The required probability distribution is

X	1	2	3	4	5	6
P(X)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

(i) Mean (μ) $= \sum_{i=1}^{6} p_i x_i$

$= 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36} + 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36}$

$$= \frac{1}{36}(1+6+15+28+45+66)$$

$$= \frac{161}{36} = 4.47$$

(ii) Variance $(\sigma^2) = \sum_{i=1}^{6} p_i x_i^2 - \mu^2$

$$= \frac{1}{36}(1)^2 + \frac{3}{36}(2)^2 + \frac{5}{36}(3)^2 + \frac{7}{36}(4)^2 + \frac{9}{36}(5)^2 + \frac{11}{36}(6)^2 - (4.47)^2$$

$$= \frac{1}{36}(1+12+45+112+225+396) - (4.47)^2$$

$$= \frac{791}{36} - 19.9808$$

$$= 21.97 - 19.981 = 1.9912.$$

Example: Find the mean and variance of the uniform probability distribution given by $f(x) = \frac{1}{n}$ for $x = 1, 2, 3 \ldots n$.

Solution: The probability distribution is

x	1	2	3	------	n
$f(x)$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	------	$\frac{1}{n}$

(i) Mean (μ) $= \sum_{i=1}^{n} x_i f(x_i)$

$= 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n \cdot \frac{1}{n}$

$= \frac{1}{n}(1+2+3+\cdots+n)$

$= \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$

(ii) Variance $= \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2$

$= 1^2 \cdot \frac{1}{n} + 2^2 \cdot \frac{1}{n} + 3^2 \cdot \frac{1}{n} + \cdots + n^2 \cdot \frac{1}{n} - \left(\frac{n+1}{2}\right)^2$

$= \frac{1}{n}(1^2+2^2+3^2+\cdots+n^2) - \frac{1}{4}(n+1)^2$

$= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \frac{1}{4}(n+1)^2$

$= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$

$= \frac{(n+1)}{2}\left(\frac{2n+1}{3} - \frac{n+1}{2}\right)$

$= \frac{n+1}{12}(4n+2-3n-3)$

$= \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}$

Example: From a lot of 10 items containing 3 defectives, a sample of 4 items is drawn at random. Let the random variable X denote the number of defective items in the sample. Find the probability distribution of X when the sample is drawn without replacement.

Solution: X takes the values 0, 1, 2 or 3

Given total no. of items = 10

No. of good items = 7

No. of defective items = 3

$P(X=0) = P(\text{No defective})$

$= \frac{^7C_4}{^{10}C_4} = \frac{7!}{4!3!} \times \frac{4!6!}{10!} = \frac{1}{6}$

$P(X=1) = P(\text{one defective and 3 good})$

$= \frac{^3C_1 \times ^7C_3}{^{10}C_4} = \frac{3 \times 7!}{3!\,4!} \times \frac{4!6!}{10!} = \frac{1}{2}$

$P(X=2) = P(\text{2 defective and 2 good})$

$= \frac{^3C_2 \times ^7C_2}{^{10}C_4} = \frac{3}{10}$

$P(X=3) = P(3 \text{ defective and } 1 \text{ good})$

$$= \frac{{}^3C_3 \times {}^7C_1}{{}^{10}C_4} = \frac{7}{{}^{10}C_4} = \frac{4!}{8\times 9\times 10} = \frac{1}{30}$$

∴ The Probability distribution of random variable X is as follows:

X	0	1	2	3
P(X)	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

Example: Find the distribution function which corresponds to the probability distribution defined by $f(x) = \frac{x}{15}$ for $x = 1, 2, 3, 4, 5$.

Solution:

Given $f(x) = \frac{x}{15}$

$\Rightarrow f(1) = \frac{1}{15}, f(2) = \frac{2}{15}, f(3) = \frac{3}{15}$

$f(4) = \frac{4}{15}, f(5) = \frac{5}{15}$

Now, $F(1) = f(1) = \frac{1}{15}$,

$F(2) = F(1) + f(2) = \frac{1}{15} + \frac{2}{15} = \frac{1}{5}$

$F(3) = F(2) + f(3) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}$

62

$$F(4) = F(3) + f(4) = \frac{2}{5} + \frac{4}{15} = \frac{2}{3}$$

and $F(5) = F(4) + f(5) = \frac{2}{3} + \frac{1}{3} = 1$

2.10 General properties of a Continuous Probability Distribution:

(i) Mean:

Mean of a continuous probability distribution is given by

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

If X is defined from a to b,

Then, $\mu = \int_{a}^{b} x f(x) dx$

(ii) Variance:

Variance of a continuous probability distribution is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

(or)

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

If X is defined from a to b,

Then, $\sigma^2 = \int_{a}^{b} x^2 f(x)\, dx - \mu^2$

(iii) Mode:

Mode is the value of x for which $f(x)$ is maximum. Thus mode is given by $f'(x)=0$ and $f''(x)<0$ for $a<x<b$.

(iv) Median:

Median is the point which divides the entire distribution into two equal parts. For continuous distribution, median is a point which divides the total area into two equal parts.

If X is defined a to b and M is the median, then

$$\int_a^M f(x)\,dx = \int_M^b f(x)\,dx = \frac{1}{2}$$

(v) Mean deviation:

Mean deviation about the mean (μ) is given by $\int_{-\infty}^{\infty} |x-\mu| f(x)\,dx$.

Example: For the continuous probability function $f(x) = Kx^2e^{-x}$ when $x \geq 0$,

find (i) K

(ii) Mean

(iii) Variance

Solution: (i) we know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_0^{\infty} Kx^2e^{-x}\,dx = 1 \quad (\because x \geq 0)$$

$$\Rightarrow K\left[x^2(-e^{-x}) - 2x(e^{-x}) + 2(-e^{-x})\right]_0^{\infty} = 1$$

$$\Rightarrow K\left\{-e^{-x}(x^2+2x+2)\right\}_0^{\infty} = 1$$

$$\Rightarrow K(0+2) = 1$$

$$\Rightarrow \boxed{K = \frac{1}{2}}$$

(ii) Mean $= \int_{-\infty}^{\infty} x f(x)\,dx$

$$= \int_0^{\infty} Kx^3e^{-x}\,dx$$

$$= K\left[x^3(-e^{-x}) - 3x^2(e^{-x}) + 6x(-e^{-x}) - 6e^{-x}\right]_0^{\infty}$$

$$= K\left[-e^{-x}(x^3+3x^2+6x+6)\right]_0^\infty$$

$$= K[0+6] = 6K = 6\left(\frac{1}{2}\right) = 3$$

$$\therefore \boxed{\mu = 3}$$

(iii) Variance $= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$

$$= \int_0^\infty x^2 \cdot Kx^2 e^{-x}\,dx - 3^2$$

$$= K\int_0^\infty x^4 e^{-x}\,dx - 9$$

$$= K\left[x^4(-e^{-x}) - 4x^3(e^{-x}) + 12x^2(-e^{-x}) - 24x(-e^{-x}) + 24e^{-x}\right]_0^\infty - 9$$

$$= \frac{1}{2}\left[-e^{-x}(x^4+4x^3+12x^2+24x+24)\right]_0^\infty - 9 \qquad \left[\because K = \frac{1}{2}\right]$$

$$= \frac{1}{2}[0+24] - 9$$

$$= 12 - 9 = 3$$

Example: Suppose a continuous random variable X has the probability density $f(x) = K(1-x^2)$ for $0 < x < 1$, and $f(x) = 0$ otherwise. Find (i) K (ii) mean (iii) variance.

Solution: (i) We know that, $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\text{i.e, } \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$\text{i.e, } 0 + \int_{0}^{1} K(1-x^2)\,dx + 0 = 1$$

$$\text{i.e, } K\left(x - \frac{x^3}{3}\right)_0^1 = 1$$

$$\text{i.e, } K\left(1 - \frac{1}{3}\right) = 1$$

$$\therefore \boxed{K = \frac{3}{2}}$$

(ii) Mean $= \mu = \int_{-\infty}^{\infty} x f(x)\,dx$

$$= \int_{0}^{1} x f(x)\,dx$$

$$= \int_{0}^{1} x \cdot K(1-x^2)\,dx$$

$$\mu = K\int_{0}^{1} (x - x^3)\,dx$$

$$\Rightarrow \mu = K\left(\frac{x^2}{2} - \frac{x^4}{4}\right)_0^1 = K\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{K}{4}$$

$$\mu = \left(\frac{3}{2}\right)\left(\frac{1}{4}\right) = \frac{3}{8} \quad \left[\because K = \frac{3}{2}\right]$$

$$\therefore \boxed{\mu = \frac{3}{8}}$$

(iii) Variance$(\sigma^2) = \int_0^1 x^2 f(x)\,dx - \mu^2$

$= \int_0^1 x^2 K(1-x^2)\,dx - \left(\frac{3}{8}\right)^2$

$= K\int_0^1 (x^2 - x^4)\,dx - \frac{9}{64}$

$\sigma^2 = K\left(\frac{x^3}{3} - \frac{x^5}{5}\right)_0^1 - \frac{9}{64}$

$= K\left(\frac{1}{3} - \frac{1}{5}\right) - \frac{9}{64}$

$= \frac{2K}{15} - \frac{9}{64}$

$= \frac{2}{15}\left(\frac{3}{2}\right) - \frac{9}{64}$

$= \frac{1}{5} - \frac{9}{64}$

$= \frac{19}{320} = 0.06$

$\therefore \boxed{\sigma^2 = 0.06}$

Example: The Probability density $f(x)$ of a continuous random variable is given by $f(x) = ce^{-|x|}$, $-\infty < x < \infty$.
Find (i) c (ii) mean (iii) variance (iv) $P(0 \le X \le 4)$

Solution: Given function $f(x) = c e^{-|x|}$, $-\infty < x < \infty$

(i) We know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-\infty}^{\infty} c e^{-|x|}\,dx = 1$$

$$\Rightarrow c \int_{-\infty}^{\infty} e^{-|x|}\,dx = 1$$

$$\Rightarrow 2c \int_{0}^{\infty} e^{-|x|}\,dx = 1 \quad [\because e^{-|x|} \text{ is even}]$$

$$\Rightarrow 2c \int_{0}^{\infty} e^{-x}\,dx = 1 \quad [\because \text{in } 0 \le x \le \infty, |x| = x]$$

$$\Rightarrow 2c(-e^{-x})_0^{\infty} = 1$$

$$\Rightarrow -2c(0-1) = 1$$

$$\Rightarrow 2c = 1$$

$$\Rightarrow \boxed{c = \frac{1}{2}}$$

$$\therefore f(x) = \frac{1}{2} e^{-|x|}$$

(ii) Mean $(\mu) = \int_{-\infty}^{\infty} x f(x)\,dx$

$$= \frac{1}{2} \int_{-\infty}^{\infty} x e^{-|x|}\,dx = 0 \quad [\because \text{Integrand is odd}]$$

(ii) Variance $= \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$

$= \int_{-\infty}^{\infty} (x-0)^2 \frac{1}{2} e^{-|x|}\,dx$

$= \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-|x|}\,dx$

$= 2 \cdot \frac{1}{2} \int_{0}^{\infty} x^2 e^{-|x|}\,dx$

[$\because$ Integrand is even]

$= \int_{0}^{\infty} x^2 e^{-x}\,dx$

$= \left(x^2 \frac{e^{-x}}{-1} - 2x \frac{e^{-x}}{1} + 2 \frac{e^{-x}}{(-1)} \right)_0^{\infty}$

$= (0 - (-2)) = 2$

(iii) $P(0 \le X \le 4) = \frac{1}{2} \int_{0}^{4} e^{-|x|}\,dx$

$= \frac{1}{2} \int_{0}^{4} e^{-x}\,dx$ [$\because$ In $0 < x < 4$, $|x| = x$]

$= -\frac{1}{2} (e^{-x})_0^4$

$= -\frac{1}{2} (e^{-4} - 1)$

$= \frac{1}{2} (1 - e^{-4}) = 0.4908$

Example: A continuous random variable has the probability density function

$$f(x) = \begin{cases} Kxe^{-\lambda x}, & \text{for } x \geq 0, \lambda > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find (i) K
(ii) mean
(iii) variance

Solution: (i) we know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-\infty}^{0} 0 \cdot dx + \int_{0}^{\infty} Kxe^{-\lambda x}\,dx = 1$$

$$\Rightarrow K\int_{0}^{\infty} xe^{-\lambda x}\,dx = 1$$

$$\Rightarrow K\left[x\left(\frac{e^{-\lambda x}}{(-\lambda)}\right) - 1\left(\frac{e^{-\lambda x}}{\lambda^2}\right)\right]_0^{\infty} = 1$$

$$\Rightarrow K\left[(0-0) - \left(0 - \frac{1}{\lambda^2}\right)\right] = 1$$

$$\Rightarrow \boxed{K = \lambda^2}$$

$$\therefore f(x) = \begin{cases} \lambda^2 xe^{-\lambda x}, & \text{for } x \geq 0, \lambda > 0 \\ 0, & \text{otherwise} \end{cases}$$

(ii) Mean $= \mu = \int_{-\infty}^{\infty} x f(x)\,dx$

$\Rightarrow \mu = \int_{-\infty}^{0} 0 \cdot dx + \int_{0}^{\infty} x \lambda^2 x e^{-\lambda x}\,dx$

$= \lambda^2 \int_0^{\infty} x^2 e^{-\lambda x}\,dx$

$= \lambda^2 \left[x^2 \left(\frac{e^{-\lambda x}}{(-\lambda)}\right) - 2x\left(\frac{e^{-\lambda x}}{\lambda^2}\right) + 2\left(\frac{e^{-\lambda x}}{-\lambda^3}\right)\right]_0^{\infty}$

$= \lambda^2 \left[(0 - 0 + 0) - \left(0 - 0 - \frac{2}{\lambda^3}\right)\right]$

$$\boxed{\mu = \frac{2}{\lambda}}$$

(iii) Variance $= \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$

$\Rightarrow \sigma^2 = \int_0^{\infty} x^2 f(x)\,dx - \left(\frac{2}{\lambda}\right)^2$

$= \lambda^2 \int_0^{\infty} x^3 e^{-\lambda x}\,dx - \frac{4}{\lambda^2}$

$= \lambda^2 \left[x^3 \left(\frac{e^{-\lambda x}}{-\lambda x}\right) - 3x^2\left(\frac{e^{-\lambda x}}{\lambda^2}\right) + 6x\left(\frac{e^{-\lambda x}}{-\lambda^3}\right) - 6\left(\frac{e^{-\lambda x}}{\lambda^4}\right)\right]_0^{\infty}$

$= \lambda^2 \left[(0 - 0 + 0 - 0) - \left(0 - 0 + 0 - \frac{6}{\lambda^4}\right)\right] - \frac{4}{\lambda^2}$

$\sigma^2 = \frac{6}{\lambda^2} - \frac{4}{\lambda^2} = \frac{2}{\lambda^2}$

Example: A continuous random variable X has the distribution function

$$F(x) = \begin{cases} 0, & \text{if } x \le 1 \\ K(x-1)^4, & \text{if } 1 < x \le 3 \\ 1, & \text{if } x > 3 \end{cases}$$

Find (i) $f(x)$

(ii) K

(iii) mean

Solution: (i) we know that $f(x) = \frac{d}{dx}[F(x)]$

$$\therefore f(x) = \begin{cases} 0, & \text{if } x \le 1 \\ 4K(x-1)^3, & \text{if } 1 < x \le 3 \\ 0, & \text{if } x > 3 \end{cases}$$

(ii) we know that $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-\infty}^{1} f(x)\,dx + \int_{1}^{3} f(x)\,dx + \int_{3}^{\infty} f(x)\,dx = 1$$

$$\Rightarrow 0 + \int_{1}^{3} 4K(x-1)^3\,dx + 0 = 1$$

$$\Rightarrow 4K\left[\frac{(x-1)^4}{4}\right]_1^3 = 1$$

$\Rightarrow K(16-0) = 1$

$\Rightarrow \boxed{K = \frac{1}{16}}$

$$\therefore f(x) = \begin{cases} 0 & , \text{if } x \leq 1 \\ \frac{1}{4}(x-1)^3 & , \text{if } 1 < x \leq 3 \\ 0 & , \text{if } x > 3 \end{cases}$$

(iii) Mean $= \mu = \int_{-\infty}^{\infty} x f(x)\, dx$

$= \int_{-\infty}^{1} x f(x)\, dx + \int_{1}^{3} x \cdot f(x)\, dx + \int_{3}^{\infty} x \cdot f(x)\, dx$

$= 0 + \int_{1}^{3} x \cdot \frac{1}{4}(x-1)^3 dx + 0$

$= \frac{1}{4}\int_{1}^{3} x(x-1)^3 dx$

$= \frac{1}{4}\int_{0}^{2} (t+1)t^3 dt \quad [\because \text{Put } x-1=t]$

$= \frac{1}{4}\int_{0}^{2} (t^4+t^3)\, dt$

$= \frac{1}{4}\left(\frac{t^5}{5} + \frac{t^4}{4}\right)_0^2$

$= \frac{1}{4}\left(\frac{2^5}{5} + \frac{2^4}{4}\right) = \frac{2^4}{4}\left(\frac{2}{5} + \frac{1}{4}\right)$

$= 4\left(\frac{13}{20}\right) = \frac{13}{5} = 2.6$

Example: If X is a continuous random variable and $Y = aX + b$, Prove that $E(Y) = aE(X) + b$ and $V(Y) = a^2 V(X)$ where V is variance and a, b are constants.

Solution: (i) By definition,

$$E(Y) = E(aX+b) = \int_{-\infty}^{\infty} (ax+b) f(x)\,dx$$

$$= a\int_{-\infty}^{\infty} x f(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx$$

$$= aE(X) + b(1) \quad [\because \text{Total Probability is unity}]$$

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

(ii) we have $E(Y) = aE(X) + b$ — ① $[\because \text{from (i)}]$

where $Y = aX + b$ — ②

② − ① $\Rightarrow Y - E(Y) = a[X - E(X)]$

squaring on Both sides

$$[Y - E(Y)]^2 = a^2 [X - E(X)]^2$$

Taking Expectation on both sides

we get $E\{[Y - E(Y)]^2\} = a^2 E\{[X - E(X)]^2\}$

$$\therefore V(Y) = a^2 V(X)$$

Example: If X is a continuous random variable and K is a constant, then prove that (i) $Var(X+K) = Var(X)$

(ii) $Var(KX) = K^2 Var(X)$

Solution:

By definition,

$$Var(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left[\int_{-\infty}^{\infty} x f(x)\,dx\right]^2$$

(i) $$Var(X+K) = \int_{-\infty}^{\infty} (x+K)^2 f(x)\,dx - \left[\int_{-\infty}^{\infty} (x+K) f(x)\,dx\right]^2$$

$$= \int_{-\infty}^{\infty} (x^2 + 2Kx + K^2) f(x)\,dx - \left[\int_{-\infty}^{\infty} x f(x)\,dx + K\int_{-\infty}^{\infty} f(x)\,dx\right]^2$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,dx + 2K\int_{-\infty}^{\infty} x f(x)\,dx + K^2 - \left[\int_{-\infty}^{\infty} x f(x)\,dx + K\right]^2$$

$$\left[\because \int_{-\infty}^{\infty} f(x)\,dx = 1\right]$$

$$= E(X^2) + 2KE(X) + K^2 - [E(X) + K]^2$$

$$= E(X^2) + 2KE(X) + K^2 - [E(X)]^2 - 2KE(X) - K^2$$

$$= E(X^2) - [E(X)]^2$$

$$= \text{var}(X)$$

(ii) $$\text{var}(KX) = \int_{-\infty}^{\infty} K^2 x^2 f(x)\,dx - \left[\int_{-\infty}^{\infty} Kx f(x)\,dx\right]^2$$

$$= K^2 \int_{-\infty}^{\infty} x^2 f(x)\,dx - K^2 \left[\int_{-\infty}^{\infty} f(x)\,dx\right]^2$$

$$= K^2 [E(X)^2 - \{E(X)\}^2]$$

$$= K^2 \text{var}(X)$$

2.11 Bionomial Distribution

Bionomial Distribution was discovered by James Bernoulli in the year 1700 and it is a Discrete Probability distribution.

Let us discuss a conceptual or particular situation where a trial or an experiment results in only two outcomes, say 'success' and 'Failure'. Further, the result of one trial does not influence the result of next trial, and the probability of success at each trial is the same from trial to trial.

Some of such situations are;

1. Tossing a coin – Head or Tail
2. Birth of a baby – Girl or Boy.
3. Auditing a Bill – Contains an error or not.

The conditions for the applicability of a Bionomial distribution are as follows,

i) There are n independent trials
ii) Each trial has only two possible outcomes.

(iii) The probabilities of two outcomes remain constant.

→ Let the number of trials be n. The trials be independent i.e the success or failure at one trial does not affect the outcome of the other trials. Thus the probability of success remains the same from trial to trial. Also let 'p' be the probability of success and 'q' be the probability of failure. Then we have $p+q=1$ or $q=1-p$.

The probability of getting such a sequence of r successes and $(n-r)$ failures is $= p^r q^{n-r}$ (using Multiplication Theorem on Probability)

However, the number of ways in which one can get r successes and n-r failures is nCr, and the probability of each of there ways is $p^r q^{n-r}$.

If x is the random variable, representing the number of successes, the probability of getting r successes and $n-r$ failures, in 'n' trials, is given by the probability function

$$P(x=r) = {}^nC_r\, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots n.$$

The probability of the number of successes so obtained is called the Bionomial Probability Distribution, because these probabilities are the successive terms in the expansion of the bionomial $(q+p)^n$. 'n' is known as the degree of the distribution

Definition: A random variable x has a Bionomial distribution if it assumes only non negative values and its probability density function is given by

$$P(X=r) = P(r) = \begin{cases} {}^nC_r\, p^r q^{n-r}; & r = 0, 1, 2, \ldots \\ & q = 1-p \\ 0; & \text{otherwise} \end{cases}$$

→ The Binomial distribution function is given by $F_x(x) = P(X \leq x) = \sum_{r=0}^{n} {}^nC_r p^r q^{n-r}$

Examples of Binomial Distribution:

i) The number of defective bolts in a box containing 'n' bolts.

ii) The number of post-graduates in a group of 'n' men.

Conditions of Binomial Distribution:

1. Trials are repeated under identical conditions for a fixed number of times, say n times.

2. There are only two possible outcomes, eg. success or failure for each trial.

3. The probability of success in each trial remains constant and does not change from trial to trial.

4. The trials are independent i.e, the probability of an event in any trial is not affected by the results of any other trial.

2.11.1 Constants of Binomial Distribution:

① Mean of the Binomial Distribution:

The Binomial probability Distribution is given by

$$P(r) = {}^nC_r\, p^r q^{n-r}; \quad r = 0, 1, 2, \ldots n \text{ and } q = 1-p.$$

Mean of X, $\mu = E(X) = \sum_{r=0}^{n} r\, P(r)$

$$= 0 \times q^n + 1 \times {}^nC_1\, p q^{n-1} + 2 \times p^2 + \cdots + n.p^n.$$

$$= npq^{n-1} + 2.\frac{n(n-1)}{2!} p^2 q^{n-2} + \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \cdots + np^n.$$

$$= np(q+p)^{n-1} \quad \text{(using Binomial distribution)}$$

$$= np(1) = np.$$

Hence Arithmatic mean of the Binomial distribution $= np$.

(2) Variance of the Binomial Distribution.

Variance, $V(X) = E(X^2) - [E(X)]^2$

$\therefore \boxed{V(X) = npq}$.

Hence the standard Deviation of the Binomial Distribution $= \sqrt{npq}$

(3). Mode of the Binomial Distribution:

Mode of the binomial distribution is the Value of x at which $p(x)$ has max. value.

$$\text{Mode} = \begin{cases} \text{integral part of } (n+1)p, & \text{if } (n+1)p \text{ is not an integer.} \\ (n+1)p \text{ and } (n+1)(p-1), & \text{,,} \end{cases}$$

Recurrence Relation for the Binomial Distribution:

$$P(r+1) = \frac{(n-r)p}{(r+1)q} \cdot P(r)$$

Example: A fair coin is tossed six times. Find the probability of getting four heads.

Solution: P = Probability of getting a head $= 1/2$

q = probability of not getting head $= 1/2$ and $n = 6, r = 4$

We know that $P(r) = {}^nC_r\, p^r q^{n-r}$

$\therefore P(4) = {}^6C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{6-4}$

$= \frac{6!}{4!2!} \left(\frac{1}{2}\right)^6$

$= \frac{6 \times 5}{2} \times \frac{1}{2^6} = \frac{15}{64} = 0.2344$

Example: Determine the probability of getting the sum 6 exactly 3 times in 7 throws with a pair of fair dice.

Solution: In a single throw of a pair of fair dice, a sum of 6 can occur in 5 ways, (1,5), (5,1), (2,4), (4,2) and (3,3) out of $6 \times 6 = 36$ ways.

Thus p = Probability of occurrence of 6 in one throw $= \frac{5}{36}$

$q = 1 - p = 1 - \frac{5}{36} = \frac{31}{36}$

n = Number of trials $= 7$

∴ Probability of getting 6 exactly thrice in 7 throws

$$= {}^7C_3 \, p^3 q^{7-3}$$

$$= {}^7C_3 \left(\frac{5}{36}\right)^3 \left(\frac{31}{36}\right)^4$$

$$= \frac{35\,(125)(31)^4}{(36)^7} = 0.0516 \text{ (nearly)}$$

Example: Ten coins are thrown simultaneou Find the probability of getting at least

(i) Seven heads

(ii) Six heads

(iii) One head.

Solution: p = Probability of getting a head $= \frac{1}{2}$

q = Probability of not getting a head $= \frac{1}{2}$

The probability of getting r heads in a throw of 10 coins is

$$P(X=r) = P(r) = {}^{10}C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r};$$

$$r = 0, 1, 2, \ldots\ldots 10$$

(i) Probability of getting at least seven heads is given by

$$P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10)$$

$$= {}^{10}C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 + {}^{10}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2 + {}^{10}C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right) + {}^{10}C_{10} \left(\frac{1}{2}\right)^{10}.$$

$$= \frac{1}{2^{10}} \left[{}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right]$$

$$= \frac{1}{2^{10}} (120 + 45 + 10 + 1)$$

$$= \frac{176}{1024} = 0.1719.$$

(ii) This is, ~~left~~ as an

$$P(X \geq 6) = \frac{1}{2^{10}} ({}^{10}C_6 + {}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 +$$

(iii) $P(\text{at least 1 head}) = P(r \geq 1)$

$= 1 - P(r=0)$

$= 1 - {}^{10}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{10-}$

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

$= 1 - \frac{1}{2^{10}}$.

Example: The mean and Variance of a binomial distribution are 2 and 8/5. Find n.

Solution: We are given,

Mean of the Binomial distribution = 2 i.e

$np = 2$ — (1)

Variance of the Binomial distribution $= \frac{8}{5}$ i.e

$npq = \frac{8}{5}$ — (2)

$\frac{(2)}{(1)} \Rightarrow \frac{npq}{np} = \frac{8/5}{2} = 4/5 \Rightarrow q = 4/5$

$p = 1 - q = 1 - 4/5 = 1/5$.

$n = \frac{2}{p} = \frac{2}{(1/5)} = 10 \quad \therefore n = 10.$

Example:

In eight throws of a die 5 or 6 is considered a success. Find the mean number of success and the standard deviation.

Solution: p = The probability of

Success $= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$.

q = The probability of failure

$= 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$.

n = Number of throws $= 8$

$\therefore$ Mean $= np = 8\left(\frac{1}{3}\right) = \frac{8}{3}$

Variance $= npq = (np)q = \left(\frac{8}{3}\right)\left(\frac{2}{3}\right)$

$= \frac{16}{9}$

Hence standard deviation $= \sqrt{\text{Variance}}$

$= \sqrt{16/9} = 4/3$

Example: Fit a binomial distribution to the following frequency distribution

x	0	1	2	3	4	5	6
f	13	25	52	58	32	16	4

Here n = number of trials = 6 and

N = total frequency = $\Sigma fi = 200$

$\therefore$ Mean = $\frac{\Sigma fixi}{\Sigma fi} = \frac{25+104+174+128+80+24}{200}$

$= \frac{535}{200} = 2.675$

Now mean of the binomial distribution $= np$

i.e $np = 6p = 2.675$

$\therefore p = \frac{2.675}{6} = 0.446$

and $q = 1-p = 1-0.446 = 0.554$

Hence the bionomial distribution to be fitted is given by

$N(q+p)^n = 200(0.554+0.446)^6$

$= 200(0.554+0.446)^6$

$= 200[(0.554)^6 + {}^6C_1(0.554)^5(0.4$

$+ {}^6C_2(0.554)^4(0.446)^2 + {}^6C_3(0.554)^3(0.446)^3 +$

${}^6C_4(0.554)^2(0.446)^4 + {}^6C_5(0.554)(0.446)^5 +$

$+ {}^6C_6 (0.446)^6]$

$= 200[0.02891 + 0.1396 + 0.2809 + 0.3016$
$+ 0.1821 + 0.05864 + 0.007866]$

$= 5.782 + 27.92 + 56.18 + 60.32 + 36.42$
$+ 11.728 + 1.5732$

The expected frequencies can be rounded off to the nearest integer to get expected frequencies as whole numbers.

∴ The successive terms in the expansion give the expected or theoretical frequencies which are

x	0	1	2	3	4	5	6
f	13	25	52	58	32	16	4
Expected or Theoretical frequency	6	28	56	60	36	12	2

2.12

Poisson Distribution

S.D Poisson introduced poisson distribution as a rare distribution of rare events i.e the events whose probability of occurence is very small but the number of trials which could lead to the occurence of the event, are very large.

Definition of Poisson Distribution:

A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and it probabilit density function is given by

$$P(x, \lambda) = P(X=x) = \begin{cases} e^{-\lambda}\lambda^{x} ; x=0,1,2,\ldots \\ 0 \quad , \text{otherwise} \end{cases}$$

Here $\lambda > 0$ is called the parameter of the distribution.

Note: 1. It should be noted that

$$\sum_{x=0}^{\infty} P(X=x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Hence equation ① is a probability function.

2. The Poisson Distribution function is

$$F_x(x) = P(X \leq x) = \sum_{r=0}^{x} P(r)$$

$$= e^{-\lambda} \sum_{r=0}^{x} \frac{\lambda^r}{r!}, \quad x = 0, 1, 2, \ldots$$

Examples of Poisson Distribution:

(i) The number of defective electric bulbs manufactured by a reputed company.

ii) The number of telephone calls per minute at a switch board.

iii) The number of particales emitted by a radio-active substance.

iv) The number of persons born blind per year in a large city.

2.12.1:

Constants of the Poisson Distribution

1. The Mean of the Poisson Distribution.

$$\text{Mean} = E(X) = \sum_{x=0}^{\infty} x.P(x) = \sum_{x=0}^{\infty} x.\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \quad \left[\because x! = x(x-1)!\right]$$

$$= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y+1}}{y!} \quad (\text{putting } y = x-1)$$

$$= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda}.\lambda.e^{\lambda}$$

$$\left[\because \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda}\right]$$

$$= \lambda \ (= np)$$

Thus the parameter λ is the Arithmetic Mean of the Poisson Distribution.

2. Variance of Poisson Distribution:

$$V(X) = E(X^2) - [E(X)]^2$$

$$= \sum_{x=0}^{\infty} x^2 p(x) - \lambda^2 \quad [\because \lambda = \text{mean of P.D}]$$

$$= \sum_{x=0}^{\infty} x^2 \cdot \frac{e^{-\lambda} \lambda^x}{x!} - \lambda^2$$

$$= e^{-\lambda} \sum_{x=1}^{\infty} \frac{x \lambda^x}{(x-1)!} - \lambda^2$$

$$= e^{-\lambda} \sum_{x=1}^{\infty} [(x-1)+1] \cdot \frac{\lambda^x}{(x-1)!} - \lambda^2$$

$$= e^{-\lambda} \left[\sum_{x=1}^{\infty} (x-1) \frac{\lambda^x}{(x-1)(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \right] - \lambda^2$$

$$= e^{-\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \right] - \lambda^2$$

$$= e^{-\lambda} \left[\sum_{y=0}^{\infty} \frac{\lambda^{y+2}}{y!} + \sum_{z=0}^{\infty} \frac{\lambda^{z+1}}{z!} \right] - \lambda^2$$

(putting $y = x-2$, $z = x-1$)

$$= e^{-\lambda} \left[\lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + \lambda \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \right] - \lambda^2$$

$$= e^{-\lambda}\left[\lambda^2 . e^{\lambda} + \lambda e^{\lambda}\right] - \lambda^2$$

$$= (\lambda^2 + \lambda) - \lambda^2$$

$$= \lambda .$$

Thus Variance $= \lambda$

Hence the variance of the distribution

= Mean of the distribution $= \lambda$

Further, standard deviation of the Poisson Distribution, $\sigma = \sqrt{\lambda}$

3. Mode of the Poisson Distribution:

Mode is the value of x for which the Probability $P(x)$ is maximum.

$\therefore$ $P(x) \geq P(x+1)$ and $P(x) \geq P(x-1)$

Now $P(x) \geq P(x+1) \Rightarrow \dfrac{e^{-\lambda} \cdot \lambda^x}{x!} \geq \dfrac{e^{-\lambda} \cdot \lambda^{x+1}}{(x+1)!}$

$\Rightarrow 1 \geq \dfrac{\lambda}{x+1}$ or $\dfrac{\lambda}{x+1} \leq 1$

$\Rightarrow \lambda \leq x+1$ or $x+1 \geq \lambda .$

$\Rightarrow x \geq \lambda - 1$

Similarly $P(x) \geq P(x-1)$

$\Rightarrow x \leq \lambda$

Combining (1) and (2), we have

$\lambda - 1 \leq x \leq \lambda$

Hence mode of the Poisson distribution lies between $\lambda - 1$ and λ.

Case (1): If λ is an integer then $\lambda - 1$ is also an integer. So we have two maximum values and the distribution is bimodel and two modes are $(\lambda - 1)$ and λ.

Case (2): If λ is not an integer, the mode of Poisson distribution is integral part of λ.

4) Recurrence Relation for the Poisson Distribution:

we have $P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$

$$P(x+1) = \frac{e^{-\lambda}\cdot\lambda^{x+1}}{(x+1)!} = \frac{\lambda}{x+1}\cdot\frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{\lambda}{x+1}P(x$$

Thus $P(x+1) = \left(\frac{\lambda}{x+1}\right)P(x)$

(or) $P(x) = \frac{\lambda}{x}\cdot P(x-1)$

which is the required relation. which this formula we can find $P(1), P(2), P(3);\cdots$ if $P(0)$ is given.

Note: 1. on successive application of the formula, we get

$$P(x+1) = \frac{\lambda^{x+1}}{(x+1)!}\cdot P(0)$$

2. $P(X \le t) = \left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^t}{t!}\right]\cdot P(0$

3. $P(a \le X \le b)\left[\frac{\lambda^a}{a!} + \frac{\lambda^{a+1}}{(a+1)!} + \cdots + \frac{\lambda^b}{b!}\right]P($

for all $0 \le a, b,\ a, b \in Z^+$.

2.12.2

Properties of Poisson Distribution

1. Range of the variable is from 0 to ∞
2. Mean and variance are equal.
3. Distribution gets more and more symmetrical about the mean as λ increases and tends to normal distribution. described in the next section.

Example: A hospital switch board receives an average of 4 emergency calls in a 10 min interval. what is the probability that (i) There are at most 2 emergency calls in a 10 minute interval (ii) There are exactly 3 emergency calls in a 10 minute interval.

Mean, $\lambda = (4 \text{ calls} / 10 \text{ minutes}) = 4 \text{ calls}$

$\therefore P(X=x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$= \frac{e^{-4}4^x}{x!} = \frac{1}{e^4} \cdot \frac{4^x}{x!}$

(i) $P(\text{at most 2 calls}) = P(X \leq 2)$

$= P(X=0) + P(X=1)$

$+ P(X=2)$

$$= \frac{1}{e^4} + \frac{1}{e^4} \cdot 4 + \frac{1}{e^4} \cdot \frac{4^2}{2!}$$

$$= \frac{1}{e^4}(1+4+8) = 13e^{-4} = 0.2381$$

(ii) P(exactly 3 calls) = P(X = 3)

$$= \frac{1}{e^4} \cdot \frac{4^2}{3!} = \frac{32}{3} e^{-4}$$

$$= 0.1954$$

Example: If a bank received on the average 6 bad cheques per day, find the probability that it will receive 4 bad cheques on any given day.

Solution: we have $P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$

Here $\lambda = 6$

$\therefore P(X=4) = \frac{e^{-6} 6^4}{4!} = \frac{54}{e^6} = 0.1339$

Example: A manufacturer knows that the Condensers he makes contain on average 1% defectives. He packs them in a boxes of 100. What is the probability that a box

picked at random will contain 3 or more finally condensers?

Solution: Here p = Probability of defective condensers = $1\% = 0.01$

n = Total number of condencers $= 100$

$\therefore \lambda = \text{mean} = np = 100(0.01) = 1.$

$$P(X=x) = p(x) = \frac{e^{-\lambda}\cdot\lambda^x}{x!} = \frac{e^{-1}\cdot 1^x}{x!}$$

$$= \frac{e^{-1}}{x!}$$

$$P(X \geq 3) = 1 - P(X<3)$$

$$= 1 - [P(X=0) + P(X=1) + P(X=2)]$$

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

$$= 1 - e^{-1}\cdot\frac{5}{2}$$

$$= 1 - 0.9197$$

$$= 0.0803.$$

Example: A manufacturer of cotter pins knows that 5% of his product is defective. pins are sold in boxes of 100. He guarantees that not more than 10 pins will be defective. what is the approximate probability that a box will fail to meet the guaranteed quality?

Solution: The probability of cotter pins to be defective $= p = 5\% = 0.05$

Total number of cotter pins, $n = 100$

$\therefore$ Mean, $\lambda = np = 100(0.05) = 5$

we have $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$

$\therefore P(X = x) = \frac{e^{-5} 5^x}{x!}$

P(a box will fail to meet the guarantee)

$= P(X > 10) = 1 - P(X \leq 10)$

$= 1 - [P(X=0) + P(X=1) + \cdots + P(X=10)]$

$= 1 - \left[\frac{e^{-5}(5)^0}{0!} + \frac{e^{-5}(5)^1}{1!} + \frac{e^{-5}(5)^2}{2!} + \frac{e^{-5}(5)^3}{3!} + \cdots \frac{e^{-5}(5)^{10}}{10!}\right] = 1 - 0.9863$

$= 0.0137$

Example: If a random variable has a Poisson distribution such that $P(1) = P(2)$. Find (i) mean of the distribution (ii) $P(4)$ (iii) $P(x \geq 1)$ (iv) $P(1 < x < 4)$

Solution: Given $P(1) = P(2)$

Hence $\frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-\lambda}\lambda^2}{2!}$

$\Rightarrow \lambda^2 = 2\lambda$

$\Rightarrow \lambda^2 - 2\lambda = 0 \Rightarrow \lambda(\lambda - 2) = 0$

$\therefore \lambda = 0$ or 2

(i) But $\lambda \neq 0$, $\therefore \lambda = 2$ Hence mean of the distribution $= 2$

(ii) $P(x) = \frac{e^{-\lambda}\cdot\lambda^x}{x!} = \frac{e^{-2}\cdot 2^x}{x!}$

$P(x = 4) = P(4) = \frac{e^{-2}\cdot 2^4}{4!} = \frac{2}{3e^2}$

$= 0.09022$

(iii) $P(x \geq 1) = 1 - P(x < 1) = 1 - P(x = 0)$

$= 1 - \frac{e^{-2}\cdot 2^0}{0!} = 1 - \frac{1}{e^2} = 0.8647.$

(iv) $P(1 < x < 4) = P(x = 2) + P(x = 3)$

$= \frac{e^{-2}\cdot 2^2}{2!} + \frac{e^{-2}\cdot 2^3}{3!} = 2e^{-2} + \frac{4}{3}e^{-2}$

$= 0.4511$

Example: Average number of accidents on any day on a national highway is 1.8, Determine the probability that the number of accidents are (i) at least one (ii) atmost one

Solution: Mean, $\lambda = 1.8$

We have $P(X=x) = P(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$

$$= \frac{e^{-1.8} \cdot (1.8)^x}{x!}$$

(i) $P(\text{at least one}) = P(X \geq 1)$

$= 1 - P(X=0)$

$= 1 - e^{-1.8}$

$= 1 - 0.1653$

$= 0.8347.$

(ii) $P(\text{at most one}) = P(X \leq 1)$

$= P(X=0) + P(X=1)$

$= e^{-1.8} + e^{-1.8}(1.8)$

$= e^{-1.8}(2.8)$

$= 0.4628.$

Example: Number of monthly breakdowns of a computer is a random variable having poisson distribution with mean equal to 1.8. Find the probability that the computer will function for a month (i) without a breakdown (ii) with only one breakdown and (iii) with at least one

Solution: Given mean, $\lambda = 1.8$

from poisson distribution

$$P(X=x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \frac{(e^{-1.8})(1.8)^x}{x!}$$

(i) $P(X=0) = \frac{(e^{-1.8})(1.8)^0}{0!} = e^{-1.8} = 0.1653$

(ii) $P(X=1) = \frac{(e^{-1.8})(1.8)^1}{1!} = e^{-1.8}(1.8) = 0.2975$

(iii) $(PX \geq 1) = 1 - P(X<1)$

$= 1 - P(X=0)$

$= 1 - 0.1653$

$= 0.8347.$

Example: A distribution of bean seeds determines from extensive tests that 5% of large batch of seeds will not germinate. He sells the seeds in packers of 200 and guarantees 90% germination. Determine the probability that a particular packet will violate the guarantee.

Solution: P = The probability of a seed not germinating = 5% = 0.05

λ = mean number of seeds in a sample of 200 = $np = 200 \times 0.05 = 10$

Let x be the number of seeds that do not germinate. Then

$$P(x) = \frac{e^{-10} \cdot 10^x}{x!}$$

A packet will violate guarantee if it contains more than 20 germination seeds.

Probability that the guarantee is violated $= P(X > 20)$

$= 1 - P(X \leq 20)$

$= \ldots$

$$= 1 - \sum_{x=0}^{20} \frac{e^{-10} 10^x}{x!}$$

$$= 1 - 0.9984$$

$$= 0.0016$$

Example: If a Poisson distribution is such that $P(X=1) \cdot \frac{3}{2} = P(X=3)$, find (i) $P(X \geq 1)$ (ii) $P(X \leq 3)$ (iii) $P(2 \leq X \leq 5)$

Solution: Given $\frac{3}{2} P(X=1) = P(X=3)$

i.e $\frac{3}{2} \cdot \frac{e^{-\lambda} \cdot \lambda^1}{1!} = \frac{e^{-\lambda} \cdot \lambda^3}{3!}$

i.e $\frac{3\lambda}{2} = \frac{\lambda^3}{6}$

i.e $\lambda^3 = 9\lambda \Rightarrow \lambda^3 - 9\lambda = 0$

$\Rightarrow \lambda(\lambda^2 - 9) = 0$

(or) $\lambda = 0, \lambda^2 = 9$

$\lambda = 0, \lambda = 3, -3.$

$\Rightarrow \underline{\lambda = 3}$ ($\because \lambda > 0$)

Hence $P(X=x) = P(x) = \frac{e^{-3} \cdot 3^x}{x!}$

(i) $P(X \geq 1) = 1 - P(X < 1)$

$= 1 - P(X = 0)$

$= 1 - \frac{e^{-3} \cdot 3^0}{0!}$

$= 1 - e^{-3}$

$= 0.950213.$

(ii) $P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$= e^{-3}\left[\frac{3^0}{0!} + \frac{3}{1!} + \frac{3^2}{2!} + \frac{3^3}{3!}\right]$

$= e^{-3}\left(1 + 3 + \frac{9}{2} + \frac{9}{2}\right)$

$= 13 e^{-3}$

$= 0.6472318.$

(iii) $P(2 \leq X \leq 5) = P(X=2) + P(X=3) + P(X=4) + P(X=5)$

$= e^{-3}\left[\frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!} + \frac{3^5}{5!}\right]$

$= 9e^{-3}\left(\frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{9}{40}\right)$

$= 9e^{-3}(1.6)$

$= 0.7169337.$

Example: If the Variance of a Poisson Variate is 3, then find the probability that (i) $x=0$ (ii) $0<x\leq 3$ (iii) $1\leq x<4$.

Solution: Given Variance of the poisson distribution, we have

mean = Variance

Hence mean of the Poisson Distribution = 3,

i.e $\lambda = 3$,

So $P(X=x) = P(x) = \frac{e^{-\lambda}.\lambda^x}{x!}$

$= \frac{e^{-3}.3^x}{x!}$

(i) $P(x=0) = \frac{e^{-3}.3^0}{0!} = e^{-3}$

$= 0.04979$

(ii) $P(0<x\leq 3) = P(x=1)+P(x=2)+P(x=3)$

$= e^{-3}.3 + e^{-3}.\frac{3^2}{2!} + e^{-3}.\frac{3^3}{3!}$

$= e^{-3}\left(3+\frac{9}{2}+\frac{9}{2}\right)$

$= e^{-3}(12) = 0.5974$

(iii) $P(1 \leq x < 4) = P(x=1) + P(x=2) + P(x=3)$

$= 0.5974$