unit. II

Moment's, Skewness, kurtosis

Moments:

definition:

The r^{th} moment about ~~the~~ any point A, denoted by μ_r' of a frequency distribution (f_i/x_i) is diffined by $\mu_r' = \frac{\sum f_i (x_i - A)^r}{N}$

When $A=0$, we get

$$\mu_r' = \frac{\sum f_i x_i^r}{N}$$

Which is the r^{th} moment about the origin

The r^{th} moment about the arithmetic mean $\bar{x}$ of a frequency distribution is given by

$$\mu_r = \frac{\Sigma f_i (x_i - \bar{x})^r}{N}$$

μ_r is also called the r^{th} central moment.

Note :-

The first moment about origin coinsides with the Arithmetic mean of the frequency distribution. and μ_2 is nothing but the varians of the frequency distribution.

Note-2

$$\mu_1' = \frac{\Sigma f_i (x_i - \bar{x})}{N}.$$

$$= \frac{\Sigma f_i x_i}{N} - \frac{\Sigma f_i \bar{x}}{N}$$

$$= \bar{x} - \frac{\cancel{N}\bar{x}}{\cancel{N}}$$

$\mu_1 = 0$

$$= \bar{x} - \bar{x}$$

$$= 0$$

Note-3:

$$\mu_1' = \frac{\Sigma f_i (x_i - A)}{N}$$

$$= \frac{\Sigma f_i x_i}{N} - \frac{A \Sigma f_i}{N}$$

$\mu_1 = \bar{x} - A$

$\bar{x} = \mu_1 + A$

$$= \bar{x} - \frac{A\cancel{N}}{\cancel{N}}$$

$$\mu_1' = \bar{x} - A$$

$$\boxed{\bar{x} = \mu_1' + A}$$

Relation between μ_r and μ_r' :

Theorem : 4.1

$$\mu_r = \mu_r' - rc_1 \mu_{r-1}' \cdot \mu_1' + rc_2 \mu_{r-2}' (\mu_1')^2 + \cdots \cdots + (-1)^{r-1} \cdot (r-1)(\mu_1')^r$$

We have,

$$\mu_r = \frac{\Sigma f_i (x_i - \bar{x})^r}{N}$$

$$= \frac{\Sigma f_i (x_i - A + A - \bar{x})^r}{N}$$

$$= \frac{\Sigma f_i \left[x_i - A - (\bar{x} - A)\right]^r}{N}$$

$$= \frac{\Sigma f_i \left[(x_i - A) - d\right]^r}{N}$$

where $d = \bar{x} - A$
$= \mu_1'$

$$= \frac{\Sigma f_i}{N} \Big[(x_i - A)^r + rc_1 (x_i - A)^{r-1} \cdot (-d) +$$

$$rc_2 (x_i - A)^{r-2} \cdot (-d)^2 + \cdots\cdots +$$

$$rc_{r-1} (x_i - A)(-d)^{r-1} + rc_r (-d)^r \Big]$$

$$= \frac{\sum f_i}{N}\Big[(x_i-A)^r - {}^rC_1 (x_i-A)^{r-1} d + {}^rC_2 (x_i-A)^{r-2}\cdot d^2 + \cdots + {}^rC_{r-1}(x_i-A)(-1)^{r-1}\cdot d^{r-1} + (-1)^r \cdot d^r\Big]$$

$$= \frac{\sum f_i (x_i-A)^r}{N} - {}^rC_1 \cdot d\, \frac{\sum f_i (x_i-A)^{r-1}}{N} + {}^rC_2 \cdot d^2 \frac{\sum f_i (x_i-A)^{r-2}}{N} + \cdots + {}^rC_{r-1}(-1)^{r-1} d^{r-1} \frac{\sum f_i (x_i-A)}{N} + (-1)^r d^r \frac{\sum f_i}{N}$$

μ_1 = central moment

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' \cdot (\mu_1')^2 + \cdots + \boxed{{}^rC_{r-1}} (-1)^{r-1} \cdot (\mu_1')^{r-1} \cdot \mu_1' + (-1)^r \cdot (\mu_1')^r$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' \cdot (\mu_1')^2 + \cdots + r(-1)^{r-1} \cdot (\mu_1')^{r-1} \cdot \mu_1' + (-1)^{r-1} \cdot (-1)(\mu_1')^r$$

$$\mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 + \cdots + r(-1)^{r-1} \cdot \mu_1'^r + (-1)^{r-1} (-1)(\mu_1')^r$$

$$= \mu_r' - {}^rC_1 \mu_{r-1}' \cdot \mu_1' + {}^rC_2 \mu_{r-2}' (\mu_1')^2 + \cdots + (-1)^{r-1} (\mu_1')^r (r-1)$$

Note:

Put $r = 1, 2, 3, 4$ we have

$$\mu_1 = \mu_1' - {}^1C_1 \cdot \mu_{1-1}' \cdot \mu_1'$$

$$= \mu_1' - \mu_0' \cdot \mu_1'$$

$$= \mu_1' - \mu_1'$$

$$= 0$$

$$\mu_2 = \mu_2' - {}^2C_1 \mu_{2-1}' \cdot \mu_1' + {}^2C_2 \mu_{2-2}' (\mu_1')^2$$

$$= \mu_2' - 2\mu_1' \cdot \mu_1' + \mu_0' \cdot (\mu_1')^2$$

$$= \mu_2' - 2(\mu_1')^2 + (\mu_1')^2$$

$$= \mu_2' - (\mu_1')^2$$

$$\mu_3 = \mu_3' - {}^3C_1 \mu_{3-1}' \cdot \mu_1' + {}^3C_2 \mu_{3-2}' \cdot (\mu_1')^2 - {}^3C_3 \mu_{3-3}' \cdot (\mu_1')^3$$

$$= \mu_3' - 3\mu_2' \cdot \mu_1' + 3\mu_1'(\mu_1')^2 - (\mu_1')^3$$

$$= \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1')^3$$

$$\mu_4 = \mu_4' - {}^4C_1 \mu_{4-1}' \cdot \mu_1' + {}^4C_2 \mu_{4-2}' (\mu_1')^2 - {}^4C_3 \mu_{4-3}' (\mu_1')^3 + {}^4C_4 \mu_{4-4}' (\mu_1')^4$$

$$= \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 4\mu_1' \cdot (\mu_1')^3 + (\mu_1')^4$$

$\cancel{\mu_4 = {}^4\mu_4' + 4\mu_3' \mu_1'}$

$\mu_4 = \mu_4' - 4\cancel{C_1 \mu_{4-1}' \cdot \mu_1'} + \cancel{4C_2 \mu_{4-2}' (\mu_1')^2}$

$= \mu_4' - \cancel{4\mu_3' \mu_1' + 6\mu_2' \mu_2'}$

$= \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 -$

$3(\mu_1')^4$

Theorem: 4.2

$$\mu_r' = \mu_r + rC_1 \mu_{r-1} (\mu_1') + rC_2 \mu_{r-2} (\mu_1')^2 + \dots + (\mu_1')^r$$

We have, $\mu_r' = \dfrac{\Sigma f_i (x_i - A)^r}{N}$

$$= \frac{\Sigma f_i (x_i - \bar{x} + \bar{x} - A)^r}{N}$$

$$= \frac{\Sigma f_i [(x_i - \bar{x}) + d]^r}{N}$$ where $d = \bar{x} - A = \mu_1'$

$$= \frac{\Sigma f_i}{N} \Big[(x_i - \bar{x})^r + rC_1 (x_i - \bar{x})^{r-1} d + rC_2 (x_i - \bar{x})^{r-2} d^2 + \dots + rC_{r-1} (x_i - \bar{x}) \cdot d^{r-1} + d^r \Big]$$

$$= \frac{\Sigma f_i}{N} \cdot \left[(x_i - \bar{x})^r + r c_1 (x_i - \bar{x})^{r-1} \cdot \mu_1' + r c_2 (x_i - \bar{x})^{r-2} (\mu_1')^2 \right.$$

$$\left. + \cdots + r c_{r-1} (x_i - \bar{x}) (\mu_1')^{r-1} + \mu_1'^r \right]$$

$$= \frac{\Sigma f_i (x_i - \bar{x})^r}{N} + r c_1 \frac{\Sigma f_i (x_i - \bar{x})^{r-1}}{N} \cdot \mu_1' + r c_2$$

$$\frac{\Sigma f_i (x_i - \bar{x})^{r-2}}{N} \cdot (\mu_1')^2 + \cdots + r c_{r-1} \underbrace{\frac{\Sigma f_i (x_i - \bar{x})}{N}}_{=0} (\mu_1')^{r-1}$$

$$+ \frac{\Sigma f_i}{N} (\mu_1')^r$$

$$\mu_r' = \mu_r + r c_1 \mu_{r-1} \cdot \mu_1' + r c_2 \mu_{r-2} (\mu_1')^2$$

$$+ \cdots + (\mu_1')^r$$

$$\mu_r = \frac{\Sigma f_i (x_i - \bar{x})^r}{N}$$

Put $r = 2, 3, 4$ we get

$$\mu_1 = \frac{\Sigma f_i (x_i - \bar{x})}{N}$$

$r = 2$

$$\mu_2' = \mu_2 + 2c_1 \mu_{2-1} (\mu_1') + 2c_2 \mu_{2-2} (\mu_1')^2$$

$$= \mu_2 + 2 \mu_1 \mu_1' + 1 \times \mu_0 (\mu_1')^2$$

$$= \mu_2 + 2 \overset{0}{\mu_1} \mu_1' + (\mu_1')^2$$

$$= \mu_2 + (\mu_1')^2$$

$r = 3$

$$\mu_3' = \mu_3 + 3C_1 \mu_{3-1} (\mu_1') + 3C_2 \mu_{3-2} (\mu_1')^2 +$$
$$3C_3 \mu_{3-3} (\mu_1')^3$$

$$= \mu_3 + 3 \mu_2 (\mu_1') + 3 \mu_1 (\mu_1')^2 +$$
$$1 \times \mu_0 (\mu_1')^3$$

$$= \mu_3 + 3 \mu_2 (\mu_1') + 3 \overset{0}{\mu_1} (\mu_1')^2 + (\mu_1')^3$$

$$= \mu_3 + 3 \mu_2 (\mu_1') + (\mu_1')^3$$

$r = 4$

$$\mu_4' = \mu_4 + 4C_1 \mu_{4-1} (\mu_1') + 4C_2 \mu_{4-2} (\mu_1')^2 +$$
$$4C_3 \mu_{4-3} (\mu_1')^3 + 4C_4 \mu_{4-4} (\mu_1')^4$$

$$= \mu_4 + 4 \mu_3 \mu_1' + 6 \mu_2 (\mu_1')^2 +$$
$$4 \overset{0}{\mu_1} (\mu_1')^3 + 1 \times \mu_0 (\mu_1')^4$$

$$= \mu_4 + 4 \mu_3 \mu_1' + 6 \mu_2 (\mu_1')^2 + (\mu_1')^4$$

Karl pearson's β and γ coefficients:

$\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$

$\gamma_1 = \sqrt{\beta_1}$ and $\gamma_2 = \beta - 3$

$\beta_1 = 0$ symme
$\beta_1 > 0$ positi
$\beta_1 < 0$ negat

If $\beta_1 = 0$ then the frequency distribution is symmetric.

If $\beta_1 > 0$ then the frequency distribution has positive skewness

If $\beta_1 < 0$ then the frequency distribution has negative skewness.

Mean - Mode and Mean - Median May be taken as measure of skewness

$\frac{\text{Mean} - \text{Mode}}{\sigma}$ and $\frac{3(\text{Mean} - \text{Median})}{\sigma}$

are called karl personi coefficient of

skewness. $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Kurtosis :-

Kurtosis is the degrees of peakedness of a distribution related to a Normal distribution

$\beta_2 = 3$ Meso
$\beta_2 < 3$ platy
$\beta_2 > 3$ lepto

For a normal curve,

If $\beta_2 = 3$ or $\gamma_2 = 0$ Then it is

Mesokurtic. $\beta_2 = 3$ mesokurtic
$\beta_2 < 3$ platykurtic

If $\beta_2 < 3$ or $\gamma_2 < 0$ then it is platy kurtic
$\beta_2 > 3$ leptokurtic

If $\beta_2 > 3$ or $\gamma_2 > 0$ then it is lepto kurtic

Problems:

1. Calculate the first four central moments for the following data to find β_1 and β_2 and discuss the nature of the ~~[illegible]~~ distribution:

Datas:

x	0	1	2	3	4	5	6
f	5	15	17	25	19	14	5

$$\bar{x} = \frac{\Sigma f_i x_i}{N \Rightarrow \Sigma f_i}$$

$$= \frac{0\times5 + 1\times15 + 2\times17 + 3\times25 + 4\times19 + 5\times14 + 6\times5}{5+15+17+25+19+14+5}$$

$$= \frac{0+15+34+75+76+70+30}{100}$$

$$= \frac{300}{100} = 3$$

x	f	$x-\bar{x}$	$f_i(x_i-\bar{x})$	$f_i(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^3$	$f_i(x_i-\bar{x})^4$
0	5	-3	-15	45	-135	405
1	15	-2	-30	60	-120	240
2	17	-1	-17	17	-17	17
3	25	0	0	0	0	0
4	19	1	19	19	19	19
5	14	2	28	56	112	224
6	5	3	15	45	135	405
	$\Sigma f_i = 100$		$\Sigma f_i(x_i-\bar{x})=0$	$\Sigma f_i(x_i-\bar{x})^2 = 242$	$\Sigma f_i(x_i-\bar{x})^3 = -6$	$\Sigma f_i(x_i-\bar{x})^4 = 1310$

$$\mu_r = \frac{\Sigma f_i (x_i-\bar{x})^r}{N}$$

$r = 1, 2, 3, 4$

$$\mu_1 = \frac{\Sigma f_i (x_i-\bar{x})}{N} = 0$$

$$\mu_2 = \frac{\Sigma f_i (x_i-\bar{x})^2}{N} = \frac{242}{100} = 2.42$$

$$\mu_3 = \frac{\Sigma f_i (x_i-\bar{x})^3}{N} = \frac{-6}{100} = -0.06$$

$$\mu_4 = \frac{\Sigma f_i (x_i-\bar{x})^4}{N} = \frac{1310}{100} = 13.1$$

~~Base β1~~ The first four ~~data~~ central moment are

$\mu_1 = 0$, $\mu_2 = 2.42$, $\mu_3 = -0.06$, $\mu_4 = 13.1$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-0.06)^2}{(2.42)^3}$$

$$= 0.0003$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{13.1}{(2.42)^2} = 2.237$$

Here $\beta_1 > 0$ then the distribution has possitive skewness.

$\beta_2 < 3$ then it is platykurtic.

2. Calculate the first four central moments for the following data to find β_1 and β_2 and discuss the nature of the distribution.

Data:

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

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

$$= \frac{0\times1 + 1\times8 + 2\times28 + 3\times56 + 4\times70 + 5\times56 + 6\times28 + 7\times8 + 8\times1}{1+8+28+56+70+56+28+8+1}$$

$$= \frac{0+8+56+168+280+280+168+56+8}{256}$$

$$= \frac{1024}{256} = 4$$

x	f	$x-\bar{x}$	$f_i(x_i-\bar{x})$	$f_i(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^3$	$f_i(x_i-\bar{x})^4$
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648

8	1	4	4	16	64	256

$\sum f_i = 256$ $\quad \sum f_i(x_i-\bar{x}) = 0$ $\quad \sum f_i(x_i-\bar{x})^2 = 512$ $\quad \sum f_i(x_i-\bar{x})^3 = 0$ $\quad \sum f_i(x_i-\bar{x})^4 = 2816$

$$\mu_r = \frac{\sum f_i (x_i-\bar{x})^r}{N}$$

$r = 1, 2, 3, 4$

$$\mu_1 = \frac{\sum f_i (x_i-\bar{x})}{N} = 0$$

$$\mu_2 = \frac{\sum f_i (x_i-\bar{x})^2}{N} = \frac{512}{256} = 2$$

$$\mu_3 = \frac{\sum f_i (x_i-\bar{x})^3}{N} = 0$$

$$\mu_4 = \frac{\sum f_i (x_i-\bar{x})^4}{N} = \frac{2816}{256} = 11$$

The first four central moment are

$\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 0$, $\mu_4 = 11$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{4} = 0$$

$$\beta_1 = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$$

Here $\beta_1 = 0$ then the frequency distribution is symmetric

$\beta_2 < 3$ then it is platykurtic.

3. Calculate the values of β_1 and β_2 for the distribution given in the following table.

Marks	0-9	10-19	20-29	30-39	40-49
Frequency	11	20	16	37	17

~~x f x-x fi(xi-x) fi(xi-x)^2~~

Mid value = 24.5 = A

marks x_i	f_i	$x_i - A$	$f_i(x_i-A)$	$f_i(x_i-A)^2$	$f_i(x_i-A)^3$	$f_i(x_i-A)^4$
4.5	11	-20	-220	4400	-88000	1760000
14.5	20	-10	-200	2000	-20000	200000
24.5	16	0	0	0	0	0
34.5	36	10	360	3600	36000	360000
44.5	17	20	340	6800	1,36000	2720000
	$\Sigma f_i = 100$		$\Sigma f_i(x_i-A) = 280$	$\Sigma f_i(x_i-A)^2 = 16800$	$\Sigma f_i(x_i-A)^3 = 64000$	$\Sigma f_i(x_i-A)^4 = 2392000$

$$\mu_3' = \frac{\Sigma f_i (x_i - A)^3}{N}$$

$$\mu_1' = \frac{\Sigma f_i (x_i - A)}{N}$$

$$= \frac{280}{100} = 2.8$$

$$\mu_2' = \frac{\Sigma f_i (x_i - A)^2}{N}$$

$$= \frac{16800}{100} = 168$$

$$\mu_3' = \frac{\Sigma f_i (x_i - A)^3}{N}$$

$$= \frac{64000}{100} = 640$$

$$\mu_4' = \frac{\Sigma f_i (x_i - A)^4}{N}$$

$$= \frac{25,92,000}{100}$$

$$= 25920$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= 168 - (2.8)^2$$

$$= 168 - 7.84$$

$$= 160.16$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1')^3$$

$$= 640 - 3 \times 168 \times 2.8 + 2 \times (2.8)^3$$

$$= -727.296$$

$$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 3(\mu_1')^4$$

$$= 25920 - 4 \times 640 \times 2.8 + 6 \times 168 \times (2.8)^2 - 3 \times (2.8)^4$$

$= 26470.3232$

The first four central moment are

$\mu_1 = 0$, $\mu_2 = 160.16$, $\mu_3 = -727.296$,

$\mu_4 = 26470.3232$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-727.296)^2}{(160.16)^3}$$

$$= 0.129 > 0$$

Here $\beta_1 > 0$ it has positive skewness

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{26470.3232}{(160.16)^2}$$

$$= 1.032 < 3$$

Here $\beta_2 < 3$ it is platykurtic.

A. The first four moments of the distribution about $x = 2$ are 1, 2.5, 5.5, and 16.

(i) Calculate the four moment about the mean

(ii) about zero

Given $A = 2$

$\mu_1' = 1$; $\mu_2' = 2.5$, $\mu_3' = 5.5$, $\mu_4' = 16$

(i) To find the moments about the mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2$

$= 2.5 - 1$

$= 1.5$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 5.5 - 3 \times 2.5 \times 1 + 2 \times 1$

$= 5.5 - 7.5 + 2$

$= 0$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= 16 - 4 \times 5.5 \times 1 + 6 \times 2.5 \times 1 - 3 \times 1$

$= 16 - 22 + 15 - 3 = 6$

(ii) To find the moments about zero

$$\mu_r' = \frac{\Sigma f_i x_i^r}{N}$$

$$\bar{x} = A + \mu_1'$$

$$= 2 + 1$$

$$= 3$$

$$\mu_1' = \frac{\Sigma f_i x_i}{N} = \bar{x}$$

$$= 3$$

$$\cancel{\mu_2' = \frac{\Sigma f_i x_i^2}{N}}$$

$$\mu_2' = \mu_2 + (\mu_1')^2$$

$$= 1.5 + 3^2$$

$$= 1.5 + 9$$

$$= 10.5$$

$$\mu_3' = \mu_3 + 3\mu_2 \mu_1' + (\mu_1')^3$$

$$= 0 + 3 \times 1.5 \times 3 + 3^3$$

$$= 40.5$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$$

$$= 6 + 4\times 0\times 3 + 6\times 1.5\times 9 + 3^4$$

$$= 6 + 6\times 1.5\times 9 + 81$$

$$= 168$$

5) The first four moments of the distribution about $x = 4$ are $-1.5, 17, -30, 108$ find the first four moment

(i) about mean

(ii) about the origin.

(iii) also calculate β_1 and β_2

Given $A = 4$

$\mu_1' = -1.5, \ \mu_2' = 17, \ \mu_3' = -30, \ \mu_4' = 108$

(i) To find the moments about the mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= 17 - (-1.5)^2$$

$$= 14.75$$

$$\mu_3 = \mu_3' - 3\mu_2' \cdot \mu_1' + 2(\mu_1')^3$$

$$= -30 - 3 \times 17 \times (-1.5) + 2(-1.5)^3$$

$$= -30 + 76.5 - 6.75$$

$$= 39.75$$

$$\mu_4 = \mu_4' - 4\mu_3' \cdot \mu_1' + 6\mu_2' \cdot (\mu_1')^2 - 3(\mu_1')^4$$

$$= 108 - 4 \times -30 \times -1.5 + 6 \times 17 \times (-1.5)^2 - 3 \times (-1.5)^4$$

$$= 142.3125$$

(ii) To find the moments about zero

$$\mu_r' = \frac{\sum f_i x_i^r}{N}$$

$$\bar{x} = A + \mu_1'$$

$$= 4 - 1.5$$

$$= 2.5$$

$$\mu_1' = \frac{\sum f_i x_i}{N} = \bar{x}$$

$$= 2.5$$

$$\mu_2' = \mu_2 + (\mu_1')^2$$

$$= 14.75 + (2.5)^2$$

$= 21$

$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$

$= 39.75 + 3 \times 14.75 \times 2.5 + (2.5)^3$

$= 166$

$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4$

$= 142.3125 + 4 \times 39.75 \times 2.5 +$

$6 \times 14.75 \times (2.5)^2 + (2.5)^3$

$= 1132$

(iii) $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(39.75)^2}{(14.75)^3}$

$= 0.49237 > 0$

Here $\beta_1 > 0$ the the frequency distribution has positive skewness.

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{142.3125}{(14.75)^2}$

$= 0.654122 < 3$

$\beta_2 < 3$ then it is platy kurtic

6) The first ~~four~~ three moments of the distribution about $x=3$ are 2, 10, and 30

(i) calculate the three moments about mean

(ii) about zero.

Let $A = 3$

$\mu_1' = 2$, $\mu_2' = 10$, $\mu_3' = 30$

(i) To find three moments about the mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2$

$= 10 - (4)$

$= 6$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 30 - 3\times 10\times 2 + 2\times 8$

$= 30 - 60 + 16$

(ii)

7) ...

origi...

$\mu_2' =$

the

$= -14$

(ii) about the origin

$$\mu_1' = \frac{\sum f_i x_i}{N}$$

$\bar{x} = A + \mu_1'$

$= 3 + 2$

$= 5$

$= 5$

$$\mu_2' = \mu_2 + \mu_1'^2$$

$= 6 + 25$

$= 31$

(5)

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$= -14 + 3 \times 6 \times 5 + 125$

$= -14 + 90 + 125$

$= 201$

7) The first three moment about the origin are given by $\mu_1' = \frac{1}{2}(n+1)$, $\mu_2' = \frac{1}{6}(n+1)(2n+1)$, $\mu_3' = \frac{1}{4}n(n+1)^2$ Examine the skewness of the distribution.

$\mu_1' = \frac{1}{2}(n+1)$, $\mu_2' = \frac{1}{6}(n+1)(2n+1)$

$\mu_3' = \frac{1}{4} n(n+1)^2$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= \frac{1}{6}(n+1)(2n+1) - \left[\frac{1}{2}(n+1)\right]^2$$

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

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

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

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

$$= \frac{1}{12}(n+1)\left[n-1\right]$$

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

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= \frac{1}{4} n(n+1)^2 - 3 \times \frac{1}{6}(n+1)(2n+1)\frac{1}{2}(n+1) + 2\left[\frac{1}{2}(n+1)\right]^3$$

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

$$= \frac{1}{4}(n+1)^2 \left[n - (2n-1) + (n+1) \right]$$

$$= \frac{1}{4}(n+1)^2 \left[n - 2n - 1 + n + 1 \right]$$

$$= \frac{1}{4}(n+1)^2 (0)$$

$$= 0$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{\frac{1}{12}(n^2-1)} = 0$$

Here $\beta_1 = 0$ then the distribution is symmetric

8) For a frequency distribution $\frac{f_i}{x_i}$ show that $\beta_2 \geq 1$

T.P $\beta_2 \geq 1$

(ie) T.P $\frac{\mu_4}{\mu_2^2} \geq 1$

$\mu_4 \geq \mu_2^2$

$\mu_4 - \mu_2^2 \geq 0$

now, $\mu_4 - \mu_2^2$

$$\frac{\sum f_i(x_i-\bar{x})^4}{N} - \left[\frac{\sum f_i(x_i-\bar{x})^2}{N}\right]^2$$

$$\frac{\sum f_i\left[(x_i-\bar{x})^2\right]^2}{N} - \left[\frac{\sum f_i(x_i-\bar{x})^2}{N}\right]^2$$

$$= \frac{\sum f_i z_i^2}{N} - \left[\frac{\sum f_i z_i}{N}\right]^2$$

where $(x_i-\bar{x})^2 = z_i$

$$= \sigma_{z_i}^2 \qquad \sigma_x = \sqrt{\frac{\sum f x^2}{N} - \left(\frac{\sum f x}{N}\right)^2}$$

≥ 0

$\therefore \mu_4 - \mu_2^2 \geq 0$

Hence $\beta_2 \geq 1$

9) Calculate the first four moments about ~~the~~ ~~po~~ $x = 4$ and Hence find the moments ~~of~~ about the main of the following distribution also find

x	0	1	2	3	4	5	6	7	8	9	10
f	5	10	30	70	140	200	140	70	30	10	5

10) The first four moments of a distribution about $x = 4$ are 1, 4, 10, 45 respectively calculate the moments about the mean

$A = 4$

$\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 45$

(i) To find the four moments about the mean:

$\mu_1 = 0$

$\mu_2 = \mu_2' - {\mu_1'}^2$

$= 4 - (1)^2$

$= 4 - 1$

$= 3$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

$$= 10 - 3\times4\times1 + 2\times1^3$$

$$= 10 - 12+2$$

$$= 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\cdot\mu_1^2 - 3(\mu_1')^4$$

$$= 45 - 4\times10\times1 + 6\times4\times1 - 3\times1$$

$$= 45-40+24-3$$

$$= 26$$

$A = 4$

9)

x	f	$x-A$	$\Sigma f(x_i-A)$	$\Sigma f(x_i-A)^2$	$\Sigma f(x_i-A)^3$	$\Sigma f(x_i-A)^4$
0	5	-4	-20	80	-320	1280
1	10	-3	-30	90	-270	810
2	20	-2	-60	120	-240	480
3	70	-1	-70	70	-70	70
4	140	0	0	0	0	0
5	200	1	200	200	200	200
6	140	2	280	560	1120	2240
7	70	3	210	630	1890	5670
8	30	4	120	480	1920	7680
[illegible]	10	5	50	250	1250	6250
[illegible]	5	6	30	180	1080	6480

N = 7

$\mu_1' =$

$\mu_1' =$

$\boxed{\mu_1' =}$

$\mu_2' =$

H

$N = 710$

$\mu_1' = \frac{\sum f_i (x_i - A)^r}{N}$

$\mu_1' = \frac{710}{710}$

$\boxed{\mu_1' = 1}$

$\mu_2' = \frac{2660}{710} = 3.75$

$\boxed{\mu_2' = 3.75}$.

$\mu_3' = \frac{6560}{710}$

$\boxed{\mu_3' = 9.24}$.

$\mu_4' = \frac{31160}{710}$

$\boxed{\mu_4' = 43.89}$.

$\mu_r' = \frac{\sum f_i (x_i - A)^r}{N}$

$\mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N}$

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

$N = \sum f_i$

$\beta_1 = 0$ symmetric

$\beta_1 > 0$ positive

$\beta_1 < 0$ Negative

$\beta_2 = 0$ meso

$\beta_2 < 0$ platy

$\beta_2 > 0$ lepto

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2$

$= 3.75 - (1)^2$

$\boxed{\mu_2 = 2.75}$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

$= 9.24 - 3(3.75)(1) + 2(1)^3$

$= 9.25 - 11.25 + 2$

$\boxed{\mu_3 = -0.01}$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

$= 43.89 - 4(9.24)(1) + 6(3.75)(1) - 3(1)$

$= 43.89 - 36.96 + 22.5 - 3$

$\boxed{\mu_4 = 26.43}$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$= \frac{(-0.01)^2}{(2.75)^3}$$

β_1

β_2

$$= \frac{0.0001}{20.796875}$$

$= 0.00004 > 0$ It has positive skewness.

β_1

β_2

β_2

β

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{26.43}{(2.75)^2}$$

$$= \frac{26.43}{20.7965}$$

$= 3.49 > 3$ lepto kurtic

curve fitting

Let x_i $i = 1, 2, \ldots, n$ be the values of independent variable and y_i $i = 1, 2, \ldots n$ be the corresponding values of dependent variable.

If the points (x_i, y_i) $i = 1, 2, \ldots n$ are plotted on a graph paper and be cupticed a diagram called scatter diagram.

If there is a functional realationship between x_i and y_i the points of the scatter diagram will be found to be consentrated round a sertern curve.

The process of finding such the functional realationship between the variables is called curve fitting example :

The lines of regression can be get by fitting a linear curve to a given bi variate distribution.

Principle of least squares:

Let (x_i, y_i) $i = 1, 2 \ldots n$ be the observed set of values of the variable (x, y) Let $y = f(x)$ be a functional relationship between the variables (x, y) Then $d_i = y_i - f(x_i)$ which is the difference between the observed values of y and the value of y determine by the functional realation is called the residuals. The principle of least squares states that the parameters involved in $f(x)$ should be chosen in such a way that

Σd_i^2 is minimum

Fitting a straight line:

Consider the fitting of a straight line $y = ax + b$ to the values (x_i, y_i) where $i = 1, 2, \ldots n$ the residual d_i is given by

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

$$d_i^2 = [y_i - (ax_i + b)]^2$$

$$d_i^2 = [y_i - ax_i - b]^2$$

$$\Sigma d_i^2 = \Sigma [y_i - ax_i - b]^2 = R \text{ (say)}$$

according to the priciple of least square we have to determine the parameters a, b, so that R is minimum.

$$\frac{\partial R}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a}\left[\Sigma(y_i - ax_i - b)\right]^2 = 0$$

$$\Rightarrow 2\Sigma(y_i - ax_i - b)\cdot(-x_i) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b)(x_i) = 0$$

$$\Rightarrow \Sigma(y_i x_i - ax_i^2 - bx_i) = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma x_i^2 - b\Sigma x_i = 0$$

$$\Rightarrow \Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \rightarrow (1)$$

$$\frac{\partial R}{\partial b} = 0 \Rightarrow \frac{\partial}{\partial b}\left[\Sigma(y_i - ax_i - b)\right]^2 = 0$$

$$\Rightarrow 2\Sigma(y_i - ax_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - \Sigma b = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - nb = 0$$

$$\Rightarrow \Sigma y_i = a\Sigma x_i + nb \rightarrow (2)$$

equation ① and ② are called normal equation. from these equations ① we have find a and b.

Fitting a second degree parabola:-

consider the fitting of the second degree parabola $y = ax^2 + bx + c$ to the values (x_i, y_i) $y_i = ax_i^2 + bx_i + c$ $i = 1, \ldots, n$. The residual d_i given by ~~$d_i = y_i - ax_i^2$~~

$$d_i = y_i - (ax_i^2 + bx_i + c)$$

$$d_i^2 = (y_i - ax_i^2 - bx_i - c)^2$$

$$\Sigma d_i^2 = \Sigma (y_i - ax_i^2 - bx_i - c)^2$$

According to the principle of least square we have to determine the parameters a, b, c so that R is minimum

$$\frac{\partial R}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a} \left[\Sigma (y_i - ax_i^2 - bx_i - c)^2 \right] = 0$$

$$\Rightarrow 2 \Sigma (y_i - ax_i^2 - bx_i - c)(-x_i^2) = 0$$

$$\Rightarrow -2 \Sigma (y_i - ax_i^2 - bx_i - c)(x_i^2) = 0$$

$\Rightarrow \sum (y_i - a x_i^2 - b x_i - c)(x_i^2) = 0$

$\Rightarrow \sum x_i^2 y_i - a \sum x_i^4 - b \sum x_i^3 - c \sum x_i^2 = 0$

$\Rightarrow \sum x_i^2 y_i = a \sum x_i^4 + b \sum x_i^3 + c \sum x_i^2 \rightarrow ②$

$\frac{\partial R}{\partial b} = 0 \Rightarrow \frac{\partial}{\partial b} \left[\sum (y_i - a x_i^2 - b x_i - c) \right]^2 = 0$

$\Rightarrow 2 \sum (y_i - a x_i^2 - b x_i - c)(-x_i) = 0$

$\Rightarrow -2 \sum (y_i - a x_i^2 - b x_i - c)(x_i) = 0$

$\Rightarrow \sum (y_i - a x_i^2 - b x_i - c)(x_i) = 0$

$\Rightarrow \sum x_i y_i - a \sum x_i^3 - b \sum x_i^2 - c \sum x_i = 0$

$\Rightarrow \sum x_i y_i = a \sum x_i^3 + b \sum x_i^2 + c \sum x_i = 0 \rightarrow ③$

$\frac{\partial R}{\partial c} = 0 \Rightarrow \frac{\partial}{\partial c} \left[\sum (y_i - a x_i^2 - b x_i - c) \right]^2 = 0 \rightarrow ④$

$\Rightarrow 2 \sum (y_i - a x_i^2 - b x_i - c)(-1) = 0$

$\Rightarrow \sum (y_i - a x_i^2 - b x_i - c) = 0$

$\Rightarrow \sum y_i - a \sum x_i^2 - b \sum x_i - c \sum = 0$

$\Rightarrow \sum y_i - a \sum x_i^2 - b \sum x_i - nc = 0$

$\Rightarrow \sum y_i - a \sum x_i^2 - b \sum x_i$

$\Rightarrow \sum y_i = a\sum x_i^2 + b\sum x_i + nc \rightarrow ③$

equation ①, ②, ③ called normal equation from these equation from a, b, c.

1) Fit a straight line to the following data.

x	0	1	2	3	4
y	2.1	3.5	5.4	7.3	8.2

Soln:

Let us fit a straight line to the given data.

$$y = ax + b \rightarrow ①$$

we have to determine parameters a, b by using normal equations.

$$\sum x_i y_i = a\sum x_i^2 + b\sum x_i \rightarrow ②$$

$$\sum y_i = a\sum x_i + nb \rightarrow ③$$

x_i	y_i	$x_i y_i$	x_i^2
0	2.1	0	0
1	3.5	3.5	1
2	5.4	10.8	4
3	7.3	21.9	9
4	8.2	32.8	16
$\Sigma x_i = 10$	$\Sigma y_i = 26.5$	$\Sigma x_i y_i = 69$	$\Sigma x_i^2 = 30$

Sub these values in (1) & (2)

$$30a + 10b = 69 \longrightarrow (4)$$

$$10a + 5b = 26.5 \longrightarrow (5)$$

$$(4) \Rightarrow 30a + 10b = 69$$

$$(5) \times 3 \Rightarrow 30a + 15b = 79.5 \longrightarrow (6)$$

$$(4) - (6) \Rightarrow \quad -5b = -10.5$$

$$b = \frac{+10.5}{+5} = 2.1$$

$$b = +2.1$$

$$b = 2.1$$

$b = 2.1$ sub (4)

$$30a + 10 \times 2.1 = 69$$

$$30a + 21 = 69$$

$$30a = 69 - 21$$

$$30a = 48$$

$$a = \frac{\overset{16}{\cancel{48}}}{\underset{10}{\cancel{30}}}$$

$$a = 1.6$$

The straight line fitted for the given data is $y = 1.6x + 2.1$

7) Fit a straight line $y = a + bx$ to the following data.

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

Let us fit a straight line to the given data

$$y = bx + a \rightarrow (1)$$

We have to determine parameters a, b by using normal equations.

$$\Sigma x_i y_i = a\Sigma x_i^2 + b\Sigma x_i \rightarrow (2)$$

$$\Sigma y_i = a\Sigma x_i + nb \rightarrow (3)$$

x_i	y_i	$x_i y_i$	x_i^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
$\Sigma x_i = 10$	$\Sigma y_i = 16.9$	$\Sigma x_i y_i = 47.1$	$\Sigma x_i^2 = 30$

$$30a + 10b = 47.1 \rightarrow (4)$$

$$10a + 5b = 16.9 \rightarrow (5)$$

$(4) \Rightarrow 30a + 10b = 47.1$

$(5)\times 2 \Rightarrow 20a + 10b = 33.8$

$$10a = 13.3$$

$$a = \frac{13.3}{10}$$

$$\boxed{a = 1.33}$$

$a = 1.33$ sub (4)

$$30 \times 1.33 + 10b = 47.1$$

$$39.9 + 10b = 47.1$$

$$10b = 47.1 - 39.9$$

$$10b = 7.2$$

$$b = 7.2/10$$

$$b = 0.72$$

The straight line fitted for the given data is

$$y = 0.72x + 1.33$$

3) fit a straight line to the following data and estimat the value of y corresponding to $x=6$

x	0	5	10	15	20	25
y	12	15	17	22	24	30

Let us fit a straight line to the given data is

$$y = ax + b \longrightarrow ①$$

we have to determine the parameters a and b by using the normal equations

$$\Sigma x_i y_i = a \Sigma x_i^2 + b \Sigma x_i \longrightarrow ②$$

$$\Sigma y_i = a \Sigma x_i + nb \longrightarrow ③$$

x_i	y_i	$x_i y_i$	x_i^2
0	12	0	0
5	15	75	25
10	17	170	100
15	22	330	225

20	24	480	400
25	30	750	625
$\Sigma x_i = 75$	$\Sigma y_i = 120$	$\Sigma x_i y_i = 1805$	Σx_i^2 1375

Sub the values ① & ②

$1375a + 75b = 1805 \rightarrow (4)$

$75a + 6b = 120 \rightarrow (5)$

$(4) \times 6 \Rightarrow 8250a + 450b = 10830 \rightarrow (6)$

$(5) \times 75 \Rightarrow 5625a + 450b = 9000 \rightarrow (8)$

$2625a = 1830$

$a = \frac{1830}{2625}$

$a = 0.6971$

Sub $a = 0.6971$ (5)

$75 \times 0.6971 + 6b = 120$

$52.2825 + 6b = 120$

$6b = 120 - 52.2825$

$= 67.7175$

$$b = \frac{67.7175}{6}$$

$$b = 11.2865$$

$$y = 0.6471x + 11.2865$$

when $x = 6$

$$y = 15.4691$$

4) fit a second degree parabola by taking x_i as a indipendent variable

x	0	1	2	3	4
y	1	5	10	22	38

Soln:

Let the second degree parabola to be fitted to the given data is

$$y = ax^2 + bx + c \longrightarrow (1)$$

we have to determine the parameters a, b, c by using the normal equations.

$$\sum x_i^2 y_i = a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 \to (3)$$

$$\sum x_i y_i = a\sum x_i^3 + b\sum x_i^2 + c\sum x_i \to (4)$$

$$\sum y_i = a\sum x_i^2 + b\sum x_i + nc \to (5)$$

x_i	y_i	$x_i^2 y_i$	$x_i y_i$	x_i^2	x_i^3	x_i^4
0	1	0	0	0	0	0
1	5	5	5	1	1	1
2	10	40	20	4	8	16
3	22	198	66	9	27	81
4	38	608	152	16	64	256
$\sum x_i = 10$	$\sum y_i = 76$	$\sum x_i^2 y_i = 851$	$\sum x_i y_i = 243$	$\sum x_i^2 = 30$	$\sum x_i^3 = 100$	$\sum x_i^4 = 354$

sub the values (3), (4), and (5)

$$354a + 100b + 30c = 851 \to (6)$$

$$100a + 30b + 10c = 243 \to (7)$$

$$30a + 10b + 5c = 76 \to (8)$$

$$(7) \Rightarrow 100a + 30b + 10c = 243 \to (7)$$

$$(8)\times 2 \Rightarrow 60a + 20b + 10c = 152 \to (8)$$

$$(-) \quad (-) \quad (-) \quad (-)$$

$$40a + 10b = 91 \to (9)$$

$354a + 100b + 80c = 851 \rightarrow (6)$

$(6)\times 3 \Rightarrow 300a + 90b + 80c = 729 \rightarrow (7)$

$(-) \quad (-) \quad (-) \quad (-)$

$54a + 10b = 122 \rightarrow (10)$

$40a + 10b = 91 \rightarrow (9)$

$54a + 10b = 122 \rightarrow (10)$

$(-) \quad (-) \quad (-)$

$-14a = -31$

$a = \frac{-31}{-14}$

$\cancel{a = 2.071}$ $\boxed{a = 2.214}$

$a = 2.071$ sub (9)

$40 \times 2.214 + 10 \times b = 91$

$88.56 + 10b = 91$

$10b = 91 - 88.56$

$10b = 2.44$ $\cancel{10b = 10.16}$

$b = \frac{2.44}{10}$ $\cancel{b = \frac{10.16}{10}}$

$\boxed{b = 0.244}$ $\cancel{\boxed{b = 1.016}}$

a, b value sub in (8)

$\cancel{30 \times 2.214 + 10 \times 1.016 + 5c = 75}$

$\cancel{72.29 + 5c = 75}$

$\cancel{5c = 75 - 72.29}$

$\cancel{5c = 2.71}$

The

data

3

the

given

$$c = \frac{2.71}{5}$$

$$c = 0.542$$

The straight line fitted for the given data is

$$y = 2.071x^2 + 1.016x + 0.542$$

a, b value sub in (8)

$$30 \times 2.214 + 10 \times 0.244 + 5c = 76$$

$$66.42 + 2.44 + 5c = 76$$

$$68.86 + 5c = 76$$

$$5c = 76 - 68.86$$

$$5c = 7.14$$

$$c = \frac{7.14}{5}$$

$$\boxed{c = 1.428}$$

the straight line fitted for the given data is

$$y = 2.214x^2 + 0.244x + 1.428$$

7) Fit a straight line to the following data regarding x as the independing variable.

years x	1911	1921	1931	1941	1951
production y	10	12	8	10	134

Soln:

Let the straight line fitted to the given data

$\boxed{u = \frac{x-A}{c}}$

$$y = ax + b$$

Put $u = \frac{x - 1931}{10}$, $v = y - 10$

The straight line fitted to the given data is $v = au + b \rightarrow (1)$

we have to determine the parameters a and b by using normal equations

$$\Sigma u_i v_i = a \Sigma u_i^2 + b \Sigma u_i \rightarrow (2)$$

$$\Sigma v_i = a \Sigma u_i + nb \rightarrow (3)$$

x_i	y_i	$u_i = \frac{x_i - 1931}{10}$	$v_i = y_i - 10$	u_i^2	$u_i v_i$
1911	10	-2	0	4	0
1921	12	-1	2	1	-2
1931	8	0	-2	0	0
1941	10	1	0	1	0
1951	14	2	4	4	8
		$\Sigma u_i = 0$	$\Sigma v_i = 4$	$\Sigma u_i^2 = 10$	$\Sigma u_i v_i = 6$

sub these values in ② & ③

$$10a + b(0) = 6$$

$$a = \frac{6}{10} = 0.6$$

$$a(0) + 5(b) = 4$$

$$5b = 4$$

$$b = 4/5 = 0.8$$

① ⇒ $v = 0.6u + 0.8$

$$y - 10 = 0.6\left(\frac{x - 1931}{10}\right) + 0.8$$

$$y - 10 = 0.06\,(x - 1931) + 0.8$$

$y - 10 = 0.06x - 115.86 + 0.8$

$y = 0.06x - 115.86 + 0.8 + 10$

$y = 0.06x - 105.06$

2) Fit the curve $\underline{y = bx^a}$ to the following data

x	1	2	3	4	5	6
y	1200	900	600	200	110	50

The given curve is $y = bx^a$

Taking log

$\log y = \log bx^a$

$\log y = \log b + \log x^a$

$\underline{\log y} = \underline{\log b} + \underline{a \log x}$

$\log y = a \log x + \log b$

If is of the form $Y = AX + B \rightarrow ①$

where $Y = \log y$, $A = a$, $X = \log x$,

$B = \log b$

we have to determine parameters A and B by using normal equations

$$\sum X_i Y_i = A \sum X_i^2 + B \sum X_i \rightarrow ②$$

$$\sum Y_i = A \sum X_i + nB \rightarrow ③$$

x_i	y_i	$X_i = \log x$	$Y_i = \log y$	X_i^2	$X_i Y_i$
1	1200	0	3.0791	0	0
2	900	0.3010	2.9542	0.0906	0.8894
3	600	0.4771	2.7781	0.2276	1.3254
4	200	0.6020	2.3010	0.3624	1.3852
5	110	0.6989	2.0413	0.4884	1.4266
6	50	0.7781	1.6989	0.6054	1.3219
		2.8571	14.8526	1.7744	6.3485
		2.86	14.85	1.77	6.35

$$1.77A + 2.86B = 6.35 \rightarrow (4)$$

$$2.86A + 6B = 14.85 \rightarrow (5)$$

$$(4) \times 6 \Rightarrow \quad 10.62A + \cancel{17.16B} = 38.10$$

$$(5) \times 2.86 \Rightarrow \quad \underset{(-)}{8.18A} + \underset{(-)}{\cancel{17.16B}} = \underset{(-)}{42.47}$$

$$2.44A \qquad = -4.4$$

$$A = \frac{-4.4}{2.44}$$

$$= -1.8032$$

$$= -1.80$$

Sub $A = -1.80$ in (5)

$$2.86(-1.80) + 6B = 14.85$$

$$6B = 14.85 + 5.148$$

$$6B = 19.998$$

$$B = \frac{19.998}{6}$$

$$B = 3.33$$

$$Y = AX + B$$

$$Y = -1.80X + 3.33$$

$A = a = -1.80$

$B = \log b = 3.333$

$b = \text{anti}\log(3.333)$

$= 2152.48$

The given curve is

$y = bx^a$

$y = (2152.48)x^{-1.80}$ // Ans.

3) Explain the method of fitting the curve good fit $y = ae^{bx}$ $(a>0)$

Taking log

$$\log y = \log ae^{bx}$$

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

$$\log y = \log a + bx \log e$$

$$\log y = (b \log e) x + \log a$$

It is of the form $Y = Ax + B \rightarrow (1)$

where $Y = \log y$, $A = b \log e$

$X = x$, $B = \log a$

We have to determine the parameters A and B of using normal equation.

$\Sigma x_i y_i = A\Sigma x_i^2 + B\Sigma x_i \rightarrow ②$

$\Sigma y_i = A\Sigma x_i + nB \rightarrow ③$

From the two normal equations we get the values of $\boxed{A \& B}$ and a & b can be obtained from $\boxed{B = \log a}$

$a = \text{antilog} B$

$A = b \log e$

$b = \frac{A}{\log e}$

4) Explain the method of fitting the curve $y = ka^{bx}$

Taking log,

$\log y = \log ka^{bx}$

$\log y = \log k + \log a^{bx}$

$\log y = \log k + bx \log a$

$$\log y = bx \log a + \log k$$

$$\log y = (b \log a)x + \log k$$

It is from $Y = Ax + B \rightarrow (1)$

where $Y = \log y$, $A = b\log a$,

$B = \log k$

we have to determine the parameters A and B by using normal equations

$$\sum x_i Y_i = A \sum x_i^2 + B \sum x_i \rightarrow (2)$$

$$\sum Y_i = A \sum x_i + nB \rightarrow (3)$$

From the two normal equations we get the values of A & B and a & b can be obtained from

$B = \log k$

$A = b \log a$

$b = \frac{A}{\log a}$

$b \log a = A$

$\log a = \frac{A}{b}$

$a = \text{anti} \log (A/b)$

1) Fit a curve of a form $y = ab^x$ the following data

Years (x)	1951	1952	1953	1954	1955	1956	1957
Production in tons (y)	201	263	314	395	427	504	612

Soln:

The given curve is $y = ab^x$

Taking log,

$\log y = \log a + \log b^x$

$\log y = \log a + x \log b$

$\log y = x \log b + \log a$

This is of the form $Y = AX + B \rightarrow (1)$

where $Y = \log y$, $A = \log b$, $B = \log a$

Put $X = x - 1954$

We have to determine the parameters A and B by using normal equations.

$\sum x_i y_i = A \sum x_i^2 + B \sum x_i \rightarrow (2)$

$\sum y_i = A \sum x_i + nB \rightarrow (3)$

x	y	$x_i = x - 1954$	$Y_i = \log y$	x_i^2	$x_i Y_i$
1951	201	-3	2.303	9	-6.909
1952	263	-2	2.419	4	-4.838
1953	314	-1	2.496	1	-2.496
1954	395	0	2.596	0	0
1955	427	1	2.63	1	2.63
1956	504	2	2.702	4	5.404
1956	612	3	2.786	9	8.358
		$\sum x_i = 0$	$\sum y_i = 17.937$	$\sum x_i^2 = 28$	$\sum x_i y_i = 2.149$

Sub these Values in ③ & ④

$28A + 0 = 2.149$

$0 + 7B = 17.932$

$B = \frac{17.932}{7}$

$\boxed{B = 2.561}$

$28A = 2.149$

$A = \frac{2.149}{28}$

$A = 0.076$

$A = \log b$

$\log b = A$

$b = \text{antilog}(A)$

$= \text{antilog}(0.076)$

$b = 1.191$

$B = \log a$

$a = \text{antilog}(B)$

$= \text{antilog}(2.561)$

$= 363.9$

The required curve

$$y = a b^{x}$$

$$y = (363.9)(1.191)^{x-1954}$$

2) Fit the exponential curve $y = a e^{bx}$ to the following data

x	0	2	4
y	5.02	10	31.62

The given curve is $y = a e^{bx}$

Taking log,

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

$$= \log a + bx \log e$$

$$= \log a + (b \log e) x$$

$$\log y = (b \log e) x + \log a$$

This is of the form $Y = Ax + B \rightarrow ①$

where $Y = \log y$, $A = b \log e$,

$B = \log a$

But so we have to determine the parameters A and B by using normal equations.

$$\Sigma x_i Y_i = A\Sigma x_i^2 + B\Sigma x_i \rightarrow ⓐ$$

$$\Sigma Y_i = A\Sigma x_i + nB \rightarrow ⓑ$$

x	y	$Y_i = \log y$	x_i^2	$x_i Y_i$
0	5.02	0.7	0	0
2	10	1	4	2
4	31.62	1.49	16	5.96
6		$\Sigma Y_i = 3.19$	$\Sigma x_i^2 = 20$	$\Sigma x_i Y_i = 7.96$

sub these values in ⓐ & ⓑ

$$7.96 = 20A + 6B \rightarrow ⓐ$$

$$3.19 = 6A + 3B \rightarrow ⓑ$$

$$ⓐ \Rightarrow 20A + 6B = 7.96$$

$$ⓑ \times 2 \Rightarrow 12A + 6B = 6.38$$

$$8A = 1.58$$

$$A = \underline{0.1975}$$

$$A = 0.20$$

Sub in ⓑ

$$3.19 = 0.20 \times 6 + 3B$$

$$3.19 = 1.2 + 3B$$

$$3B = 1.99$$

$$B = 0.66$$

$$a = \text{antilog}(B)$$

$$= \text{antilog}(0.66)$$

$$= 4.57$$

$$b = \frac{A}{\log e}$$

$$= \frac{0.20}{0.43}$$

$$= 0.465$$

$$y = 4.57\, e^{0.46x}$$

Unit - II

correlation

consider a set of ~~by~~ bivariate data x_i, y_i $i = 1, 2 \dots n$ if their is a change in one variable corresponding to change a other variable we say that the variable that correated.

If the two variable deviate in the same direction the correlation is said to be direct or positive. If they always deviate in the opposite direction the correlation is set to be inverse or negative. If the change in one variable corresponce to the proposal to the other variable then the correlation is perfect.

correl

betwe

difined

co - V

1) The

Student

Height

Weight

~~ ~~ Karl pearson's coefficiant of correlation:-

Karl pearson's coefficiant of correlation between the variable x and y is difined by $r_{(x,y)} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x \sigma_y}$

where $\bar{x} = \frac{\Sigma x_i}{n}$, $\bar{y} = \frac{\Sigma y_i}{n}$

$\sigma_x = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}}$ & $\sigma_y = \sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n}}$

co-variance between x & y is difined by

$$\text{covariance}_{(x,y)} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n}$$

Hence $r_{(x,y)} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$

1) The heights and weights ~~are~~ of 5 students are given below.

Height in cm (x)	160	161	162	163	164
Weight in kg (y)	50	53	54	56	57

find the correlation between x & y

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$= \frac{160 + 161 + 162 + 163 + 164}{5}$$

$$= 162$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{50 + 53 + 54 + 56 + 57}{5}$$

$$= 54$$

x	y	$x - \bar{x}$ $x - 162$	$y - \bar{y}$ $y - 54$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
160	50	-2	-4	4	16	8
161	53	-1	-1	1	1	1
162	54	0	0	0	0	0
163	56	1	2	4	4	2
164	57	2	3	1	9	6
		$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(y_i - \bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 10$	$\Sigma(y-\bar{y})^2 = 30$	$\Sigma(x-\bar{x})(y-\bar{y}) = 17$

$$\sigma_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

$$= \frac{10}{5} = 2$$

$\sigma x = \sqrt{2}$

$$\sigma y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n}$$

$$= \frac{30}{5} = 6$$

$\sigma y = \sqrt{6}$

correlation between x & y is

$$r_{(x,y)} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n\,\sigma x\,\sigma y}$$

$$\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\,\sigma x\,\sigma y}$$

$$= \frac{17}{5\times\sqrt{2}\times\sqrt{6}} = \frac{17}{5\times 2\sqrt{3}}$$

$$\simeq \frac{17}{5\times 3.464} = \frac{17}{17.320} = 0.98$$

Theorem (1)

1) Proove that $r_{(x,y)} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[n\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2}\,[n\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}}$

$$r_{xy} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma x\,\sigma y}$$

Proof:

we have $r_{(x,y)} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n\,\sigma x\,\sigma y} \to (1)$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = \Sigma[x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}]$$

$$= \Sigma x_i y_i - \Sigma x_i\bar{y} - \Sigma\bar{x}y_i + \Sigma\bar{x}\bar{y}$$

$$= \Sigma x_i y_i$$

$$= \sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}$$

$$= \sum x_i y_i - \bar{y} n\bar{x} - \bar{x} n\bar{y} + n\bar{x}\bar{y}$$

$$= \sum x_i y_i - n\bar{x}\bar{y} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}$$

$$\boxed{\bar{x} = \frac{\sum x_i}{n},\ \bar{y} = \frac{\sum y_i}{n}}$$

$$= \sum x_i y_i - \cancel{n}\frac{\sum x_i}{\cancel{n}} \cdot \frac{\sum y_i}{n}$$

$$= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n} \longrightarrow ②$$

$$\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$= \frac{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2}{n}$$

$$= \frac{\sum x_i^2}{n} - \frac{2\bar{x}\sum x_i}{n} + \frac{\sum \bar{x}^2}{n}$$

$$= \frac{\sum x_i^2}{n} - \frac{2\bar{x}\,\overset{\sum x_i}{\cancel{n\bar{x}}}}{n} + \frac{\cancel{n}\bar{x}^2}{\cancel{n}}$$

$$= \frac{\Sigma x_i^2}{n} - \frac{2\bar{x}\,\cancel{n}\bar{x}}{\cancel{n}} + \bar{x}^2$$

$$= \frac{\Sigma x_i^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{\Sigma x_i^2}{n} - \bar{x}^2$$

$$= \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2$$

$$= \frac{\Sigma x_i^2}{n} - \frac{(\Sigma x_i)^2}{n^2}$$

$$= \frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n^2}$$

$$\sigma_x = \frac{\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]^{1/2}}{n} \longrightarrow ③$$

|||ly

$$\sigma_y = \frac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n} \longrightarrow ④$$

sub ②, ③ & ④ in ①

$$r_{(x,y)} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\cancel{n}} \times \frac{n\cancel{^2}}{\cancel{n}\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]^{1/2} \times \left[n\Sigma y_i^2 - (\Sigma y_i)^2\right]^{1/2}}$$

$$r_{(x,y)} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2}\left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$$

Hence proved.

Theorem(2)

2) Prove that the correlation coefficient is indipened of the change of origin and scale.

$A \& B \rightarrow$ origin

$h \& k \rightarrow$ scale

Let $u_i = \frac{x_i - A}{h}$

$v_i = \frac{y_i - B}{k}$

~~we have~~ To prove $r_{(x,y)} = r_{(u,v)}$

$$u_i = \frac{x_i - A}{h}$$

$$hu_i = x_i - A$$

$$x_i = hu_i + A$$

$$\frac{x_i}{n} = \frac{hu_i}{n} + \frac{A}{n}$$

$$\frac{\Sigma x_i}{n} = \frac{\Sigma h u_i}{n} + \frac{\Sigma A}{n}$$

$$\bar{x} = h\bar{u} + \frac{nA}{n}$$

$$\bar{x} = h\bar{u} + A$$

$$x_i - \bar{x} = hu_i + \cancel{A} - h\bar{u} - \cancel{A}$$

$$= hu_i - h\bar{u}$$

$$= h(u_i - \bar{u}) \qquad u_i = \frac{x_i - A}{h}$$

$$(x_i - \bar{x})^2 = h^2 (u_i - \bar{u})^2 \qquad v_i = \frac{y_i - B}{k}$$

$$\frac{(x_i - \bar{x})^2}{n} = \frac{h^2 (u_i - \bar{u})^2}{n}$$

$$\frac{\Sigma (x_i - \bar{x})^2}{n} = h^2 \frac{\Sigma (u_i - \bar{u})^2}{n}$$

$$\sigma_x^2 = h^2 \sigma_u^2$$

$$\sigma_x = h \sigma_u$$

III^{ly} $y_i - \bar{y} = k(v_i - \bar{v})$

$$\sigma_y = k \sigma_v$$

$$\frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$u_i = \frac{x_i - A}{h}$$

$$v_i = \frac{y_i - B}{k}$$

$$r_{(x,y)} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma_x\,\sigma_y}$$

$$= \frac{\sum \cancel{h}(u_i-\bar{u})\,\cancel{k}(v_i-\bar{v})}{n(\cancel{h}\sigma_u)(\cancel{k}\sigma_v)}$$

$$= \frac{\sum (u_i-\bar{u})(v_i-\bar{v})}{n\,\sigma_u\,\sigma_v}$$

$$r_{(x,y)} = r_{(u,v)}$$

Hence proved.

Theorem(3)

3) prove that $-1 \le r \le 1$

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

$$= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n\sqrt{\frac{\sum (x_i-\bar{x})^2}{n}}\cdot\sqrt{\frac{\sum (y_i-\bar{y})^2}{n}}}$$

$$= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n\,\frac{\sqrt{\sum (x_i-\bar{x})^2}}{\sqrt{n}}\cdot\frac{\sqrt{\sum (y_i-\bar{y})^2}}{\sqrt{n}}}$$

$$= \frac{\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}}{\sigma}$$

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

Let $(x_i-\bar{x}) = a_i$, $y_i-\bar{y} = b_i$

$$r = \frac{\Sigma a_i b_i}{\sqrt{\Sigma a_i^2}\sqrt{\Sigma b_i^2}}$$

squaring,

$$r^2 = \frac{(\Sigma a_i b_i)^2}{\Sigma a_i^2 \, \Sigma b_i^2} \rightarrow (1)$$

By schwartz inequality

we have, $(\Sigma a_i b_i)^2 \leq \Sigma a_i^2 \, \Sigma b_i^2$

$$(1) \Rightarrow r^2 \leq \frac{\Sigma a_i^2 \, \Sigma b_i^2}{\Sigma a_i^2 \, \Sigma b_i^2}$$

$$r^2 \leq 1$$

(ie) $|r| \leq 1$

(ie) $-1 \leq r \leq 1$

Hence proved

Note:

(i) If $r = 1$ the correlation is perfect and positive

(ii) If $r = -1$ the correlation is perfect and negative.

(iii) If $r = 0$ the variables are uncorrelated.

(iv) If the variables x & y are uncorrelated then $cov(x, y) = 0$

Theorem: (4)

Prove that $r(x,y) = \dfrac{\sigma_x^2 + \sigma_y^2 - (\sigma_{x-y}^2)}{2\sigma_x \sigma_y}$ $2r_x$

$\sigma_{x-y}^2 = \dfrac{\Sigma[(x_i - y_i) - (\bar{x} - \bar{y})]^2}{n}$

$\sigma_{x-y}^2 = \dfrac{\Sigma[(x_i - y_i) - (\bar{x} - \bar{y})]^2}{n}$

$= \dfrac{\Sigma[x_i - y_i - \bar{x} + \bar{y}]^2}{n}$

$$= \frac{\Sigma[(x_i - \bar{x}) - (y_i - \bar{y})]^2}{n}$$

$$= \frac{\Sigma[(x_i-\bar{x})^2 - 2(x_i-\bar{x})(y_i-\bar{y}) + (y_i-\bar{y})^2]}{n}$$

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

$$\to r_{xy} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y}$$

$$= \sigma_x^2 - \frac{2 r_{xy} \cancel{n} \sigma_x \sigma_y}{\cancel{n}} + \sigma_y^2$$

$$\sigma_{x-y}^2 = \sigma_x^2 - 2r_{xy}\sigma_x\sigma_y + \sigma_y^2$$

$$2r_{xy}\sigma_x\sigma_y = \sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2$$

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

∴ Hence proved

1) Ten students obtained the following % of mark in the college internal test (x) and in the final university exam (Y) find the correlation co-efficient between the markes of two test

X	51	63	63	49	58	60	65	63	46	50
Y	49	72	75	50	48	60	70	48	60	56

We have

$$r_{(x,y)} = r_{uv}$$

Let $u_i = x_i - 50$

$v_i = y_i - 48$

$$r_{uv} = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\left[n\sum u_i^2 - (\sum u_i)^2\right]^{1/2}\left[n\sum v_i^2 - (\sum v_i)^2\right]^{1/2}}$$

x_i	y_i	$u_i = x_i - 50$	$v_i = y_i - 48$	u_i^2	v_i^2	$u_i v_i$
51	49	1	1	1	1	1
63	72	13	24	169	576	312
63	75	13	27	169	729	351
49	50	-1	2	1	4	-2
50	48	0	0	0	0	0
60	60	10	12	100	144	120
65	70	15	22	225	484	330
63	48	13	0	169	0	0
46	60	-4	12	16	144	-48
50	56	0	8	0	64	0
		$\Sigma u_i = 60$	$\Sigma v_i = 108$	$\Sigma u_i^2 = 850$	$\Sigma v_i^2 = 2146$	$\Sigma u_i v_i = 1064$

$$r_{uv} = \frac{10 \times 1064 - 60 \times 108}{\left[10 \times 850 - (60^2)\right]^{1/2} \left[10 \times 2146 - 108^2\right]^{1/2}}$$

$$= \frac{10640 - 6480}{(8500 - 3600)^{1/2} \left[21460 - 11664\right)^{1/2}}$$

$$= \frac{4160}{(4900)^{1/2}(9796)^{1/2}}$$

$$= \frac{4160}{70 \times 98.97} = \frac{4160}{6927.9}$$

$$= 0.6$$

2) Find the correlation coefficient between two variables

X	300	350	400	450	500	550	600	650	700
Y	800	900	1000	1100	1200	1300	1400	1500	1600

We have

$$r_{xy} = r_{uv}$$

Let $u_i = \frac{x_i - 500}{50}$

$v_i = \frac{y_i - 1200}{100}$

$$r_{uv} = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\left[n\sum u_i^2 - (\sum u_i)^2\right]^{1/2}\left[n\sum v_i^2 - (\sum v_i)^2\right]^{1/2}}$$

x_i	y_i	$u_i = \frac{x_i - 500}{50}$	$v_i = \frac{y_i - 1200}{100}$	u_i^2	v_i^2	$u_i v_i$
300	800	−4	−4	16	16	16
350	900	−3	−3	9	9	9
400	1000	−2	−2	4	4	4
450	1100	−1	−1	1	1	1
500	1200	0	0	0	0	0
550	1300	1	1	1	1	1
600	1400	2	2	4	4	4
650	1500	3	3	9	9	9
700	1600	4	4	16	16	16
		$\Sigma u_i = 0$	$\Sigma v_i = 0$	$\Sigma u_i^2 = 60$	$\Sigma v_i^2 = 60$	$\Sigma u_i v_i = 60$

$$r_{uv} = \frac{9 \times 60 - 0 \times 0}{\left[9 \times 60 - (60)\right]^{1/2}\left[9 \times 60 - (60)^2\right]^{1/2}}$$

$$r_{uv} = \frac{9 \times 60 - 0}{(9 \times 60 - 0)^{1/2}(9 \times 60 - 0)^{1/2}}$$

$$= \frac{540}{(540)^{1/2}(540)^{1/2}}$$

$$= \frac{9 \times 60}{\left[9 \times 60 - 3600\right]^{1/2}\left[9 \times 60 - 3600\right]^{1/2}}$$

$$= \frac{540}{(540 - 3600)^{1/2}(540 - 3600)^{1/2}}$$

$$= \frac{540}{(-3060)^{1/2}(-3060)^{1/2}}$$

$$= \frac{540}{}$$

$= \frac{540}{\sqrt{540}\sqrt{540}}$

$= \frac{540}{540} = 1$

$r = 1$ then the correlation between perfect and positive.

3) A ~~pro~~ programmes while writing a program for correlation co-efficient between 2 variables x & y from 30 pairs of observations obtained the following result $\Sigma x = 300$, $\Sigma x^2 = 3710$, $\Sigma y = 210$, $\Sigma y^2 = 2000$, $\Sigma xy = 2100$ at the time of checking it was found that he had copied down 2 pairs (x_i, y_i) as $(10, 20)$ and $(12, 10)$ instead of the correct value $(10, 15)$ and $(20, 15)$ Obtain the correct value of the correlation co-efficient.

$\Sigma x = 300$, $\Sigma x^2 = 3718$, $\Sigma y = 210$, $\Sigma y^2 = 2000$

$\Sigma xy = 2100$

(x_i, y_i) as $(18, 20)$ & $(12, 10)$ → wrong values

$(10, 15)$ & $(20, 15)$ → correct values

corrected

$\Sigma x = 300 - 18 - 12 + 10 + 20$

$= 300$

$\Sigma x^2 = 3718 - 18^2 - 12^2 + 10^2 + 20^2$

$= 3750$

$\Sigma y = 210 - 20 - 10 + 15 + 15$

$= 210$

$\Sigma y^2 = 2000 - 20^2 - 10^2 + 15^2 + 15^2$

$= 1950$

$\Sigma xy = 2100 - (18 \times 20) - (12 \times 10) + (10 \times 15) + (20 \times 15)$

$= 2070$

The corrected values are $\Sigma x = 300$, $\Sigma x^2 = 3750$

$\Sigma y = 210$, $\Sigma y^2 = 1950$, $\Sigma xy = 2070$

Here $n = 30$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{[n\sum x_i^2 - (\sum x_i)^2]^{1/2} [n\sum y_i^2 - (\sum y_i)^2]^{1/2}}$$

$$= \frac{30 \times 2070 - 300 \times 210}{[30 \times 3750 - (300)^2]^{1/2} [30 \times 1950 - (210)^2]^{1/2}}$$

$$= \frac{-900}{(22500)^{1/2} (14400)^{1/2}}$$

$$= \frac{-900}{150 \times 120}$$

$$= \frac{-900}{18000}$$

$$= -1/20$$

$$= -0.05$$

17) If x and y are two variable prove that the correlation co-efficient between $ax+b$ & $cy+d$ is $r_{ax+b, cy+d}$

$$r_{ax+b, cy+d} = \frac{ac}{|ac|} r_{xy} \text{ if } ac \neq 0$$

Let $u_i = ax_i + b$, $v_i = cy_i + d$

$$\bar{u} = \frac{\sum u_i}{n}$$

$$\bar{u} = \frac{\sum (ax_i + b)}{n}$$

$$= \frac{\sum ax_i}{n} + \frac{\sum b}{n}$$

$$= \frac{a\sum x_i}{n} + \frac{\cancel{n}b}{\cancel{n}}$$

$$= a\frac{\sum x_i}{n} + b$$

$$\bar{u} = a\bar{x} + b$$

$$\frac{n\sum x_i y_i - \sum x_i \sum y_i}{\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2}\left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$$

III^{ly} $\bar{v} = c\bar{y} + d$

$$r_{uv} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{n\sigma_u \sigma_v}$$

$$\sigma_u^2 = \frac{\sum (u_i - \bar{u})^2}{n}$$

$$= \frac{\sum \left[(ax_i + b) - (a\bar{x} + b)\right]^2}{n}$$

$$= \frac{\sum \left[ax_i + \cancel{b} - a\bar{x} - \cancel{b}\right]^2}{n}$$

$$= \frac{\sum \left[ax_i - a\bar{x}\right]^2}{n}$$

$$= \frac{a^2 \Sigma (x_i - \bar{x})^2}{n}$$

$$= a^2 \sigma_x^2$$

$|||^{ly}$ $\sigma_v^2 = c^2 \cdot \sigma_y^2$

$$\sigma_u^2 \sigma_v^2 = a^2 \sigma_x^2 \cdot c^2 \cdot \sigma_y^2$$

$$(\sigma_u \cdot \sigma_v)^2 = (ac)^2 (\sigma_x \cdot \sigma_y)^2$$

Taking square root

$$\sigma_u \sigma_v = |ac| \sigma_x \sigma_y$$

$$r_{uv} = \frac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v}$$

$$= \frac{\Sigma [(ax_i + \cancel{b} - a\bar{x} - \cancel{b})(cy_i + \cancel{d} - c\bar{y} - \cancel{d})]}{n |ac| \sigma_x \sigma_y}$$

$$= \frac{\Sigma [(ax_i - a\bar{x})(cy_i - c\bar{y})]}{n |ac| \sigma_x \sigma_y}$$

$$= \frac{ac \, \Sigma (x_i - \bar{x})(y_i - \bar{y})}{n |ac| \sigma_x \sigma_y}$$

$$r_{uv} = \frac{ac}{|ac|} \; r(x,y)$$

Hence proved

2) If x, y and z are uncorrelated variables each having same standard deviation obtain the correlation co-efficient between $x+y$ & $y+z$

x, y & z are uncorrelated variables

x and y are uncorrelated $\Rightarrow cov(x,y) = 0$

$$(ie)\; \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n} = 0$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 0$$

y and z are uncorrelated $\Rightarrow cov(y,z) = 0$

$$\Sigma(y_i - \bar{y})(z_i - \bar{z}) = 0$$

z and x are uncorrelated $\Rightarrow cov(z,x) = 0$

$$\Sigma(z_i - \bar{z})(x_i - \bar{x}) = 0$$

Also given $\sigma_x = \sigma_y = \sigma_z = \sigma$

To find correlation co-efficient between $x+y$ & $y+z$

Let $u_i = x_i + y_i$, $v_i = y_i + z_i$

$\bar{u} = \bar{x} + \bar{y}$, $\bar{v} = \bar{y} + \bar{z}$

$$r_{uv} = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{n \sigma_u \sigma_v} \rightarrow ①$$

~~= (Σ x_i + y_i ...)~~

$$\sum (u_i - \bar{u})(v_i - \bar{v}) = \sum \left\{ \left[(x_i + y_i) - (\bar{x} + \bar{y}) \right] \left[(y_i + z_i) - (\bar{y} + \bar{z}) \right] \right\}$$

$$= \sum \left[(x_i + y_i - \bar{x} - \bar{y}) \right] \left[(y_i + z_i - \bar{y} - \bar{z}) \right]$$

$$= \sum \left[(x_i - \bar{x}) + (y_i - \bar{y}) \right] \left[(y_i - \bar{y}) + (z_i - \bar{z}) \right]$$

$$= \sum \left[(x_i - \bar{x})(y_i - \bar{y}) + (x_i - \bar{x})(z_i - \bar{z}) + (y_i - \bar{y})(y_i - \bar{y}) + (y_i - \bar{y})(z_i - \bar{z}) \right]$$

$$= \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (x_i - \bar{x})(z_i - \bar{z}) + \sum (y_i - \bar{y})^2 + \sum (y_i - \bar{y})(z_i - \bar{z})$$

$$= 0 + \sum (y_i - \bar{y})^2 + 0$$

$$= \Sigma(y_i-\bar{y})^2 \qquad \left[\sigma_y^2 = \frac{\Sigma(y_i-\bar{y})^2}{n}\right.$$

$$\Sigma(y_i-\bar{y})^2 = \sigma_y^2 n$$

$$\Sigma(u_i-\bar{u})(u_i-\bar{u}) = n\sigma_y^2$$

$$= n\sigma^2$$

$$\sigma_u^2 = \frac{\Sigma(u_i-\bar{u})^2}{n}$$

$$= \frac{\Sigma[(x_i+y_i)-(\bar{x}+\bar{y})]^2}{n}$$

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

$$= \frac{\Sigma[(x_i-\bar{x})+(y_i-\bar{y})]^2}{n}$$

$$= \frac{\Sigma[(x_i-\bar{x})^2+(y_i-\bar{y})^2+2(x_i-\bar{x})(y_i-\bar{y})]}{n}$$

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

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2$$

$$= \sigma^2 + \sigma^2 = 2\sigma^2$$

$$\sigma_u = \sqrt{2}\,\sigma$$

$|||^{ly}$ $\sigma_v = \sqrt{2}\sigma$

$$① \Rightarrow r_{uv} = \frac{\cancel{n}\sigma^2}{\cancel{n}\sqrt{2}\cancel{\sigma}\cdot\sqrt{2}\cancel{\sigma}}$$

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

1) Show that the variables $u = x\cos\alpha + y\sin\alpha$ and $v = y\cos\alpha - x\sin\alpha$ are uncorrelated if $\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$

Let $u_i = x_i\cos\alpha + y_i\sin\alpha$

$v_i = y_i\cos\alpha - x_i\sin\alpha$

$$\bar{u} = \frac{\Sigma(x_i\cos\alpha + y_i\sin\alpha)}{n}$$

$$= \bar{x}\cos\alpha + \bar{y}\sin\alpha$$

$|||^{ly}$ $\bar{v} = \bar{y}\cos\alpha - \bar{x}\sin\alpha$

$u_i - \bar{u} = (x_i\cos\alpha + y_i\sin\alpha) - (\bar{x}\cos\alpha + \bar{y}\sin\alpha)$

$\cancel{u_i - \bar{u}} = (x_i - \bar{x})\cos\alpha + (y_i - \bar{y})\sin\alpha$

$v_i - \bar{v} = (y_i - \bar{y})\cos\alpha - (x_i - \bar{x})\sin\alpha$

u & v are uncorrelated

(i.e) $r_{uv} = 0$

(i.e) $\dfrac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n} = 0$

(i.e) $\Sigma (u_i - \bar{u})(v_i - \bar{v}) = 0$

$$\Sigma \left[[(x_i - \bar{x})\cos\alpha + (y_i - \bar{y})\sin\alpha] [(y_i - \bar{y})\cos\alpha - (x_i - \bar{x})\sin\alpha] \right] = 0$$

$$\Sigma \left[(x_i - \bar{x})(y_i - \bar{y})\cos^2\alpha - (x_i - \bar{x})^2 \sin\alpha\cos\alpha + (y_i - \bar{y})^2 \sin\alpha\cos\alpha - (y_i - \bar{y})(x_i - \bar{x})\sin^2\alpha \right] = 0$$

$$\Sigma \left[[(x_i - \bar{x})(y_i - \bar{y})](\cos^2\alpha - \sin^2\alpha) - [(x_i - \bar{x})^2 - (y_i - \bar{y})^2][\sin\alpha\cos\alpha] \right] = 0$$

$$(\cos^2\alpha - \sin^2\alpha)\, \Sigma (x_i - \bar{x})(y_i - \bar{y}) - \sin\alpha\cos\alpha\, \Sigma [(x_i - \bar{x})^2 - (y_i - \bar{y})^2] = 0$$

$$\cos 2\alpha \cdot n\, r_{xy}\, \sigma_x \sigma_y - \frac{2\sin\alpha\cos\alpha}{2} \left[\Sigma (x_i - \bar{x})^2 - \Sigma (y_i - \bar{y})^2 \right] = 0$$

$$\cos 2\alpha \cdot n\, r_{xy}\, \sigma_x \sigma_y - \frac{\sin 2\alpha}{2} \left[n\sigma_x^2 - n\sigma_y^2 \right] = 0$$

$$\cos 2\alpha \cdot n\, r_{xy} \cdot \sigma_x \sigma_y = \frac{1}{2} \sin 2\alpha \left[n(\sigma_x^2 - \sigma_y^2) \right]$$

$$2\cos 2\alpha \cdot n \cdot r_{xy}\, \sigma_x \sigma_y = \sin 2\alpha\; n(\sigma_x^2 - \sigma_y^2)$$

$$\frac{2\, r_{xy}\, \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2} = \frac{\sin 2\alpha}{\cos 2\alpha}$$

$$\frac{2\, r_{xy}\, \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2} = \tan 2\alpha$$

$$2\alpha = \tan^{-1}\left(\frac{2\, r_{xy}\, \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$$

$$\alpha = \frac{1}{2}\tan^{-1}\left(\frac{2\, r_{xy}\, \sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$$

Hence proved.

2) Show that if X', Y' are the deviation of the random variable X, Y from the respective mean. Then

(i) $r = 1 - \frac{1}{2N} \Sigma \left(\frac{x_i'}{\sigma_x} - \frac{y_i'}{\sigma_y}\right)^2$ and

(ii) $r = -1 + \frac{1}{2N} \Sigma \left(\frac{x_i'}{\sigma_x} + \frac{y_i'}{\sigma_y}\right)^2$

(iii) Deduce that $-1 \leq r \leq 1$

soln

vari

Soln:

x' & y' are the deviation of the random variable x & y

$\therefore x' = x_i - \bar{x}$, $y_i = y_i - \bar{y}$

$$RHS = 1 - \frac{1}{2N} \Sigma \left(\frac{x_i'}{\sigma x} - \frac{y_i'}{\sigma y} \right)^2$$

$x' = x_i - \bar{x}$

$y = y_i - \bar{y}$

$$= 1 - \frac{1}{2N} \Sigma \left[\left(\frac{x_i'}{\sigma x} \right)^2 - \frac{2 x_i' y_i'}{\sigma x \sigma y} + \left(\frac{y_i'}{\sigma y} \right)^2 \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{\Sigma (x_i')^2}{\sigma x^2} - \frac{2 \Sigma x_i' y_i'}{\sigma x \sigma y} + \frac{\Sigma (y_i')^2}{\sigma y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{\Sigma (x_i - \bar{x})^2}{\sigma x^2} - \frac{2 \Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sigma x \sigma y} + \frac{\Sigma (y_i - \bar{y})^2}{\sigma y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[\frac{N \sigma x^2}{\sigma x^2} - \frac{2N \gamma \sigma x \sigma y}{\sigma x \sigma y} + N \frac{\sigma y^2}{\sigma y^2} \right]$$

$$= 1 - \frac{1}{2N} \left[2N - 2N\gamma \right]$$

$$= 1 - \frac{1}{2\cancel{N}} \cdot 2\cancel{N}(1-\gamma)$$

$$= 1 - 1 + \gamma$$

$$= \gamma$$

$$= \text{LHS}$$

(ii) $\text{RHS} \Rightarrow -1 + \frac{1}{2N} \sum \left[\frac{x_i'}{\sigma x} + \frac{y_i'}{\sigma y} \right]^2$

$$= -1 + \frac{1}{2N} \sum \left[\frac{(x_i')^2}{\sigma x^2} + \frac{(y_i')^2}{\sigma y^2} + \frac{2 x_i' y_i'}{\sigma x \sigma y} \right]$$

(iii)

$$\cancel{= -1 + \frac{1}{2N} \sum [\,]}$$

$$= -1 + \frac{1}{2N} \left[\frac{\sum (x_i')^2}{\sigma x^2} + \frac{\sum (y_i')^2}{\sigma y^2} + \sum \frac{2 x_i' y_i'}{\sigma x \sigma y} \right]$$

$$= -1 + \frac{1}{2N} \left[\frac{\sum (x_i - \bar{x})^2}{\sigma x^2} + \frac{\sum (y_i - \bar{y})^2}{\sigma y^2} + \frac{2 \sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma x \sigma y} \right]$$

$$= -1 + \frac{1}{2N} \left[\frac{N \cancel{\sigma x^2}}{\cancel{\sigma x^2}} + \frac{N \cancel{\sigma y^2}}{\cancel{\sigma y^2}} + 2N\gamma \right]$$

$$= -1 + \frac{1}{2N} \left[N + N + 2N\gamma \right]$$

$$= -1 + \frac{1}{2N}\left[2N + 2N r\right]$$

$$= -1 + \frac{1}{\cancel{2N}} \cdot \cancel{2N}\left[1 + r\right]$$

$$= \cancel{-1} + \cancel{1} + r$$

$$= r$$

$\cancel{RHS} = LHS$

(iii) From ①

$$1 - \frac{1}{2N} \sum \left(\frac{x_i'}{\sigma x} - \frac{y_i'}{\sigma y}\right)^2 \leq 1$$

$$r \leq 1 \rightarrow ⓐ$$

$$r = 1 - \frac{1}{2N} \sum \left(\frac{x_i'}{\sigma x} - \frac{y_i'}{\sigma y}\right)$$

From ②

$$-1 + \frac{1}{2N} \sum \left(\frac{x_i'}{\sigma x} + \frac{y_i'}{\sigma y}\right)^2 \geq -1$$

$$r \geq -1$$

$$\text{i.e., } -1 \leq r \rightarrow ⓑ$$

From ⓐ & ⓑ we get

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

Hence proved

1) Let x, y be two variable with standard deviation σ_x & σ_y respectively. If $u = x + ky$, $v = x + \left(\frac{\sigma_x}{\sigma_y}\right) y$ & $r_{uv} = 0$ then find the value of k.

<u>soln:</u>

$u_i = x_i + k y_i$ $\qquad$ $v_i = x_i + \left(\frac{\sigma_x}{\sigma_y}\right) y_i$

$\bar{u} = \bar{x} + k\bar{y}$ $\qquad$ $\bar{v} = \bar{x} + \left(\frac{\sigma_x}{\sigma_y}\right) \bar{y}$

$$u_i - \bar{u} = x_i + k y_i - \bar{x} - k\bar{y}$$
$$= (x_i - \bar{x}) + k(y_i - \bar{y})$$

$$v_i - \bar{v} = \cancel{x_i + k y_i}\ x_i + (\sigma_x/\sigma_y) y_i - \bar{x} - \left(\frac{\sigma_x}{\sigma_y}\right)\bar{y}$$
$$= (x_i - \bar{x}) + \left(\frac{\sigma_x}{\sigma_y}\right)(y_i - \bar{y})$$

$r_{uv} = 0$

$cov(u,v) = 0$

$\Sigma(u_i - \bar{u})(v_i - \bar{v}) = 0$

$$\Sigma\left[\left[(x_i - \bar{x}) + k(y_i - \bar{y})\right]\left[(x_i - \bar{x}) + \left(\frac{\sigma_x}{\sigma_y}\right)(y_i - \bar{y})\right]\right] = 0$$

$$\Sigma\left[(x_i-\bar{x})^2+\left(\frac{\sigma x}{\sigma y}\right)(x_i-\bar{x})(y_i-\bar{y})+k(y_i-\bar{y})(x_i-\bar{x})+\right.$$
$$\left.k\left(\frac{\sigma x}{\sigma y}\right)(y_i-\bar{y})^2\right]=0$$

$$\left(\Sigma(x_i-\bar{x})^2+\left(\frac{\sigma x}{\sigma y}\right)\Sigma(x_i-\bar{x})(y_i-\bar{y})+k\,\Sigma(y_i-\bar{y})(x_i-\bar{x})+\right.$$
$$\left.k\left(\frac{\sigma x}{\sigma y}\right)\Sigma(y_i-\bar{y})^2\right]=0$$

$$n\sigma x^2+\left(\frac{\sigma x}{\sigma y}\right)n\,\sigma x\,\sigma y\,r_{xy}+k\,n\,\sigma x\,\sigma y\,r_{xy}+$$
$$k\left(\frac{\sigma x}{\sigma y}\right)n\sigma y^2=0$$

$$n\sigma x^2+n\sigma x^2 r_{xy}+k\,n\,\sigma x\,\sigma y\,r_{xy}+k\,\sigma x\,\sigma y\,n=0$$

$$n\sigma x\left[\sigma x+\sigma x\, r_{xy}+k\,\sigma y\, r_{xy}+k\sigma y\right]=0$$

$$\sigma x\left[\sigma x+\sigma x\, r_{xy}+k\,\sigma y\, r_{xy}+k\sigma y\right]=0$$

$$\sigma x\left[\sigma x(1+r_{xy})+k\,\sigma y\,(1+r_{xy})\right]=0$$

$$\sigma x\,(1+r_{xy})(\sigma x+k\sigma y)=0$$

$$\sigma x=0 \text{ (or) } 1+r_{xy}=0 \text{ (or) } \sigma x+k\sigma y=0$$

$$\sigma x+k\sigma y=0$$

$$k\sigma y=-\sigma x$$

$$k=-\left(\frac{\sigma x}{\sigma y}\right)$$

If $r_{xy}\neq -1$, $\sigma x\neq 0$ we get $k=-\left(\frac{\sigma x}{\sigma y}\right)$

Internal.

Rank correlation:

Let (x_i, y_i) be the ranks of the i^{th} individual in the first & II ranking respectively in the coefficient of correlation between the rank (x_i, y_i) are called the rank correlation co efficient is denoted by ρ (rou)

$$\rho = 1 - \frac{\sum(x-y)^2}{n(n^2-1)}$$

Therom:

P.T the rank correlation co.efficient ρ is $1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$

consider the collection of n individuals

Let x_i and y_i be the ranks of the i^{th} individuals

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{1+2+\dots+n}{n}$$

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

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

$$\sigma_x^2 = \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2$$

$$\Sigma x_i^2 = 1^2 + 2^2 + \dots + n^2$$

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

$$\sigma_x^2 = \frac{n(n+1)(2n+1)}{6n} - \left[\frac{n+1}{2}\right]^2$$

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

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

$$= \frac{(n+1)\left[2(2n+1) - 3(n+1)\right]}{12}$$

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

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

$$\sigma_x^2 = \frac{n^2-1}{12}$$

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

$$\sigma_x^2 = \sigma_y^2 = \frac{n^2-1}{12}$$

Now $\Sigma(x-y)^2 = \Sigma(x-\bar{x}+\bar{y}-y)^2$ $[\because \bar{x}=\bar{y}]$

$$= \Sigma[(x-\bar{x})-(y-\bar{y})]^2$$

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

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

$$= n\sigma_x^2 - 2n\rho\sigma_x\sigma_y + n\sigma_y^2$$

$[\because \sigma_x^2 = \sigma_y^2 = \sigma^2]$

$$= n\sigma^2 - 2n\rho\sigma^2 + n\sigma^2$$

$$= 2n\sigma^2 - 2n\rho\sigma^2$$

$$\Sigma(x-y)^2 = 2n\sigma^2(1-\rho)$$

$$1-\rho = \frac{\Sigma(x-y)^2}{2n\sigma^2}$$

$$\rho = 1 - \frac{\Sigma(x-y)^2}{2n\sigma^2}$$

$$= 1 - \frac{\Sigma(x-y)^2}{2n \cdot \frac{n^2-1}{12}}$$

$$= 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$\rho = \sqrt{\frac{\Sigma(x-y)^2}{N}}$

$\frac{n^2}{6}$

This is called the spearman's ρ or formula for the rank correlation

1) Find the rank correlation co-efficient between the height in cm and weight in kg of 6 soldiers in Indian Army.

Height (in cm)	165	167	166	170	169	172
Weight (in kg)	61	60	63.5	63	61.5	64

Height (in cm)	Rank for height (x)	Weight (in kg)	Rank for weight (y)	$x-y$	$(x-y)^2$
165	6	61	5	1	1
167	4	60	6	-2	4
166	5	63.5	2	3	9
170	2	63	3	-1	1
169	3	61.5	4	-1	1
172	1	64	1	0	0

Here, $n = 6$

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

$$= 1 - \frac{6\times16}{6(36-1)}$$

$$= 1 - \frac{96}{6\times35}$$

$$= 1 - \frac{96}{210}$$

$$\rho = 0.542 \text{ //Ans.}$$

Note:

If two (or) more individuals get the same Rank in the ranking process we assign the common rank to the repeated values. This common rank is the average of the ranks, and the next item will get the rank next to the rank already assumed. As a result of this in the formula for the ρ we add the factor $\frac{m(m^2-1)}{12}$ to $\Sigma(x-y)^2$. Where m is the number of times an item has repeated values. This correlation factor added for each repeated rank of the variables (x,y).

Find from the following for data in marks uptaining to the ten students in Physics and chemistry.

Calculate the rank correlation.

Physics 35 56 50 65 44 38 44 50 15 26

Chemistry 50 35 70 25 35 58 75 60 55 35

Physics	Rank (in Physics) (x)	Chemistry	Rank in Chemistry (y)	$x-y$	$(x-y)^2$
35	8	50	6	2	4
56	2	35	8	-6	36
50	3.5	70	2	1.5	2.25
65	1	25	10	-9	81
44	5.5	35	9	-2.5	6.25
38	7	58	4	3	9
44	6.5	75	1	4.5	20.25
50	3.5	60	3	0.5	0.25
15	10	55	5	5	25
26	9	35	8	1	1

$\Sigma(x-y)^2 = 18[illegible]$

Physics

thrice in

$n = 10$

Here the marks 50 and 44 occur ~~...~~ repeated twice in x and marks 35 occur thrice in y

$$\sum(x-y)^2 = \sum(x-y)^2 + \frac{m(m^2-1)}{12}$$

Hence corrected $\sum(x-y)^2$ = actual $\sum(x-y)^2 + \frac{m(m^2-1)}{12}$

$$\frac{m(m^2-1)}{12}$$

$$\sum(x-y)^2 = 185 + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}$$

$$= 185 + \frac{2(3)}{12} + \frac{2(3)}{12} + \frac{3(8)}{12}$$

$$= 185 + \frac{6}{12} + \frac{6}{12} + \frac{24}{12}$$

$$= 185 + \frac{1}{2} + \frac{1}{2} + 2$$

$$= 185 + 3$$

$$= 188$$

$$\rho = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 188}{10(100-1)}$$

$$= 1 - \frac{1128}{10(99)}$$

$$= 1 - \frac{1128}{990}$$

$$= \frac{990 - 1128}{990}$$

$$= -0.139 \text{ Ans.}$$

2) Calculate the rank correlation co-efficient for the following data

x	20	25	60	45	80	25	55	65	25	75
y	52	50	55	50	60	70	72	78	80	60

x	Rank in x	y	Rank in y	$x-y$	$(x-y)^2$
20	10	52	8	2	4
25	8	50	9.5	-1.5	2.25
60	4	55	7	-3	9
45	6	50	9.5	-3.5	12.25

80	1	60	6	-5	25
25	8	70	4	4	16
55	5	72	3	2	4
65	3	78	2	1	1
25	8	80	1	7	49
75	2	63	5	-3	9

$\Sigma(x-y)^2 = 131.5$

$n=10$

Here the 25 occur thrice in x and 80 occur twice in y.

Hence corrected $\Sigma(x-y)^2$ = actual $\Sigma(x-y)^2 + \frac{m(m^2-1)}{12}$

$= 131.5 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$

$= 131.5 + \frac{3(9-1)}{12} + \frac{2(4-1)}{12}$

$= 131.5 + \frac{3(8)}{12} + \frac{2(3)}{12}$

$= 131.5 + \frac{24}{12} + \frac{6}{12}$

$= 131.5 + 2 + \frac{1}{2}$

$= \frac{131.5\times2+2\times2+1}{2}$

$= 134$

cient

25 75

80 63

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

$$= 1 - \frac{6 \times 134}{10 \times (100-1)}$$

$$= 1 - \frac{804}{10 \times 99}$$

$$= 1 - \frac{804}{990}$$

$$= \frac{990-804}{990}$$

$$= 0.1878$$

7) Three Judges Asign the ranks to 8 entries in a beauty contest

Judge Mr. X	1	2	4	3	7	6	5	8
Judge Mr. Y	3	2	1	5	4	7	6	8
Judge Mr. Z	1	2	3	4	5	7	8	6

Which pair of judges has a nearest approach to common taste in beauty.

We have to find

ρ_{xy}, ρ_{yz}, ρ_{zx}

x	y	z	$x-y$	$(x-y)^2$	$y-z$	$(y-z)^2$	$z-x$	$(z-x)^2$
1	3	1	-2	4	2	4	0	0
2	2	2	0	0	0	0	0	0
4	1	8	3	9	-2	4	-1	1
3	5	4	-2	4	1	1	1	1
7	4	5	3	9	-1	1	-2	4
6	7	7	-1	1	0	0	1	1
5	6	8	-1	1	-2	4	-3	9
1	8	6	0	0	2	4	-2	4
				$\Sigma(x-y)^2 = 28$		$\Sigma(y-z)^2 = 18$		$\Sigma(z-x)^2 = 20$

$n = 8$

$$\rho_{xy} = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 28}{8 \times (64-1)}$$

$$= 1 - \frac{6 \times 28}{8 \times 63}$$

$$= 1 - \frac{168}{504} = \frac{504-168}{504}$$

$$= 0.6666\ldots$$

$$\rho_{(yz)} = 1 - \frac{6\sum(y-z)^2}{n(n^2-1)}$$

$$= 1 - \frac{6(18)}{8(64-1)}$$

$$= 1 - \frac{6\times 18}{8(63)}$$

$$= 1 - \frac{108}{504}$$

$$= \frac{504-108}{504}$$

$$= 0.7857$$

$$\rho_{(zx)} = 1 - \frac{6\sum(z-x)^2}{n(n^2-1)}$$

$$= 1 - \frac{6(20)}{8\times 63}$$

$$= 1 - \frac{120}{504} = \frac{504-120}{504}$$

$$= 0.7619$$

Since $\rho_{yz} > \rho_{xy}$ and ρ_{zx}

Hence the judges Mr. y and Mr. z have nearest approach to common taste in beauty.

2) The co-efficient of Rank correlation of marks obtained by 10 students in Maths and physics was found to be 0.8. It was latest discovered that the difference in ranks in two subjects obtained by one of the student was wrongly taken as five instead of 8. Find the correct co-efficient of rank correlation.

Soln:

Here $n = 10$

$\rho = 0.8$

$$0.8 = 1 - \frac{6\sum(x-y)^2}{10(10^2-1)}$$

$$0.8 = 1 - \frac{6\sum(x-y)^2}{990}$$

$$\frac{6\sum(x-y)^2}{990} = 1 - 0.8$$

$$\frac{6\Sigma(x-y)^2}{990} = 0.2$$

$$6\Sigma(x-y)^2 = 0.2 \times 990$$

$$6\Sigma(x-y)^2 = 198$$

$$\Sigma(x-y)^2 = \frac{198}{6}$$

$$\Sigma(x-y)^2 = 33$$

corrected $\Sigma(x-y)^2 = 33 - 5^2 + 8^2$

$$= 33 - 25 + 64$$

$$= 72$$

$$\rho = 1 - \frac{6 \times 72}{10(10^2-1)}$$

$$= 1 - \frac{432}{990}$$

$$= \frac{990 - 432}{990}$$

$$= 0.564$$

Q. Let $(x_1, x_2, \ldots x_n)$ be the ranks of n individuals according to the character of A and $y_1, y_2 \ldots y_n$ be the ranks of same individuals according to another character B. It is given that $x_i + y_i = 1+n$ for $i = (1, 2 \ldots n)$. Show that the values of the rank correlation co-efficient ρ between the character A & B is (-1)

Given:

$$x_i + y_i = 1+n \longrightarrow ①$$

Let d_i be the difference between the two ranks x_i & y_i for $i = 1, 2, \ldots n$.

$$d_i = x_i - y_i \longrightarrow ②$$

$$① - ② \Rightarrow x_i + y_i - d_i = (1+n) - (x_i - y_i)$$

$$x_i + y_i - (x_i - y_i) = (1+n) - d_i$$

$$\cancel{x_i} + y_i - \cancel{x_i} + y_i = (1+n) - d_i$$

$$2y_i = (1+n) - d_i$$

$$d_i = (1+n) - 2y_i$$

Rank correlation $\rho = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)} \rightarrow (2)$

$$= 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\sum d_i^2 = \sum\left[(1+n) - 2y_i\right]^2$$

$$= \sum\left[(1+n)^2 - 2(1+n)2y_i + (2y_i)^2\right]$$

$$= \sum\left[(1+n)^2 - 4(n+1)y_i + 4y_i^2\right]$$

$$= \sum(1+n)^2 - 4(n+1)\sum y_i + 4\sum y_i^2$$

$$= n(n+1)^2 - 4(n+1)\sum y_i + 4\sum y_i^2$$

now $\sum y_i = 1+2+\cdots+n = \frac{n(n+1)}{2}$

$$\sum y_i^2 = 1^2+2^2+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum d_i^2 = n(n+1)^2 - \cancel{4}^{2}(n+1)\frac{(n+1)n}{\cancel{2}} + \cancel{4}^{2}\frac{n(n+1)(2n+1)}{\cancel{6}_{3}}$$

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

1) The c

moves

captain

0.8, 9

$$= n(n+1)\left[\frac{3n+3-6n-6+4n+2}{3}\right]$$

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

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

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

$$= \frac{1 - \frac{\overset{2}{\cancel{6}} \times n(n^2-1)}{\cancel{3}}}{n(n^2-1)}$$

$$= 1 - 2\cancel{n}\cancel{(n^2-1)} \times \frac{1}{\cancel{n}\cancel{(n^2-1)}}$$

$$= 1 - 2$$

$$\rho = -1$$

Hence proved.

The co-efficient of rank correlation between marks in statistics and mathematics in certain group of student is obtained by a 0.2. If the sum of the squares of the difference in ranks is given to be 33. Find the number of students in a group.

$\rho = 0.8$

$\Sigma(x-y)^2 = 33$

$n = ?$

We have $\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$

$$0.8 = 1 - \frac{6 \times 33}{n(n^2-1)}$$

$$0.8 = 1 - \frac{198}{n(n^2-1)}$$

$$\frac{198}{n(n^2-1)} = 1 - 0.8$$

$$\frac{198}{n(n^2-1)} = 0.2$$

$$n(n^2-1) = \frac{198}{0.2}$$

$$n(n^2-1) = 990$$

$n(n^2-1) = 10(10^2-1)$

Here $n(n^2-1)$ is ~~of~~ the form $10(10^2-1)$

$\boxed{n=10}$

2) The co-efficient of rank correlation ~~between~~ of the marks ~~in~~ ~~[illegible]~~ obtained by 10 students in physics and chemistry pass to be 0.5 it was latter discovered that the differents in ranks the two subjects ~~[illegible]~~ obtained by one of the students for strongly taken as 3 instead of 7. Find the correct co-efficient of rank correlation.

$n = 10, \quad \rho = 0.5$

we have $\rho = 1 - \frac{6\Sigma(x-y)^2}{n(n^2-1)}$

$0.5 = 1 - \frac{6\Sigma(x-y)^2}{10(10^2-1)}$

$\frac{6\Sigma(x-y)^2}{10(100-1)} = 1-0.5$

$\frac{6\Sigma(x-y)^2}{10\times 99} = 0.5$

$$\Sigma(x-y)^2 = \frac{0.5\times 990}{6}$$

$$\Sigma(x-y)^2 = 82.5$$

corrected $\Sigma(x-y)^2 = 82.5 - 3^2+7^2$

$$= 82.5 - 9+49$$

$$= 122.5$$

correct rank correlation

$$\rho = 1-\frac{6\,\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1-\frac{6\times 122.5}{990}$$

$$= 1-0.742$$

$$= 0.258$$

3) Following on the marks explained by 10 student is first 3 semester is 3 ancillary papers out of 75

Semester I (Ancillary I)	60	55	75	45	69	45	72	39	35	45

Semester II ~~80~~ 70 58 73 49 60 49 60 55 60 48
(Ancillary II)

Semester III 55 61 68 40 58 60 50 38 50 60
(Ancillary III)

~~Semester I Rank Semester II Rank Semester III Rank~~

Semester I	Rank x	II	Rank y	III	Rank z	x-y	y-z	x-z	$(x-y)^2$	$(y-z)^2$	$(x-z)^2$
60	4	70	2	55	6	2	-4	-2	4	16	4
55	5	58	6	61	2	-1	4	3	[illegible]	16	9
75	1	73	1	68	1	0	0	0	0	0	0
45	7	49	8.5	40	9	-1.5	-0.5	-2	2.25	0.25	4
64	3	60	4	58	5	-1	-1	-2	1	1	4
45	7	49	8.5	60	3.5	-1.5	5	3.5	2.25	25	12.25
72	2	60	4	60	7.5	-2	-3.5	5.5	4	12.25	30.25
39	9	55	7	38	10	2	-3	-1	4	9	1
35	10	60	4	50	7.5	6	-3.5	2.5	36	12.25	6.25
45	7	48	10	60	3.5	-3	6.5	3.5	9	42.25	12.25

$$\rho_{xy} = 1 - \frac{6 \Sigma (x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 63.5}{990}$$

$$= 1 - 0.385$$

$= 0.615$

$$\rho_{yz} = 1 - \frac{6\Sigma(y-z)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 134}{990}$$

$$= 1 - 0.810$$

$$= 0.188$$

$$\rho_{zx} = 1 - \frac{6\Sigma(z-x)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 83}{990}$$

$$= 1 - 0.503$$

$$= 0.497$$

(7) A computer while calculating the correlation co-efficient between two variables x and y obtained the following constants $n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$,

$\Sigma y = 100$, $\Sigma y^2 = 460$ & $\Sigma xy = 508$. It was later discovered at the time of checking that he had copied down 2 pairs of observations (x_i, y_i) as $(6,14)$ & $(8,6)$ instead of the correct values $(8,12)$ & $(6,8)$. Obtained the correct value of the correlation co-efficient between $(x \& y)$

Soln:

$\Sigma x = 125$, $\Sigma x^2 = 650$

$\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 508$

(x_i, y_i) as $(6,14)$ & $(8,6)$ - wrong

$(8,12)$ & $(6,8)$ - ~~right~~ correct

corrected

$\Sigma x = 125 - \cancel{6} - \cancel{8} + \cancel{8} + \cancel{6}$

$= 125$

$\Sigma x^2 = 650 - \cancel{6^2} - \cancel{8^2} + \cancel{8^2} + \cancel{6^2}$

$= 650$

$\Sigma y = 100 - 14 - 6 + 12 + 8$

$= 100$

$\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2$

$= 460$

The corrected values are

$\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 520$

$\Sigma xy = 508 - (6\times 12) - (8\times 6) + (8\times 12) + (6\times 6)$

$= 520$

Here $n = 25$

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\left[n\Sigma x_i^2 - (\Sigma x_i)^2\right]^{1/2}\left[n\Sigma y_i^2 - (\Sigma y_i)^2\right]^{1/2}}$$

$$= \frac{25\times 520 - 125\times 100}{\left[(25\times 650) - (125)^2\right]^{1/2}\left[(25\times 460) - (100)^2\right]^{1/2}}$$

Stu

achi

possi

of

Labo

Lectu

$$= \frac{13000 - 12500}{\left[(16250)^{1/2} - (125)^2\right]\left[(11500)^{1/2} - (100)^2\right]}$$

$$= \frac{13000 - 12500}{(625)^{1/2}\,(900)^{1/2}}$$

$$= \frac{500}{25 \times 30}$$

$$= \frac{500}{750}$$

$$r_{xy} = 0.66$$

1) The following table shows how 10 students were ranked according to their achievements in the laboratory and lecture portions of biology course find the coefficients of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

Laboratory (x)	~~Rank in~~	Lecture (y)	~~Rank in~~	$x-y$	$(x-y)^2$
8	~~8~~	9		-1	1
3	~~8~~	5		-2	4
9	~~2~~	10		-1	1
2	~~9~~	1		1	1
7	~~4~~	8		-1	1
10	~~1~~	7		3	9
4	~~7~~	3		1	1
6	~~5~~	4		2	4
1	~~10~~	2		-1	1
5	~~6~~	6		-1	1

$$\Sigma(x-y)^2 = 24$$

$$\rho = 1 - \frac{6\,\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 24}{10(10^2-1)}$$

$$= 1 - \frac{6 \times 24}{10(100-1)}$$

$$= 1 - \frac{6 \times 24}{10(99)}$$

$$= 1 - \frac{144}{990}$$

$$= \frac{990 - 144}{990}$$

$$= 0.855$$

2) 10 students got the following % of marks in 2 subjects Economics & Statistics

Economics 78 65 36 98 25 75 82 90 62 39

Statistics 84 53 51 91 60 68 62 86 58 47

Calculate the rank correlation co-efficient

Economics	Rank in x	Statistics	Rank in y	$x-y$	$(x-y)^2$
78	4	84	3	1	1
65	6	53	8	-2	4
36	9	51	9	0	0

98	1	91	1	0	0
25	10	60	6	4	16
75	5	68	4	1	1
82	3	62	5	-2	4
90	2	86	2	0	0
62	7	58	7	0	0
39	8	47	10	-2	4

$\Sigma(x-y)^2 = 30$

$$\rho = 1 - \frac{6\,\Sigma(x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(10^2-1)}$$

$$= 1 - \frac{6 \times 30}{10(99)}$$

$$= 1 - \frac{6 \times 30}{990}$$

$$= 1 - \frac{180}{990}$$

$$= \frac{990 - 180}{990}$$

$$= 0.818$$

0
16
1
4
0
0
4
$(x-y)^2 = 30$

Regression

If there is a fuctional relationship between the variables x_i & y_i the points in the scater diagram will cluster around some curve called the curve of regression If a curve is a straight line it is called a line of regression between the two variables

If we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to the y axis (x axis) from the points to the line to minimished we obtain a line of best fit for the data and it is called the regression line of y on x (x on y)

Theorem :- 1

The equation of the regression line of y on x is given by $y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x})$

Let $y = ax + b$ be the regression line of y on x

$$y_i = ax_i + b$$

$$y_i - ax_i - b = 0$$

$$(y_i - ax_i - b)^2 = 0$$

$$\Sigma (y_i - ax_i - b)^2 = 0$$

Let $S = \Sigma (y_i - ax_i - b)^2$

According to the principle of least Squares we have to determine the parameters a and b so that So minimum

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

$$\Rightarrow 2\Sigma(y_i - ax_i - b)(-x_i) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b)(x_i) = 0$$

$$\Rightarrow \Sigma(x_i y_i - ax_i^2 - bx_i) = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma x_i^2 - b\Sigma x_i = 0$$

$$\Rightarrow a\Sigma x_i^2 + b\Sigma x_i = \Sigma x_i y_i \rightarrow (1)$$

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

$$\Rightarrow 2\Sigma(y_i - ax_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma(y_i - ax_i - b) = 0$$

$$\Rightarrow \Sigma y_i - a\Sigma x_i - nb = 0$$

$$\Rightarrow a\Sigma x_i + nb = \Sigma y_i \rightarrow (2)$$

Equation (1) & (2) are called the normal equation

÷ing ② by n

② ⇒ $a\frac{\Sigma x_i}{n} + \frac{nb}{n} = \frac{\Sigma y_i}{n}$

$a\bar{x} + b = \bar{y}$

the regression of line passes

through $(\bar{x}, \bar{y})$

Now shifting the origin to this point

$(\bar{x}, \bar{y})$ by giving the transformation

$X_i = x_i - \bar{x}, \; Y_i = y_i - \bar{y}$

$X_i = x_i - \bar{x}$

$\Sigma X_i = \Sigma(x_i - \bar{x})$

$= \Sigma x_i - \Sigma\bar{x}$

$= n\bar{x} - n\bar{x}$

$= 0$

///ly

$\Sigma Y_i = 0$

(2) $\Rightarrow a\sum x_i + nb = \sum y_i$

$\Rightarrow a \times 0 + n \times b = 0$

$\Rightarrow nb = 0$

Here $n \neq 0$, $b = 0$

Hence the line of regression becomes

$$Y = aX \longrightarrow (3)$$

(1) $\Rightarrow a\sum x_i^2 + b\sum x_i = \sum x_i y_i$

$\Rightarrow a\sum x_i^2 + 0\sum x_i = \sum x_i y_i$

$a\sum x_i^2 = \sum x_i y_i$

$$a = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \frac{r \cancel{n} \cancel{\sigma_x} \sigma_y}{\cancel{n} \sigma_x^{\cancel{2}}}$$

$$a = r\frac{\sigma_y}{\sigma_x}$$

(3) $\Rightarrow Y = aX$

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

Which is the Regression line of y on x

Hence proved

Theorem :- ②

The equation of regression line of x on y is given by $(x-\bar{x}) = r\frac{\sigma x}{\sigma y}(y-\bar{y})$

Let $x = ay+b$ be the regression line of x on y

$$x_i = ay_i + b$$

$$x_i - ay_i - b = 0$$

$$(x_i - ay_i - b)^2 = 0$$

$$\Sigma(x_i - ay_i - b)^2 = 0$$

Let $S = \Sigma(x_i - ay_i - b)^2$

According to the principle of least squares we have to determine the parameters a and b so that S is minimum

$$\frac{\partial s}{\partial a} = 0$$

$$\Rightarrow 2\Sigma(x_i - ay_i - b)(-y_i) = 0$$

$$\Rightarrow -2\Sigma(x_i - ay_i - b)y_i = 0$$

$$\Rightarrow \Sigma(x_i - ay_i - b)y_i = 0$$

$$\Rightarrow \Sigma x_i y_i - a\Sigma y_i^2 - b\Sigma y_i = 0$$

$$\Rightarrow a\Sigma y_i^2 + b\Sigma y_i = \Sigma x_i y_i \longrightarrow (1)$$

$$\frac{\partial s}{\partial b} = 0$$

$$\Rightarrow 2\Sigma(x_i - ay_i - b)(-1) = 0$$

$$\Rightarrow -2\Sigma(x_i - ay_i - b) = 0$$

$$\Rightarrow \Sigma(x_i - ay_i - b) = 0$$

$$\Rightarrow \Sigma x_i - a\Sigma y_i - nb = 0 \longrightarrow (2)$$

$$a\Sigma y_i + nb = \Sigma x_i \longrightarrow (2)$$

Equation (1) & (2) are called the normal equation

$\div$ origin by n

(2) $\Rightarrow a\frac{\sum y_i}{n} + \frac{nb}{n} = \frac{\sum x_i}{n}$

$\Rightarrow \quad a\bar{y} + b = \bar{x}$

The regression of line passes through $(\bar{y}, \bar{x})$

Now shifting the origin to this point $(\bar{y}, \bar{x})$ by giving the transformation

$X_i = x_i - \bar{x}, \quad Y_i = y_i - \bar{y}$

$X_i = x_i - \bar{x}$

$\sum X_i = \sum (x_i - \bar{x})$

$= \sum x_i - \sum \bar{x}$

$= n\bar{x} - n\bar{x}$

$= 0$

III^{ly} $\sum Y_i = 0$

(2) $\Rightarrow a\Sigma y_i + nb = \Sigma x_i$

$\Rightarrow a \times 0 + nb = 0$

$\Rightarrow nb = 0$

Here $n \neq 0$, $b = 0$

Hence the line of Regression becomes

$$x = ay \rightarrow (3)$$

(1) $\Rightarrow a\Sigma y_i^2 + b\Sigma y_i = \Sigma x_i y_i$

$\Rightarrow a\Sigma y_i^2 + 0\,\Sigma y_i = \Sigma x_i y_i$

$\Rightarrow a\Sigma y_i^2 = \Sigma x_i y_i$

$$a = \frac{\Sigma x_i y_i}{\Sigma y_i^2}$$

$$a = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2}$$

$$a = \frac{r\cancel{n}\,\sigma x\,\sigma y}{\cancel{n}\,\sigma y^2}$$

$$a = \frac{r\,\sigma x}{\sigma y}$$

(2) $x = ay$

$$x - \bar{x} = r\frac{\sigma x}{\sigma y}(y - \bar{y})$$

Hence proved

Note :- $\bar{x}, \bar{y}$ is the point of intersection of a 2 regression line

The slope of the regression line of y on x is called the regression co-efficient of y on x an it is denoted by b_{yx}.

Hence $b_{yx} = r\frac{\sigma y}{\sigma x}$

$|||^{ly}$

The regression co-efficient of x on y is given by

$$b_{xy} = r\frac{\sigma x}{\sigma y}$$

Theorem: 3 :-

correlation co-efficient is the geometric mean between the regression co-efficients

(ie) $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$

Proof:-

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_{yx} \cdot b_{xy} = r\frac{\cancel{\sigma_y}}{\cancel{\sigma_x}} \cdot r\frac{\cancel{\sigma_x}}{\cancel{\sigma_y}}$$

$$b_{yx} \cdot b_{xy} = r^2$$

$$r^2 = b_{yx} \cdot b_{xy}$$

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

Hence Proved

The sign of the correlation co-efficient they same as the regression co-efficient

Theorem: 4

If one of the the regression co-efficient is greater than unity, the other is less than unity.

We have $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$b_{yx} \cdot b_{xy} = r\frac{\cancel{\sigma_y}}{\cancel{\sigma_x}} \cdot r\frac{\cancel{\sigma_x}}{\cancel{\sigma_y}}$$

$$= r^2$$

$$= r^2 \leq 1$$

$$b_{yx} \cdot b_{xy} \leq 1$$

If $b_{yx} > 1$ then $b_{xy} < 1$

If $b_{xy} > 1$ then $b_{yx} < 1$

Theorem: 5)

Arithmetic mean of the regression co-efficient is greater than (or) equal to the correlation co-efficient.

Let b_{yx} & b_{xy} be the correlation co-efficient

(ii). P $\frac{b_{yx}+b_{xy}}{2} \ge r$

$b_{yx}+b_{xy} \ge 2r$

$r\frac{\sigma y}{\sigma x} + r\frac{\sigma x}{\sigma y} \ge 2r$

$\cancel{r}\left(\frac{\sigma y}{\sigma x} + \frac{\sigma x}{\sigma y}\right) \ge 2\cancel{r}$

$\frac{\sigma y}{\sigma x} + \frac{\sigma x}{\sigma y} \ge 2$

$\frac{\sigma y^2 + \sigma x^2}{\sigma x\, \sigma y} \ge 2$

$\sigma y^2 + \sigma x^2 \ge 2\sigma x\, \sigma y$

$\sigma y^2 + \sigma x^2 - 2\sigma x\sigma y \ge 0$

$(\sigma x^2 - \sigma y^2)^2 \ge 0$

This condition is always true.

Theorem 6

Regression co-efficient are indipenend of the change of origin but dependent on the change of the ~~sca~~ Scale.

Let $u_i = \dfrac{x_i - A}{h}$, $v_i = \dfrac{y_i - B}{k}$

$$h u_i = x_i - A \qquad\qquad k v_i = y_i - B$$

$$x_i = h u_i + A \qquad\qquad y_i = k v_i + B$$

$$\bar{x} = h\bar{u} + A \qquad\qquad \bar{y} = k\bar{v} + B$$

$$x_i - \bar{x} = h u_i + \cancel{A} - h\bar{u} - \cancel{A} \qquad\qquad (y_i - \bar{y}) = k(v_i - \bar{v})$$

$$= h u_i - h\bar{u}$$

$$= h(u_i - \bar{u})$$

$$(x_i - \bar{x})^2 = h^2 (u_i - \bar{u})^2$$

$$\cancel{(y_i - \bar{y})^2 = k^2 (v_i - \bar{v})^2}$$

$$\frac{(x_i - \bar{x})^2}{n} = \frac{h^2 (u_i - \bar{u})^2}{n}$$

$$\frac{\Sigma (x_i - \bar{x})^2}{n} = h^2 \frac{\Sigma (u_i - \bar{u})^2}{n}$$

$$\sigma x^2 = h^2 \sigma u^2$$
$$\sigma x = h \sigma u$$
lly
$$\sigma y^2 = k^2 \sigma v^2$$
$$\sigma y = k \sigma v$$

$$r_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma x \sigma y}$$

$$= \frac{\Sigma \cancel{h}(u_i - \bar{u})\cancel{k}(v_i - \bar{v})}{n \cancel{h} \sigma u \cancel{k} \sigma v}$$

$$= \frac{\Sigma (u_i - \bar{u})(v_i - \bar{v})}{n \sigma u \sigma v}$$

$$r_{xy} = r_{uv}$$

$$b_{xy} = r \frac{\sigma x}{\sigma y}$$
$$= r \frac{h(\sigma u)}{k(\sigma v)}$$
$$= \frac{h}{k} b_{uv}$$

$$b_{yx} = r \frac{\sigma y}{\sigma x}$$
$$= r \left(\frac{k}{h} \frac{\sigma v}{\sigma u}\right) = \frac{k}{h} b_{vu}$$

Hence the regression co-efficient are independent of origin A & B but dependent of the scale h & k

Hence proved

1) The following data relate to the marks of 10 students in the internal test and university Examination for the maximum of 50 each

internal	25	28	30	32	35	36	38	39	42	[illegible]
uni. marks	20	26	29	30	25	18	26	35	35	[illegible]

(i) Obtain the two regression equation & determine

(ii) The most likely internal mark for the university mark of 25

(iii) The most likely for the internal mark of 30

Let x be the internal mark & y the university mark

now ~~xcc~~ ~~25×20+28×26~~

$$\bar{x} = \frac{25+28+30+32+35+36+38+39+42+45}{10}$$

$$= 35$$

$$\bar{y} = \frac{20+26+29+30+25+18+26+35+35+46}{10}$$

$$= 29$$

x	y	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
25	20	-10	-9	100	81	90
28	26	-7	-3	49	9	21
30	29	-5	0	25	0	0
32	30	-3	1	9	1	-8
35	25	0	-4	0	16	0
36	18	1	-11	1	121	-11
38	26	3	-3	9	9	-9
39	35	4	6	16	36	24
42	35	7	6	49	36	42
45	46	10	17	100	289	70

$\Sigma(x_i-\bar{x})$ $= 0$ $\Sigma(y_i-\bar{y})$ $= 0$ $\Sigma(x_i-\bar{x})^2 = 358$ $\Sigma(y_i-\bar{y})^2 = 598$ $\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 324$

Regr

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 324$

$\Sigma(x_i-\bar{x})^2 = 358$

$\Sigma(y_i-\bar{y})^2 = 598$

$$\sigma x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n}$$

$$= \frac{358}{10}$$

$$= 35.8$$

$\sigma x = 5.98$

$$\sigma y^2 = \frac{\Sigma(y_i-\bar{y})^2}{n}$$

$$= \frac{598}{10}$$

$$= 5.98$$

$\sigma y = 7.733$

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma x\,\sigma y}$$

$$= \frac{324}{10 \times 5.98 \times 7.733}$$

R

$= 0.7$ (app)

Regression line of y on x

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

$$y-\bar{y} = r\frac{\sigma y}{\sigma x}(x-\bar{x})$$

$$y-29 = 0.7 \times \frac{7.733}{5.98}(x-35)$$

$$y-29 = \frac{5.4131}{5.98}(x-35)$$

$$y-29 = 0.905(x-35$$

$$y-29 = 0.905x - 31.675$$

$$y = 0.905x - 31.675 + 29$$

$$y = 0.905x - 2.675$$

Regression line of x on y

$$x-\bar{x} = r\frac{\sigma x}{\sigma y}(y-\bar{y})$$

$$x-35 = 0.7 \times \frac{5.98}{7.733}(y-29)$$

$x - 35 = 0.5413(y - 29)$

$x - 35 = 0.5413y - 15.697$

$x = 0.5413y - 19.303 \rightarrow (2)$

when $y = 25$, $x = ?$

Sub in (2)

$x = 0.54 \times 25 - 19.34$

$= 32.84$

The most likely internal mark for u.m 25 is 32.84

when $x = 30$, $y = ?$

Sub in (3)

$y = 0.9 \times 30 - 2.5$

$y = 24.5$

The most likely u.marks for internal mark 35 is 24.5

Students obtain the following in the college internal test x and in the final uni. marks y

x	51	63	63	49	50	60	65	63	46	50
y	49	72	75	50	48	60	70	48	60	56

Estimate the uni. marks of a student who get 61 in the internal test

$x \quad y \quad x-\bar{x} \quad y-\bar{y}$

$\bar{x} = \frac{\sum x_i}{n}$

$= \frac{51+63+63+49+50+60+65+63+46+50}{10}$

$= 56$

$\bar{y} = \frac{\sum y_i}{n}$

$= \frac{49+72+75+50+48+60+70+48+60+56}{10}$

$= 58.8$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
51	49	-5	-9.8	25	96.04	49
62	72	7	12.2	49	174.2	92.4
63	75	7	16.2	49	262.4	113.4
49	50	-7	-8.8	49	-14.44	61.6
50	48	-6	-10.8	36	116.6	64.8
60	60	4	1.2	16	1.44	4.8
65	70	9	11.2	81	125.4	100.8
63	48	7	-10.8	49	116.6	-75.6
46	60	-10	1.2	100	1.44	-12
50	56	-6	-2.8	36	7.84	16.8
				$\Sigma(x_i-\bar{x})^2$ $=490$	$\Sigma(y_i-\bar{y})^2$ $=999.6$	$\Sigma(x_i-\bar{x})(y_i-\bar{y})$ $=416$

$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = 416$

$\Sigma(x_i-\bar{x})^2 = 490$

$\Sigma(y_i-\bar{y})^2 = 999.6$

$$\sigma_x^2 = \frac{\Sigma(x_i-\bar{x})^2}{n}$$

$$= \frac{490}{10}$$

$$= 49$$

$\sigma x = 7$

$$\sigma y^2 = \frac{\Sigma (y_i - \bar{y})^2}{n}$$

$$= \frac{979.6}{10}$$

$$= 97.96$$

$\sigma y = 9.9$

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n \sigma x \sigma y}$$

$$= \frac{416}{10 \times 9.9 \times 7}$$

$$= \frac{416}{693}$$

$$= 0.60$$

The regression line of y on x is

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$y - 58.8 = 0.60 \times \frac{9.90}{7} (x - 56)$$

$$y - 58.8 = 0.85 (x - 56)$$

$y - 58.8 = 0.85x - 47.6$

$y = 0.85x - 47.6 + 58.8$

$y = 0.85x + 11.2$

when $x = 61$, $y = ?$

$y = 0.85x + 11.2$

$= 0.85 \times 61 + 11.2$

$= 51.85 + 11.2$

$\boxed{y = 63.05}$

2) Out of the two lines of regression $\underbrace{x+2y-5=0}_{x}$ & $\underbrace{2x+3y-8=0}_{y}$ which one is the regression line of x on y

b_{xy}

b_{yx}

$y^2 = b_{xy}\ b_{yx}$

The given two regression lines are

$x+2y-5=0$, $2x+3y-8=0$

$x = -2y+5$

$x = -2y+5$

$3y = -2x+8$

$y = \frac{-2}{3}x + 8/3$

Suppose the regression line of x on y

$x = -2y+5$

Here $b_{xy} = -2$

The regression line of y on x

$y = -2/3\,x + 8/3$

Here $b_{yx} = -2/3$

We have $r^2 = b_{xy} \cdot b_{yx}$

$= -2 \cdot -2/3$

$= 4/3$

$r^2 = 1.33 > 1$

This is not possible.

∴ Our assumption is wrong

Hence regression line of x on y is

$2x+3y-8=0$

1) The 2 variables x and y for the regression line $3x+2y-26=0$ & $6x+y-31=0$

find (i) The mean values of x & y

(ii) Prove that the correlation co-efficient between x & y

(iii) The variens of y if the variens of x is 25.

(i) Since the two lines passes through $(\bar{x}, \bar{y})$

$$3\bar{x}+2\bar{y}=26 \rightarrow (1)$$

$$6\bar{x}+\bar{y}=31 \rightarrow (2)$$

$(1)\times 2 \Rightarrow \quad \cancel{6\bar{x}}+4\bar{y}=52$

$(2) \Rightarrow \quad \cancel{6\bar{x}}+\bar{y}=31$

$\quad (-) \quad (-) \quad (-)$

$$3\bar{y}=21$$

$\boxed{\bar{y} = 7}$

Sub $\bar{y} = 7$ in ①

$$3\bar{x} + 2 \times 7 = 26$$

$$3\bar{x} = 26 - 14$$

$$3\bar{x} = 12$$

$$\boxed{\bar{x} = 4}$$

(ii) suppose $3x + 2y - 26 = 0$ is the regression line of x on y

$$3x = -2y + 26$$

$$x = \frac{-2}{3} y + \frac{26}{3}$$

$$\boxed{bxy = -2/3}$$

$6x + y - 31$ is the Regression line of y on x

~~6x = -y + 31~~ $y = -6x + 31$

$$\boxed{byx = -6}$$

$$r^2 = byx \cdot bxy$$

$$= -\cancel{6}^{2} \times -2/\cancel{3}$$

$= 4 > 1$

$\therefore$ our assumption is wrong

$3x + 2y - 26 = 0$ is the regression line of y on x

$2y = -3x + 26$

$y = \frac{-3}{2}x + \frac{26}{2}$

$\boxed{b_{yx} = -3/2}$

$6x + y - 31 = 0$ is the regression line of x on y

$6x = -y + 31$

$x = \frac{-y}{6} + \frac{31}{6}$

$\boxed{b_{xy} = -1/6}$

$\gamma^2 = b_{xy} \cdot b_{yx}$

$= -1/6 \times -3/2$

$= 1/4$

$r = \pm 0.5$

$r = -0.5$

...c of y on x

(iii) Varience of $x = 25$

$\sigma x^2 = 25$

$\sigma x = 5$

To find σy

we have,

$b_{xy} = r \cdot \frac{\sigma x}{\sigma y}$

$\frac{-1}{6} = -0.5 \cdot \frac{5}{\sigma y}$

...of x on y

$+\frac{\sigma y}{6} = +2.5$

$\sigma y = 2.5 \times 6$

$\sigma y = 15$

$\sigma y^2 = 225$

2) If $x = 4y+5$ & $y = kx+4$ are the regression line of x on y & y on x respectively (i) show that $0 \le k \le 1/4$

(ii) If $k = 1/6$ find the means of the two variables x & y and the correlation co-efficient between them

(i) The regression line of x on y is

$$x = 4y+5$$

$$\boxed{b_{xy} = 4}$$

Regression line of y on x is

$$y = kx+4$$

$$\boxed{b_{yx} = k}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= 4k$$

$$r^2 = 4k$$

we have,

$$0 \leq r^2 \leq 1$$

$$0 \leq 4k \leq 1$$

$$0 \leq k \leq 1/4$$

∴ Hence proved

(ii) If $k = 1/8$

$$r^2 = 4k$$

$$= 4 \times 1/8$$

$$r^2 = 1/2$$

$$r = \pm\sqrt{\frac{1}{2}}$$

$$r = \pm 0.707$$

$r = +0.707$ [∵ b_{yx} & b_{xy} are positive]

$$x = 4y+5$$

$$x - 4y - 5 = 0 \rightarrow (1)$$

$$y = kx+4$$

$$y = \frac{1}{8}x+4$$

$$y = \frac{x+32}{8}$$

$8y = x+4$

$x - 8y + \cancel{32} = 0 \rightarrow (b)$

The two regression lines passes through $(\bar{x}, \bar{y})$

$\bar{x} - 4\bar{y} - 5 = 0 \rightarrow (a)$

$\bar{x} - 8\bar{y} + 32 = 0 \rightarrow (b)$

$(-) \quad (+) \quad (-)$

$4\bar{y} - 37 = 0$

$4\bar{y} = 37$

$\bar{y} = 37/4$

$\boxed{\bar{y} = 9.25}$

sub $\bar{y} = 9.25$ in (a)

$\bar{x} - 4 \times 9.25 - 5 = 0$

$\bar{x} - 37 - 5 = 0$

$\bar{x} - 42 = 0$

$\boxed{\bar{x} = 42}$

9) The variable x & y are connected by the equation $ax+by+c=0$. Show that $r_{xy} = -1$ (or) 1 according a & b are of the same sign or of opposite sign.

Writing $ax+by+c=0$ is of the form

$$ax = -by - c$$

$$x = \frac{-b}{a}y - c/a$$

$$\boxed{b_{xy} = -b/a}$$

Writing $ax+by+c=0$ in the form

$$by = -ax - c$$

$$y = -\frac{a}{b}x - c/b$$

$$\boxed{b_{yx} = -a/b}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= \frac{-b}{a} \times \frac{-a}{b}$$

$$= 1$$

$r^2 = 1$

$r = \pm 1$

Suppose a & b are of same sign then

$r^2 = 1$

Hence $r = -1$ [$\because$ b_{xy} & b_{yx} are Negative]

Suppose a & b are of opposite sign

then $r^2 = 1$

Hence $r = 1$ [$\because$ b_{xy} & b_{yx} are positive]

4) The following table shows the ages x & blood pressures y are given of 12 women

(i) find the correlation co-efficient between x & y

(ii) determine the regression equation of y on x

(iii) estimate the blood pressure of a women whose age is 45

Age (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood Pressure (y)	147	125	160	118	149	128	150	145	115	140	152	155

5) The equations of two regression lines obtained in a correlation analysis are $4x - 5y + 33 = 0$ & $20x - 9y - 107 = 0$. If the variance $y = 16$, find

(i) The mean values of x & y

(ii) The correlation coefficient between x & y

(iii) Standard deviation of x.

5) (i) Since the two lines passes through $\bar{x}, \bar{y}$

$$4\bar{x} - 5\bar{y} = -33 \rightarrow ①$$

$$20\bar{x} - 9\bar{y} = 107 \rightarrow ②$$

$$① \times 5 \Rightarrow 20\bar{x} - 25\bar{y} = -165$$

$$② \Rightarrow \underline{\cancel{20\bar{x}} - 9\bar{y} = 107}$$

$$-16\bar{y} = -272$$

$16\bar{y} = 272$

$\bar{y} = \frac{272}{16}$

$\boxed{\bar{y} = 17}$

Sub $\bar{y} = 17$ in ①

$4\bar{x} - 5(17) = -33$

$4\bar{x} - 85 = -33$

$4\bar{x} = -33 + 85$

$4\bar{x} = 52$

$\bar{x} = 52/4$

$\boxed{\bar{x} = 13}$

(ii) suppose $4x - 5y + 33 = 0$ is the regression line of x on y

$4x = 5y - 33$

$x = \frac{5y}{4} - \frac{33}{4}$

$\boxed{b_{xy} = 5/4}$

suppose $20\bar{x} - 9\bar{y} = 107$ is the regression line of y on x

$$-9\bar{y} = -20\bar{x} + 107$$

$$\bar{y} = \frac{+20}{+9}\bar{x} - \frac{107}{9}$$

$$\bar{y} = \frac{20}{9}\bar{x} - \frac{107}{9}$$

$$\boxed{b_{yx} = 20/9}$$

Now, $r^2 = b_{xy} \cdot b_{yx}$

$$= \frac{5}{\cancel{4}} \times \frac{\cancel{20}^{5}}{9}$$

$$r^2 = 25/9 \qquad r^2 = 2.78 > 1$$

~~$r = \pm 5/3$~~

our assumption is wrong

$4x - 5y + 33 = 0$ is ~~$\boxed{r = 5/3}$~~ the regression line of y on x

$$-5y = -4x - 33$$

$$5y = 4x + 33$$

$$y = 4/5\, x + 33/5$$

$$\boxed{b_{yx} = 4/5}$$

$20x - 9y - 107 = 0$ is the regression line of x on y

$$20x = 9y + 107$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

$$\boxed{b_{xy} = \frac{9}{20}}$$

$$r^2 = b_{yx} \cdot b_{xy}$$

$$= \frac{4}{5} \cdot \frac{9}{20}$$

$$r^2 = \frac{9}{25} = 0.36$$

$$r = \pm 0.6$$

$$r = 0.6$$

(iii) Given varience of $y = 16$

$$\sigma_y^2 = 16$$

$$\sigma_y = 4$$

To find the varience of x

$$\sigma_x^2 = ?$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

$$\frac{4}{5} = 0.6 \times \frac{4}{\sigma_x}$$

$$\frac{4}{5} = \frac{2.4}{\sigma_x}$$

$$4\sigma_x = 2.4 \times 5$$

$$4\sigma_x = 12$$

$$\sigma_x = \frac{12}{4}$$

$$\sigma_x = 3$$

$$\sigma_x^2 = 9$$

standard deviation of x $\boxed{\sigma_x = 3}$

reg x on y

deviation

$\boxed{=3}$

(a) (i) Let x be the Age

and y be the blood pressure

Now $\bar{x} = \frac{\Sigma x_i}{n}$

$$= \frac{56+42+72+36+63+47+55+49+38+42+68+60}{12}$$

$$= \frac{628}{12}$$

$$\bar{x} = 52.33$$

$$\bar{y} = \frac{\Sigma y_i}{n}$$

$$= \frac{147+125+160+118+149+128+150+145+115+140+152+155}{12}$$

$$= \frac{1684}{12} = 140.33$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
56	147	3.67	6.67	24.48	13.47	44.49
42	125	-10.33	-15.33	158.36	106.71	235.01
72	160	19.67	19.67	386.91	386.91	386.91
36	118	-16.33	-22.33	364.65	266.67	498.63
63	149	10.67	8.67	92.51	113.85	75.63
47	128	-5.33	-12.33	65.72	28.41	152.03
55	150	2.67	9.67	25.82	7.13	93.51
49	145	-3.33	4.67	-15.55	11.09	21.81
38	115	-14.33	-25.33	362.98	205.35	641.61
42	140	-10.33	-0.33	3.41	106.71	0.11
68	152	15.67	11.67	182.87	245.55	136.19
60	152	7.67	14.67	112.52	58.83	215.21
				$\Sigma(x-\bar{x})(y-\bar{y})$ $= 1764.68$	$\Sigma(x-\bar{x})^2$ $= 1550.68$	$\Sigma(y-\bar{y})^2$ $= 2800.68$

$$\sigma x^2 = \frac{\Sigma(x-\bar{x})^2}{n}$$

$$= \frac{1550.68}{12}$$

$$\sigma x^2 = 129.2$$

$$\sigma x = 11.37$$

$(y-\bar{y})^2$

4.49

55.01

386.91

98.63

5.63

2.03

3.51

.81

1.61

.11

38.19

15.21

24.57

2800.68

$$\sigma y^2 = \frac{\Sigma(y-\bar{y})^2}{n}$$

$$= \frac{2800.68}{12}$$

$$= 14.44$$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma x \sigma y}$$

$$= \frac{1764.68}{12 \times 11.37 \times 14.44}$$

$$= \frac{1764.68}{1970.19}$$

$$r = 0.90$$

(ii) The regression line of y on x

$$y - \bar{y} = r\frac{\sigma y}{\sigma x}(x-\bar{x})$$

$$y - 140.33 = 0.9\left(\frac{14.44}{11.37}\right)(x - 52.33)$$

$$y - 140.33 = 0.9(1.27)(x - 52.33)$$

$$y - 140.33 = 1.146x - 59.66$$

$$y = 1.146x - 59.66 + 140.33$$

$$y = 1.14x + 80.67$$

(iii) When $x = 45$, $y = ?$

$$y = 1.14(45) + 80.67$$

$$= 51.3 + 80.67$$

$$= 131.97$$

Theorem :-

The angle between ~~the~~ to regression line is given by $\theta = \tan^{-1}\left[\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma x \, \sigma y}{\sigma x^2 + \sigma y^2}\right)\right]$

Regression line of x on y is

$$x - \bar{x} = r\frac{\sigma x}{\sigma y}(y - \bar{y})$$

$$x - \bar{x} = r\frac{\sigma x}{\sigma y}y - r\frac{\sigma x}{\sigma y}\bar{y}$$

$$r\frac{\sigma x}{\sigma y} \cdot y = x - \bar{x} + r\frac{\sigma x}{\sigma y}\bar{y}$$

$$y = \frac{\sigma y}{r\sigma x}\left[x - \bar{x} + r\frac{\sigma x}{\sigma y}\bar{y}\right]$$

$$y = \frac{\sigma y}{r \sigma x} x - \frac{\sigma y}{r \sigma x}\left(\bar{x} - r \frac{\sigma x}{\sigma y} \bar{y}\right)$$

$$\boxed{m_2 = \frac{\sigma y}{r \sigma x}}$$

Regression line of y on x

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

1

$$y = r \frac{\sigma y}{\sigma x} x - r \frac{\sigma y}{\sigma x} \bar{x} + \bar{y}$$

$$\boxed{m_1 = r \frac{\sigma y}{\sigma x}}$$

Let θ be the obtuse angle between two regression lines

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$$

$$= \frac{r \frac{\sigma y}{\sigma x} - \frac{\sigma y}{r \sigma x}}{1 + r \frac{\sigma y}{\sigma x} \times \frac{\sigma y}{r \sigma x}}$$

$$= \frac{r\, \sigma y / \sigma x - \frac{\sigma y}{r \sigma x}}{1 + \frac{\sigma y^2}{\sigma x^2}}$$

$$= \frac{\dfrac{\gamma^2 \delta y - \delta y}{\gamma \delta x}}{\boxed{\dfrac{\delta x^2 + \delta y^2}{\delta x^2}}}$$

$$\gamma^2 = \frac{\delta y}{\delta x}$$

$$= \frac{\dfrac{\gamma^2 \delta y - \delta y}{\gamma \delta x}}{\dfrac{\delta x^2 + \delta y^2}{\delta x^2}}$$

$$= \frac{\dfrac{\delta y(\gamma^2 - 1)}{\gamma \delta x}}{\dfrac{\delta x^2 + \delta y^2}{\delta x^2}}$$

$$= \frac{\delta y}{\gamma \delta x}(\gamma^2 - 1) \times \frac{\delta x^2}{\delta x^2 + \delta y^2}$$

$$= \frac{\delta y(\gamma^2 - 1)\delta x}{\gamma(\delta x^2 + \delta y^2)}$$

$$= \frac{\sigma x \sigma y (r^2 - 1)}{r(\sigma x^2 + \sigma y^2)}$$

$$\tan\theta = \left(\frac{r^2-1}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)$$

$$\theta = \tan^{-1}\left[\left(\frac{r^2-1}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)\right]$$

Hence Proved

Note:- 1

The accute angle between the regression line is given by $\theta = \tan^{-1}\left[\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)\right]$

Note: 2

If $r = 0$ the $\theta = \tan^{-1}(\infty) =$

$= \pi/2$

Thus If the two Variables are uncorrelated then the lines of regression are perpenticular to each other.

Note: 3.

If $r = \pm 1$ then $\theta = \tan^{-1}(0)$

$\theta = 0$ (or) π

$\therefore$ The two regression lines are parallel

The two lines have the common point $(\bar{x}, \bar{y})$ Then the two line must be co-incident $\therefore$ If their is a perfect correlation (Positive or Negative) between the 2 variables then the two lines of regression co-inside

1) If θ is a accute angle between the two regression line show that $\sin\theta \leq 1 - r^2$

We have if θ is a accute angle between the two regression line

then $\theta = \tan^{-1}\left[\left(\frac{1 - r^2}{r}\right) \cdot \left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}\right)\right]$

We assume that $\sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y \rightarrow (1)$

Suppose if it is not true

$$\sigma_x^2 + \sigma_y^2 < 2\sigma_x\sigma_y$$

$$\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y < 0$$

$$(\sigma_x - \sigma_y)^2 < 0$$

So this is impossible $\left[\because (\sigma_x - \sigma_y)^2 > 0\right]$

Our assumption is wrong

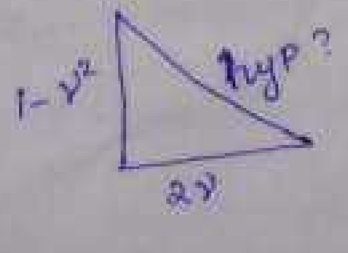
$$\sigma_x^2 + \sigma_y^2 \geq 2\sigma_x\sigma_y$$

$$\frac{1}{2} \geq \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$$

$$\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \leq \frac{1}{2}$$

$$\tan\theta \leq \frac{1-r^2}{r} \cdot \frac{1}{2}$$

$$\tan\theta \leq \frac{1-r^2}{2r}$$



$$(hyp)^2 = (1-x^2)^2 + (2x)^2$$

$$hyp = \sqrt{(1-x^2)^2 + 4x^2}$$

$$= \sqrt{1+x^4-2x^2+4x^2}$$

$$= \sqrt{1+x^4+2x^2}$$

$$= \sqrt{(x^2+1)^2}$$

$$= x^2+1$$

we have $\sin\theta \leq \dfrac{1-x^2}{x^2+1}$

$$\sin\theta \leq 1-x^2$$

UNIT III : ASSOCIATION OF ATTRIBUTES

Association of Attributes - Coefficient of Association - Consistency - Time Series – Definition - Components Of Time Series - Seasonal and cyclic variations.

THEORY OF ATTRIBUTES

Attributes:

The qualitative characteristics of a population are called attributes and they cannot be measured by numeric quantities. Hence the statistical treatment required for attributes is different from that of quantitative characteristic.

Suppose the population is divided into two classes according to the presence or absence of a single attribute. The positive class denotes the presence of the attributes and the negative class denotes the absence of the attribute. Capital Roman letter such as A,B,C,D... are used to denote positive Greek letters such as $\alpha, \beta, \gamma, \delta$ ...... are used to denote negative classes.

For example If A represents the attribute richness then α represents the attribute non-richness (poor).

A class represented by n attributes is called a class of n^{th} order.

For example,

A,B,C, $\alpha, \beta, \gamma, \delta$ are all of first order, AB, $A\beta, \alpha B, \alpha\beta$ are of second order, and ABC, $A\beta\gamma, A\beta C, \alpha\beta\gamma$ are of the third order.

The number of individuals possessing the attributes in a class of n^{th} order is called a class frequency of order 'n' and class frequencies are denoted by bracketing the attributes.

Thus (A) stands for the frequency of A the number of individuals possessing the attribute A and $(A\beta)$ stands for the number of individuals possessing of the attributes A and not B.

Note:

1. Class frequencies of the type (A), (AB), (ABC) are known as positive class frequencies.
2. Class frequencies of the type $(\alpha),(\beta),(\alpha\beta),(\alpha\beta\gamma)\ldots are$ known as negative class frequencies.
3. Class frequencies of the type $(\alpha B),(A\beta),(A\beta\gamma),(\alpha\beta C)\ldots$ are known as contrary frequencies.
4. The classes of highest order are called the ultimate classes and their frequencies are called the ultimate class frequencies.

Examples:

1. AB = (ABC) + (ABγ)

Consider, (ABγ) $= AB\gamma.N$

=AB(1-C).N

=AB.N − ABC.N

=(AB) − (BC)

$\therefore (AB) = (ABC) + (AB\gamma)$

2. If there are two attributes A and B we have,

N = (A) + (α) = (B) + (β)

Hence N = (A) + (α)

N = (AB) + (Aβ) + (αB) + ($\alpha\beta$)

And N = (B) +(β) = (AB) + $(\alpha B) + (A\beta) + (\alpha\beta)$

If there are three attributes A,B,C we have N = (A) + (α)

We have N = (A) + (α)

$\Rightarrow$ N = (AB) + (Aβ) +($\alpha\beta$) + $(\alpha\beta)$

Thus,

N = (ABC) + (ABγ) + $A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$

3. Consider two attributes A and B

Now, $(\alpha\beta) = \alpha\beta.N$

$= (1\text{-}A)\,(1\text{-}B).N$
$= (1\text{-}A\text{-}B + AB).\,N$
$= N\text{-}A.N\text{-}\,B.N + AB.N$
$= N\text{-}(A) - (B) + (AB)$

4. $(AB) = AB.N$

$= (1\text{-}\gamma)\,(1 - \beta).N$

$= (1\text{-}\alpha - \beta + \alpha\beta).N$

$= N - \alpha.N - \beta.N + \alpha\beta.N$

$= N\text{-}\,(\alpha) - (\beta) + (\alpha\beta)$

5. $(\alpha\beta\gamma) = \alpha\beta\gamma.N = (1 - A)(1 - B)(1 - C).N = N\text{-}A.N\text{-}B.N - C.N + AB.N + AC.N + BC.N - ABC.N$

$= N\text{-}(A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$

6. $N = (A) + (B) + (C) - (AB) - (BC) - (AC) + (ABC) + (\alpha\beta\gamma)$

Problem :

Given $(A) = 30$, $(B) = 25$, $(\alpha) = 30$ $(\alpha\beta) = 20$.

Find i) (N) (ii) (β) (iii) (AB) (iv) $(A\beta)$ (V) (αB)

Solution:

i) $N = (A) + (\alpha) = 30 + 30 = 60$
ii) $(\beta) = N - (B) = 60 - 25 = 35$
iii) $(AB) = AB.N$

$= (1\text{-}\alpha)\,(1 - \beta).N$

$= N\text{-}(\alpha) - (\beta) + (\alpha\beta)$

$= 60\text{-}30\text{-}35 + 20$

$= 15$

iv. $(A\beta) = A\beta.N = A(1 - B).N$
$= (A) - (AB)$
$= 30\text{-}15$
$= 15$

v. $(\alpha B) = \alpha B.N = (1 - A)B.N$

$= (B)\text{-}(AB)$

$= 25\text{-}15$

$=10$

Problem :

Given the following ultimate class frequencies of two attributes A and B. Find the frequencies of positive and negative class frequencies and the total number of observations.

$$(AB) = 975, (\alpha B) = 100, (A\beta) = 25, (\alpha\beta) = 950.$$

Solution:

Positive class frequencies are (A) and (B)

$(A) \quad = (AB) + (A\beta) = 975 + 25 = 1000$

$(B) \quad = (AB) + (\alpha B) = 975 + 100 = 1075$

Negative class frequencies are (α) *and* (β)

$(\alpha) = (\alpha B) + (\alpha\beta) = 100 + 950 = 1050$

$(\beta) = (A\beta) + (\alpha\beta) = 235 + 950 = 975$

$N = (A) + (\alpha) = (B) + (\beta)$

Taking,

$N = (A) + (\alpha) = 1000 + 1050 = 2050$

Problem :

Given the following positive class frequencies find the remaining class frequencies N = 20 (A) = 9; (B) = 12; (C) = 8; (AB) = 6; (BC= 4); (CA) = 4; (CA) = 4; (ABC) = 3

Solution:

There are three attributes A,B,C.

$\therefore$ The total number of class frequencies is $3^3=27$.

We are given only 8 class frequencies and we have to find the remaining 19 class frequencies. They are

Order 1:

$(\alpha) = N - (A) = 20 - 9 = 11.$

$(\beta) = N-(B) = 20-12=8$

$(\gamma) = N - (C) = 20 - 8 = 12$

Order 2:

$(A\beta) = A(1 - B).N$

$= (A) - (B)$

$= 9-6= 3$

$(\alpha B) = (1 - A)B.N$

$= (B) - (AB)$

$= 12-6 = 6$

$(A\gamma) = A(1 - C).N$

$= (A) - (AC)$

$= 9-4$

$= 5$

$(\alpha C) = (1 - A)C.N$

$=(C) - (AC)$

$= 8-4=4$

$(B\gamma) = B(1 - C)N$

$= (B) - (BC)$

$= 12-4 = 8$

$(\beta C) = (1 - B)\, C.N$

$= (C) - (BC)$

$= 8\text{-}4 = 4$

$(\alpha\beta) = (1 - A)(1 - B).N = N - (A) - (B) + (AB)$

$= 20\text{-}9\text{-}12 + 6 = 5$

$(\beta\gamma) = (1 - B)(1 - C).N$

$= N - (B) - (C) + (BC)$

$= 20\text{-}12\text{-}8+4$

$= 4$

$(\alpha\gamma) = (1 - A)(1 - C).N$

$= N\text{-} (A) - (C) + (AC)$

$= 20\text{-}9\text{-}8+4$

$= 7$

Order 3:

$(A\beta\gamma) = AB(1 - C).N$

$= (AB) - (ABC)$

$= 6\text{-}3 = 3$

$(A\beta C) = A(1 - B)C.N$

$= (AC) - (ABC)$

$= 4\text{-}3 = 1$

$(A\beta\gamma) = A\,(1 - B)(1 - C).N$

$= (A) - (AC) - (AB) + (ABC)$

$= 9\text{-}4\text{-}6+3 = 2$

$(\alpha BC) = (1 - A)BC.N$

$= (BC) - (ABC)$

$= 4\text{-}3\text{=}1$

$(\alpha B\gamma) = (1 - A)(1 - C).B.N$

$= (B) - (BC) - (AB) + (ABC)$

$= 12\text{-}4\text{-}6\text{+}3$

$= 5$

$(\alpha\beta C) = (1 - A)(1 - B)C.N$

$=(C) - (AC) - (BC) + (ABC)$

$= 8\text{-}4\text{-}4\text{+}3\text{=}3$

$(\alpha\beta\gamma) = (1 - A)(1 - C).N$

$= N\text{-}(A)\text{-}(B) - (C) + (AB) + (BC) + (CA) - (ABC)$

$= 20\text{-}9\text{-}12\text{-}8\text{+}6\text{+}4\text{+}4\text{-}3 = 2$

Problem :

In a class text in which 135 candidates were examined for proficiency in English and Maths. It was discovered that 75 students failed in English, 90 failed in Maths and 50 failed in both. Find how many candidates i) have passed in Maths ii) have passed in English, failed in Maths iii) have passed in both.

Solution:

Let A denote pass in English and B denote pass in Maths .

$\therefore$ (α) denotes fail in English and (β) denotes fail in Maths.

Given (α)= 75; $(\beta) = 90; (\alpha\beta) = 50; N = 135$

We have to find (i) (B) (ii) $(A\beta)$ $(iii)(AB)$

i) (B) = N-(β)

= 135-90

= 45

ii) Consider, $(\beta) = (A\beta) + (\alpha\beta)$

$$\Rightarrow (A\beta) = (\beta) - (\alpha\beta)$$
$$= 90 - 50$$
$$= 40$$

iii) (AB) = (1-α)$(1 - \beta).N$

= N- $(\alpha) - (\beta) + (\alpha\beta)$

= 135-75-90 + 50

= 20

Problem :

Given N = 1200; (ABC) = 600; $(\alpha\beta\gamma) = 50$; $(\gamma) = 270$;

$(A\beta) = 36$; $(\beta\gamma) = 204$; $(A) - (\gamma) = 192$; $(B) - (\beta) = 620$.

Find the remaining ultimate class frequencies .

Solution:

Since there are 3 attributes there are 2^3=8.Ultimate class frequencies we are given two.

Hence we have find the remaining six

They are (i) $(AB\gamma)$ (ii)$(A\beta C)$

(iii) (αBC) $(iv)(A\beta\gamma)$ $(v)(\alpha B\gamma)$ and $(vi)(\alpha\beta C)$

To find the frequencies of positive classes: (A), (B), (C); (AB), (BC), (AC).

First order: $(A) - (\alpha) = 192$

$$(A) + (\alpha) = 1200(= N)$$

Adding,

2(A)=1200+192

2(A)=1392

(A)=696

$(B) - (\beta) = 620$

$(B) - (\beta) = 620$ (=N)

Hence (B) = 910

Now, (C) = N- (γ)

= 1200 – 270

= 930.

Second order:

$(AB) = (A) - (A\beta) = 696 - 36$

= 660

$(BC) = (B) - (B\gamma) = 910 - 204$

= 706

We have, $N = (A) + (B) + (C) - (AB) - (BC) - (AC) + (ABC) + (\alpha\beta\gamma)$

$(AC) = (A) + (B) + (C) - (AB) - (BC) + (ABC) + (\alpha\beta\gamma)$

= 696 + 910 + 930 – 660 – 706 + 600 + 50= 620

Third order:

i. $(AB\gamma) = AB\,(1-C).N$

$= (AB) - (ABC)$

$= 660 - 600$

$= 60$

ii. $(A\beta C) = AC(1-B).N$

$= (AC) - (ABC)$

$= 620 - 600$

$= 20$

iii. $(\alpha BC) = (1-A)BC.N$

$= (BC) - (ABC)$

$= 706 - 600$

$= 106$

iv. $(A\beta\gamma) = A(1-B)(1-C).N$

$= (A) - (AB) - (AC) + (ABC)$

$= 696 - 660 - 620 + 600$

$= 16$

v. $(\alpha B\gamma) = (1-A)\,(1-C)B.N$

$= (B) - (AB) - (BC) + (ABC)$

$= 910 - 660 - 706 + 600$

$= 144.$

vi. $(\alpha\beta C) = (1-A)(1-B)C.N$

$= (C) - (AC) - (BC) + (ABC)$

$= 930 - 620 - 706 + 600 = 204$

Problem :

Given that $(A) = (\alpha) = (B) = (\beta) = N/2$

Show that i) (AB) ii) $(\alpha\beta)$ $(ii)(A\beta) = (\alpha B)$

Solution:

i. $(AB) = AB.N$

$= (1\text{-}\alpha)\,(1-\beta).N$

$= N\text{-}\,(\alpha) - (\beta) + (\alpha\beta)$

$= N - N/2 - N/2 + (\alpha\beta)$

$(AB) = (\alpha\beta)$

ii. $(A\beta) = A\beta.N$

$= (1\text{-}\alpha)\,(1-B).N$

$= N - (\alpha) - (B) + (AB)$

$= N - N/2 - N/2 + (\alpha B)$

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

Problem :

Of 500 men in a locality exposed to cholera 172 in all were attacked, 178 were inoculated and of these 128 were attacked. Find the number of persons.

i) not inoculated not attacked

ii) inoculated not attacked

iii) not inoculated attacked

Solution:

Denote the attribute A as attacked and the attribute B as inoculated.

Hence α denote "NOT ATTACKED"; β DENOTES "NOT INOCULATED".

Given, N= 500; (A) = 172; (B) = 178; (AB) = 128

To find (i) $(\alpha\beta)$ $(ii)(\alpha B)$ $(iii)(A\beta)$

i. $(\alpha\beta) = \alpha\beta.N$

= (1-A) (1-B).N

= N-(A) –(B) + (AB)

= 500 -172-178 + 128

= 278

i. $(\alpha B) = \alpha B.N = (1-A)B.N$

= (B) – (AB)

= 178 – 128 = 50

iii). $(A\beta) = A\beta.N = A(1-B).N$

= (A) – (AB)

= 172 – 128 = 44

Problem:

There were 200 students is a college whose results in the first semester, second semester and the third semester are as follows: 80 passed in the first semester; 75 passes in the second semester. 96 passed in the third semester 25 passed in all the three semester 46 failed in all the three semester 29 passed in the first two and failed in the third semester 42 failed in the first two semester but passed in the third semester. Find how many students passed in atleast two semesters

Solution:

Denoting "pass in first semester" as "A' Pass in second semester 'B' and pass in the third semester as 'C' we get.

N = 200; (A) = 80, (B) = 75 ; (C) = 96

$(ABC) = 25;\ (\alpha\beta\gamma) = 46; (AB\gamma) = 29; (\alpha\beta C) = 42$

We have to find $(AB\gamma) + (\alpha BC) + (A\beta C) + (ABC)$

Consider, $(C) = (AC) + (\alpha C)$

$= (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C)$

$\therefore (ABC) + (\alpha BC) + (A\beta C) = (C) - (\alpha\beta C)$

$= 96 - 42 = 54$

$\therefore (ABC) + (\alpha BC) + (A\beta C) + (AB\gamma) = 54 + 29 = 83$

Thus the number of students who passed in atleast two semester is 83.

Problem :

Given $(ABC) = 149;\ (AB\gamma) = 738; (A\beta C) = 225; (A\beta\gamma) = 1196; (\alpha BC) = 204; (\alpha B\gamma) = 1762; (\alpha\beta C) = 171; (\alpha\beta\gamma) = 21842.\ find\ (A), (B), (C), (AB), (AC), (BC) and\ N.$

Solution:

$N = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$

$= 149 + 738 + 225 + 1196 + 204 + 1762 + 171 + 21842.$

$= 26287$

(A) $= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = 149 + 738 + 225 + 1196$

$= 2308$

(B) $= (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) = 149 + 738 + 204 + 1762$

$= 2853$

(C) $= 749$

$(AB) = (ABC) + (AB\gamma) = 149 + 738 = 887$

$(AC) = (ABC) + (A\beta C) = 149 + 225 = 374$

$(BC) = (ABC) + (\alpha BC) = 353$

Problem :

In a very hotly fought battle 70% of the solders at least lost an eye 75% at least lost an ear 80% at least an arm and 85% at least lost a leg. How many at least must have lost all the four?

Solution:

Denoting "loosing an eye" A, "loosing a ear by B" "loosing an arm by C" and "loosing a leg by D"

We have

$N = 100, (A) \geq 70, (B) \geq 75, (C) \geq 80, (D) \geq 85.$

To find the least value of ABCD

$(ABCD) \geq (A) + (B) + (C) + (D) - 3N$

$\geq 70 + 75 + 80 + 85 - 300$

$= 10$

$(ABCD) \geq 10$

At least 10% of the soldiers lost all the four.

Problem :

A company producers tube lights and conducts a test on 5000 lights for production defects of frames (F); chokes (C); starters (S) and tubes (T). The following are the records of defects.

(F) = 130, (C)=120, (S) = 115, (T) = 86

(FC) = 100, (CS) = 130, (ST) = 75, (FT) = 60

(CT) = 54, (FS) = 37, (FCS) = 90 , (CST) = 85

(FST) = 112, (FCT) = 108, (FCST) = 5.

Find the percentage of the tube lights which pass all the four tests.

Solution:

Number of tube lights passing the four tests

$$= (1\text{-}F)\,(1\text{-}C)\,(1\text{-}S)\,(1\text{-}T)\,.N$$

$$= [1\text{-}(F+C+S+T) + (FC+CS+ST+FT+CT+FS) - (FCS+CST+STF+FCT) + FCST].N$$

$$= N\text{-}[(F)+(C)+(S)+(T)]+ [(FC)+(CS)+(ST) + (FT)+(CT)+(FS)] - [(FCS)+(FCT)+((FST)+(CST)] + (FCST)$$

$$= 5000\text{-} (130+20+115+86) + (100+130 +75+60+54+37)\text{-}(90+108+112+85) +5$$

$$= 5000\text{-}451+456\text{-}395+5$$

$$= 5461\text{-}846=4615$$

Out of 5000 tube lights 4615 pass the four tests for defects.

Percentage of tube lights which pass the four tests

$= \frac{4615}{5000} \times 100 = 92.3\%$

Exercises:

1.Given the frequencies (A) = 1150, $(\alpha) = 1120$, $(AB) = 1075$ $(\alpha\beta) = 985$. *Find remaining* class frequencies and total number of observations.

2. Given the following ultimate class frequencies find the frequencies of the positive and negative classes and the total number of observations.

(AB) = 733, $(A\beta) = 840$, $(\alpha B) = 699$; $(\alpha\beta) = 783$.

3. A survey reveals that out of 1000 people in locality 800 like coffee, 700 like tea, 660 like both coffee and tea. Find how many people like neither coffee nor tea.

4. An examination result shows the following data. 56% at least failed in part I Tamil, 76% at least failed in part II English 82% at least failed in major – chemistry and 88% at least failed ancillary maths. How many at least failed in all the four?

5. In a university examination 95% of the candidates passed partI, 70% passed in part II, 65% passed part III. Find how many at least should have passed the whole examination.

Consistency of data:

Definition:

A set of class frequencies is said to the consistent if none of them is negative otherwise the given set of class frequencies is said to be inconsistent.

We have the following set of criteria for testing the consistency in the case of single attributes and three attributes.

Attributes	Condition consistency	Equivalent positive class condition	Number of conditions
A	$(A)\geq 0$ $(\alpha)\geq 0$	$(A)\geq 0$ $(A)\leq N$(Since $(\alpha)=(1-A)N\geq 0$)	2
A,B	$(AB)\geq 0$ $(A\beta)\geq 0$ $(\alpha B)\geq 0$ $(\alpha\beta)\geq 0$	$(AB)\geq 0$ $(AB)\leq A$ $(AB)\leq B$ $(AB)\geq(A)+(B)-N$	2^2
A,B,C	$(ABC)\geq 0$ $(AB\gamma)\geq 0$ $(A\beta C)\geq 0$ $(\alpha BC)\geq 0$ $(A\beta\gamma)\geq 0$ $(\alpha B\gamma)\geq 0$ $(\alpha\beta C)\geq 0$ $(\alpha\beta\gamma)\geq 0$	i) $(ABC)\geq 0$ ii) $(ABC)\leq(AB)$ iii) $(ABC)\leq(AC)$ iv) $(\alpha BC)\leq(BC)$ v) $(ABC)\geq(AB)+(AC)-(A)$ vi) $(ABC)\geq(AB)+(BC)-(B)$ vii) $(ABC)\geq(AC)+(BC)-(C)$ viii) $(ABC)\leq(AB)+(BC)+(AC)-(A)-(C)+(N)$	2^3

Note:

In the case of 3 attributes conditions

(i) and (Viii)

$$\Rightarrow (AB)+(BC)+(AC)\geq (A)+(B)+(C)-N \;\ldots\ldots(ix)$$

Similarly,

(ii) and (vii)

$\Rightarrow$ (AC) +(BC) − (AB)$\leq (C)$ ……………………………………(x)

(iii) and (vi)

$\Rightarrow$(AB)+(BC)-(AC) $\leq$(B) ………………………………… ….(xi)

iv) and (v)

$\Rightarrow$ (AB) +(AC) − (BC)   (A) ……………………….(xii)

conditions (ix) to (xii) can be used to check the consistency of data when the class of first and second order alone are known.

Problem :

Find whether the following data are consistent. N= 600; (A) = 300;(B) = 400; (AB)=50.

Solution:

We calculate the ultimate class frequency $(\alpha\beta)$, $(\alpha B) and\ (A\beta)$

$(\alpha\beta) = \ \alpha\beta.N = (1 - A)(1 - B).N$

= N-(A) −(B) +(AB)

= 600 − 400 +50

= -50

Since $(\alpha\beta) < 0$, the data are inconsistent.

Problem :

Show that there is some error in the following data: 50% of people are wealthy and healthy 35% are wealthy but not healthy 20% are healthy but not wealthy.

Solution:

Taking "wealth" as A and "health as "B" we get the following data

N=100, (AB) =50; $(A\beta)$ = 35, (αB)=20

To check the consistency of data we find $(\alpha\beta)$

$(\alpha\beta) = \alpha\beta . N = (1 - A)(1 - B). N$

=N-(A) – (B) + (AB)

But (A) = (AB) +$(A\beta)$

=50+35=85

(B) = (AB) + (αB)

=50+20

=70

$(\alpha\beta) = 100 - 85 - 70 + 50$

=-5

$(\alpha\beta)$<0

Hence there is error in the data.

Problem :

Of 2000 people consulted 1854 speak Tamil; 1507 speak Hindi; 572 Speak English; 676 speak Tamil and Hindi; 286 speak Hindi and English; 114 speak Tamil; Hindi and English. Show that the information as it stands is incorrect.

Solution:

Let A,B,C denote the attribution of speaking Tamil, Hindi, English respectively.

Given, N= 2000, (A) = 1854, (B) = 1507 (C) = 572;

(AB)= 676; (AC)= 286, (BC) = 270, (ABC)= 114

Consider $(\alpha\beta\gamma) = \alpha\beta\gamma . N$

=(1-A) (1-B) (1-C).N

=N – (A) – (B) –(C) + (AB) +(BC) +(AC) – (ABC)

=2000 – 1854 – 1507 – 572 +676 + 270 + 286 – 114

=- 815

$$\therefore (\alpha\beta\gamma) < 0.$$

Hence the data are inconsistent.

∴The information is incorrect.

Problem :

Find the limits of (BC) for the following available data.

N = 125, (A) = 48, (B) = 62, (C) = 45

$(A\beta) = 7 \ and \ (A\gamma) = \ 18$

Solution:

To find (AB) and (AC)

$(AB) = (A) - (A\beta)$

$= 48-7 = 41$

$(AC) = (A) - (A\gamma)$

$= 48-18 = 30$

Now, by condition of consistency (ix)

$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$

$41 + (BC) + 30 \geq 48 + 62 + 45 - 125$

$$(BC) \geq -41 \ldots\ldots\ldots\ldots\ldots (i)$$

Also using (xii)

$(AB) + (AC) - (BC) \leq (A)$

$\Rightarrow (BC) \geq (AB) + (AC) - (A)$

$= 41+30-48 = 23$

$(BC) \geq 23 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$

Using (xi), $(AB) + (BC) - (AC) \leq (B)$

$\Rightarrow (BC) \leq (B) + (AC) - (AB)$

$= 62+ 30-41$

$=51$

$\therefore (BC) \leq 51 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (iii)$

Using (x), $(AC) + (BC) - (AB) \le (C)$

$\Rightarrow (BC) \le (C) + (AB) - (AC)$

$= 45+41-30$

$= 56$

$\therefore (BC) = 56$ ………………………………(iv)

From (i), (ii), (iii) and (iv) we get

$23 \le (BC) \le 56$

Problem :

Find the greatest and least value of (ABC) if (A)=50, (B)=60, (C)= 80, (AB) = 35, (AC)= 45 and (BC)=42

Solution:

The problem involves 3 attributes and we are given positive class frequencies of first order and second order only.

Using positive class conditions (ii), (iii), (iv) of consistency for 3 attributers

$(ABC) \le (AB) \Rightarrow (ABC) \le 35$

$(ABC) \le (BC) \Rightarrow (ABC) \le 42$

$(ABC) \le (AC) \Rightarrow (ABC) \le 45$

$$\Rightarrow (ABC) \le 45 \ldots\ldots\ldots\ldots\ldots\ldots (i)$$

Using (v) (vi) and (vii)

$(ABC) \ge (AB) + (AC) - (A)$

$$\Rightarrow (ABC) \ge 35 + 45 - 50 = 30$$

$(ABC) \geq (AB) + (BC) - (B)$

$\Rightarrow (ABC) \geq 35 + 42 - 60 = 17$

$(ABC) \geq (AC) + (BC) - (C)$

$\Rightarrow (ABC) \geq 45 + 42 - 80 = 7$

Thus $(ABC) \geq 30$

$(ABC) \geq 17$

$(ABC) \geq 7$

$\Rightarrow (ABC) \geq 30 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$

From (1) and (2) we get $30 \leq (ABC) \leq 35$

$\therefore$The least value of (ABC) is 30 and the greatest value of (ABC) is 35.

Problem :

If $\frac{(A)}{N} = x; \ \frac{(B)}{N} = 2x, \frac{(C)}{N} = 3x$ and

$\frac{(AB)}{N} = \frac{(AC)}{N} = \frac{(BC)}{N} = y$. prove that neither x *nor* y can exceed ¼ .

Solution:

Clearly x *and* y are positive integers. The condition of consistency

$$(AB) \leq (A)$$

$$\Rightarrow \frac{(AB)}{N} \leq \frac{(A)}{N}$$

$$y \leq x$$

Similarly,

$(BC) \leq (B) \ \Rightarrow \ y \leq 2x$

$\Rightarrow y \leq x$ ……………………………(1)

Now, $(AB) \geq (A) + (B) - N$

$\Rightarrow \frac{(AB)}{N} \geq \frac{(A)}{N} + \frac{(B)}{N} - 1$

Thus, $(AB) \geq (A) + (B) - N$

$y \geq 3x - 1$

Similarly

$(BC) \geq (B) + (C) - N$

$\Rightarrow y \geq 5x - 1$

$\Rightarrow y \geq 5x - 1$ ……………………(2)

$(AC) \geq (A) + (C) - N$

By (1) and (2) $5x - 1 \leq y \leq x$.

Taking $5x-1 \leq x$ $we get$ $x \leq \frac{1}{4}$

Taking $y \leq x$ $we get$ $y \leq \frac{1}{4}$

Neither x nor y can exceed ¼.

Exercises:

1. Examine the consistency of data when

i) (A)=800; (B)= 700, (AB)=660; (N)= 1000

ii) (A)=600; (B)= 500, (AB)= 50; N= 1000

iii) N=2100; (A)=1000, (B)=1300; (AB)=1100

iv) N=100; (A)=45; (B)=55, (C) = 50; (AB)=15 , (BC)= 25, (AC)= 20, (ABC)=12

v)N=1800; (A)=850; (B)=780; (C)=326; (AB)=250; (BC)=122; (AC)=144;(ABC)= 50

2. A market investigator returns the following data of 2000 people consulted 1754 liked chocolates 1872 liked toffee and 572 liked biscuits, 678 liked chocolate and coffee, 236 liked chocolates and biscuits, 270 liked chocolates and biscuits, 270 liked toffee and biscuits, 114 liked all the three .Show that the information it started must be incorrect.

3. If (A) = 50; (B)= 60; (C)=50; $(A\beta) = 5$;

$(A\gamma) = 20$ *and* $N = 100$. Find the least and greatest value of (BC).

Independence and Association of Data:

Two attributes A and B are said to be independent if there is same proportion of A's amongst B as amongst β's.

Thus A and B are independent iff

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \text{.......................................(i)}$$

or

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \text{.......................................(ii)}$$

From (i) we get

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

$$\therefore (AB) = \frac{(A)(B)}{N} \text{..........................(1)}$$

$$\text{And } (A\beta) = \frac{(A)(\beta)}{N} \text{ ....(2)}$$

Again from (1) we get

$1-\frac{(AB)}{(B)} = 1-\frac{(A\beta)}{(\beta)}$

$\frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$

$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$

$\therefore \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$

$= \frac{(\alpha\beta)+(\alpha B)}{(\beta)+(B)}$

$= \frac{(\alpha)}{N}$

$(\alpha\beta) = \frac{(\alpha)(\beta)}{N}$ ……………………………(3)

And $(\alpha B) = \frac{(\alpha)(B)}{N}$ ……………………………(4)

(1),(2),(3),(4) are all equivalent conditions for independent of the attribute A and B.

Association and Coefficient of Association:

If (AB) $\neq \frac{(A)(B)}{N}$ we say that A and B are associated. There are two possibilities.

If (AB) > $\frac{(A)(B)}{N}$ we say that A and B are positively associated and If (AB) < $\frac{(A)(B)}{N}$ we say that A and B are negatively associated.

Let us denote $\delta = (AB) - \frac{(A)(B)}{N}$

ie. $\delta = \frac{1}{N}[\,(AB)\,(\alpha\beta) - (A\beta)(\alpha\beta)]$

Note:

i. A and B are independent if $\delta = 0$.

ii. A and B are positively associated if $\delta > 0$ and negatively associated if $\delta < 0$.

Coefficient of association:

There are several measures indicating the intensitivity of association between two attribution

A and B.

The most commonly used measures are the Yule's coeficiency of association Q and coefficient of colligation Y which are defined as follows.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Y = \frac{\left[1 - \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}{\left[1 + \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}$$

Problem :

Check whether the attributes A and B are independent given that (i) = 30 (B)= 60, (AB)= 12, N= 150

(ii)(AB) = 256, $(\alpha B) = 768, (A\beta) = 48, (\alpha\beta) = 144$.

Solution:

Given class frequencies are of first order condition for independence is

(AB) = $\frac{(A)(B)}{N}$

Consider,

$$= \frac{(A)(B)}{N} = = \frac{30\times60}{150} = 12 = (AB)$$

$$\therefore (AB) = \frac{(A)(B)}{N}$$

Hence A and B are independent.

ii) $(A) = (AB) + (A\beta) = 256 + 48 = 304$

$(B) = (AB) + (\alpha B) = 256 + 768 = 1024$

$(\alpha) = (\alpha B) + (\alpha\beta) = 768 + 144 = 912$

$(\beta) = (A\beta) + (\alpha\beta) = 48 + 144 = 192$

$N = (A) + (\alpha) = 304 + 912 = 1216$

$$\text{Now} = \frac{(A)(B)}{N} = \frac{304 \times 1024}{1216} = 256 = (AB)$$

$$\therefore (AB) = \frac{(A)(B)}{N}$$

Hence A and B are independent.

Problem :

In a class test in which 135 candidates were examined for proficiency in physics and chemistry, it was discovered that 75 students failed in physics, 90 failed in chemistry and 50 failed in both. Find the magnitude of association and state if there is any association between failing in physics and chemistry.

Solution:

Denoting "fail in Physics" as A and "fail in Chemistry" as B we get

$(A) = 75$, $(B) = 90$, $(AB) = 50$, $N = 135$

The magnitude of association is measured by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)\beta(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$(\alpha) = N - (A) = 135 - 75 = 60$$

$$(\beta) = N - (B) = 135 - 90 = 45$$

$$(\alpha B) = (B) - (AB) = 90 - 50 = 40$$

$$(A\beta) = (A) - (AB) = 75 - 50 = 25$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = 60 - 40 = 20$$

$$Q = \frac{50\times20 - 25\times40}{50\times20 + 20\times40}$$

$$Q = 0$$

∴ A and B are independent hence failure in physics and chemistry are completely independent of each other.

Problem :

Show whether A and B are independent or positively associated or negatively associated in the following cases.

i) N = 930, (A) = 300, (B) = 400, (AB) = 230

ii) (AB) = 327, $(A\beta) = 545, (\alpha B) = 741, (\alpha\beta) = 235$

iii) (A) = 470, (AB) = 300, $(\alpha) = 530, (\alpha B) = 150$

iv. (AB) = 66, $(A\beta) = 88, (\alpha B) = 102; (\alpha\beta) = 136$

Solution:

i) $\frac{(A)(B)}{N} = \frac{300 \times 400}{930} = 129.03$

Now, $\delta = (AB) - \frac{(A)(B)}{N}$

$= 230 - 129.03$

$= 100.97$

Here $\delta > 0$

Hence A and B are positively associated.

ii) $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$

$= \frac{327 \times 235 - 545 \times 741}{327 \times 235 + 545 \times 741}$

$= \frac{76845 - 4038845}{76845 + 403845}$

$= \frac{-32700}{480690}$

$= -0.6803$

$Q < 0.$

Hence A and B are negatively associated.

iii) N= (A) + (α)

$= 470 + 530$

$= 1000$

(A)= (AB) + (αB)

$= 300 + 150$

$= 450$

Now, $\frac{(A)(B)}{N} = \frac{470 \times 450}{1000} = 2115$

$\therefore \delta = (AB) - \frac{(A)(B)}{N}$

$= 300 - 2155$

$= -1825$

$\therefore \delta < 0.$

Hence A and B are negatively associated.

iv. $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) - (A\beta)(\alpha B)}$

$= \frac{66\times136 - 88\times102}{66\times136 + 88\times102} = 0. \therefore$ A and B are independent.

Problem :

Calculate the co-efficient of associate between intelligence of father and son from the following data.

Intelligent father with intelligent sons 200.Intelligent fathers with dull sons 50. Dull fathers with intelligence sons 110. Dull fathers with dull sons 600. Comment on the result.

Solution:

Denoting the "intelligence of fathers" as A and intelligence of sons" by B

we have

(AB) = 200, $(A\beta) = 50$, $(\alpha B) = 110$, $(\alpha\beta) = 600$

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{200 \times 600 - 50 \times 110}{200 \times 600 + 50 \times 110}$$

$$= 0.91235$$

Since Q is positive it means that intelligent fathers are likely to have intelligent sons.

Problem :

Investigate from the following data between inoculations against small pox prevention from attack.

	Attacked	Not attacked	Total
Inoculated	25	220	245
Not inoculated	90	160	250
Total	115	380	495

Solution:

Denoting A as "inoculated" and B as "attacked" we have (AB)= 25, $(A\beta) = 220, (\alpha B) = 90$ *and*

$(\alpha\beta) = 160.$

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{25\ X\ 160 - 220\ X\ 90}{25\ X\ 160 + 220\ X\ 90}$$

$$= \frac{400 - 19800}{400 + 19800}$$

$$= \frac{-15800}{23800}$$

$= -0.6638.$

Attributes A and B have negative association.

i.e. "Inoculation" and "attack from small pox" are negatively associated.

Thus inoculation against small pox can be taken as the preventive measure.

Problem :

From the following data compare the association between marks in physics and chemistry in MKU and MSU

University	MSU	MKU
Total number of candidate	200	1600
Pass in physics	80	320
Pass in chemistry	40	90
Pass in physics and chemistry	20	30

Solution:

Denoting "pass in physics" as A and "pass in chemistry" as B.

We have,

MKU	MSU
N=1600	N=200
(A) = 320	(A)=80
(A)= 90	(A)= 40
(AB) = 30	(AB) = 20

From the above data we get the rest of the class frequencies for MKU and MSU.

MKU	MSU
$(A\beta) = (A) - (AB)$	$(A\beta) = (A) - (AB)$
$= 320 - 30$	$= 80 - 20$
$= 290$	$= 60$
$(\alpha B) = (B) - (AB)$	$(\alpha B) = (B) - (AB)$
$= 90\text{-}30$	$= 40\text{-}20$
$= 60$	$= 20$
$(\alpha\beta) = N - (A) - (B) + (AB)$	$(\alpha\beta) = N - (A) - (B) + (AB)$
$= 1600\text{-}\ 320 - 90 + 30$	$= 200 - 80 - 40 + 20$
$= 1220$	$= 100$

We now find the coefficient of association between A and B for MKU and MSU

	Passed	Failed	Total
Married	90	65	155
Unmarried	260	110	370
Total	350	175	525

3. From the figures given in the following table compare the association between literacy and un employment in rural and urban areas- and given reasons for the difference if any

	Urban	Rural
Total adult males	25 lakhs	200 lakhs
Literate males	10 lakhs	40 lakhs
Unemployed males	5 lakhs	12 lakhs
Literate and unemployed males	3 lakhs	4 lakhs

Time series:

Definition:

Time series is a series of values of a variable over a period of time arranged chronologically

Components of a time series:

The various forces affecting the values of a phenomenon in a time series may be broadly classified into the following three categories generally known as the components of a time series.

1. Longtime trend (or) secular trend
2. Short term fluctuations (or) periodic movements
3. Irregular fluctuations

1. Long time trend:

The general tendency of a time series is to increase or decrease or stagnate over a period of several years. Such a long run tendency of a time series to increase or decrease over a period of time is known as secular trend or simply trend. Though the term "long" is a relative term it depends upon the nature of the series under consideration.

The long term trend does not mean that the series should continuously move in one direction only. It is possible that different tendencies of increases and decrease persist together. A graphical representation indicating a long term increase or decrease or stability is given is the following figures.

2. Short term fluctuations:

In most of the time series a number of forces repeat themselves periodically over a period of time preventing the values of the series to move in a particular direction. The variations caused by such forces are called short term fluctuations. This short term fluctuations may broadly be classified into (a) seasonal variation (B) cyclical variation

a) Seasonal variation:

Generally seasonal variations are considered as short term fluctuations that occur within a year. These fluctuations may be regular as well as irregular with in a period of one year

b) Cyclical variation

If the period of oscillation for the periodic movements is a time series is greater than one year then it is called cyclical variation. Generally oscillatory movement is nay business activity is due to the out time of the business cycles normally having four phases namely prosperity recession, depression and recovery. The period between two successive peaks or though is known as the period of the cycle. In cyclical variation generally the period of a cycle is three to eleven years.

3. Irregular fluctuations

The fluctuation which are purely random and due to unforeseen and unpredictable forces are called Irregular fluctuations

Measurement of trends

A graphical representation of a time series exhibits the general upwards and downward tendencies

The following are the four study of measurement of the trend in a time series

i) Graphic method
ii) Method of curve fitting by the principles of least squares.
iii) Method of semi averages
iv)Method of moving averages.

i) Graphic Method

This is the simplest method of determining the trend. In this method all values of the time series are plotted on a graph paper and a smooth curve is drawn by free hand to pass through as many points as possible. The smoothing of the curve eliminates the other components such as seasonal, cyclic and random variations.

ii) The method of curve fitting:

This is the best method of fitting a trend and it is commonly used in practice.

iii) Method of semi averages

In this method the whole time series data is classified into two equal parts with respect to time. Having divided the given series into two equal parts we calculate the arithmetic mean for each part. These means are called semi-averages. Then these average are plotted against the mid values of the respective period covered by each part. The line joining these points give the straight line trend for the time series.

iv) Method of moving averages

This method for measuring the trend consists of obtaining a series of moving average of successive m terms of the time series. This averaging process smoothens the fluctuations and the UPS and down in the given data. It has been observed and proved mathematically that if a trend is liner the period of the moving average is taken to be the period of oscillation.

Measurement for seasonal variation

There is a simple method for measuring the seasonal variation which involves simple averages.

Simple average method

Step 1:

All the data are arranged by years and months.

Step 2:

Compute the simple average $\bar{x}_i$ for i^{th} months

Step 3:

Obtain the overall average x of these average $\bar{x}_i$ and

$$\bar{x} = \frac{\bar{x}1 + \bar{x}2 + \cdots \ldots + \bar{x}12}{12}$$

Step4:

Seasonal indices for different months are calculated by expressing monthly average as the percentage of the overall average x

Thus seasonal index for i^{th} month $= \frac{\bar{x}_i}{\bar{x}} \times 100$ Take X = x - 1987 and Y = y-42

Then the line of best fit become

Y=ax + b

The normal equations are $\sum xy = a\sum x^2 + b\sum x$

$$\sum y = a\sum x + nb, where\ n = 11$$

From the table,

-19 = 110 a

$\Rightarrow a = \frac{-19}{110} = -0.17$

17 = 11b

$\Rightarrow b\frac{17}{11} = 1.55$

∴ The line of best fit is Y= - 0.17 x + 1.55

$$ie. Y - 42 = -0.17\,(x - 1987) + 1.55$$

$$y = -0.17x + 1987 \times 0.17 + 1.55 + 42$$

$$y = -0.17x + 381.34\ is\ the\ straight\ line\ trend$$

Problem:

Use the method least squares and fit a straight line trend to the following data given from 82 to 92. Hence estimate the trend values for 1993.

Year	82	83	84	85	86	87	88	89	90	91	92
Production in quintals	45	46	44	47	42	41	39	42	45	40	48

Solution:

Let the line of best fit be

Y=ax + b Take X = x – 1987 and Y = y-42

Then the line of best fit become

Y=ax + b

The normal equations are $\sum xy = a\sum x^2 + b\sum x$

$$\sum y = a\sum x + nb, where\ n = 11$$

From the table, -19 = 110 a $\Rightarrow a = \frac{-19}{110} = -0.17$

X	X= x-1987	Y	Y= y-42	XY	X^2
1982	-5	45	3	-15	25
1983	-4	46	4	-16	16
1984	-3	44	2	-6	9
1985	-2	47	5	-10	4
1986	-1	42	0	0	1

1987	0	41	-1	0	0
1988	1	39	-3	-3	1
1989	2	42	0	0	4
1990	3	45	3	9	9
1991	4	40	-2	-8	18
1992	5	48	6	30	25
1993	0	-	17	-19	110

$$17 = 11b \Rightarrow b\frac{17}{11} = 1.55$$

∴ The line of best fit is Y= - 0.17 x + 1.55

$$\text{ie, } Y - 42 = -0.17(x - 1987) + 1.55$$

$$y = -0.17x + 1987 \times 0.17 + 1.55 + 42$$

$$y = -0.17x + 381.34 \textit{ is the straight line trend}$$

From the line trend

When x =1982, y=44.4

X = 1983, y=44.23, x=1984, y=44.06

X= 1985, y=43.89, x= 1986, y=43.72

X=1987, y= 43.55, x=1988, y=43.38 X=1989, y=43.21, x=1990, y= 43.04

X = 1991, y = 42.87, x= 1992, y=42.7

Thus the trend values are 44.4, 44.23, 44.06, 43.89, 43.72, 43.58, 43.38, 43.21, 43.04, 43.04, 42.87, 42.7

Problem:

Calculate the seasonal variation indices from the following data

Month	Monthly sales in lakhs				Total	$\bar{x}_i$	Seasonal indices $\frac{\bar{x}_i}{\bar{x}}$x 100
	I 1991	II 1992	III 1993	IV 1994			
January	10	11	11.5	13.5	46	11.5	$\frac{11.5}{12}$x100=95.8
February	8.5	8.5	9	10	36	9	$\frac{9}{12}$x100 = 75
March	10.5	12	11	12.5	46	11.5	$\frac{11.5}{12}$x100 = 95.8
April	12	14	16	18	60	15	$\frac{15}{12}$x100 = 125
May	10	9	12	15	46	11.5	$\frac{11.5}{12}$x100 = 95.8
June	10.5	10.5	11	14	46	11.5	$\frac{11.5}{12}$x100 = 95.8
July	12	14	13	17	56	14	$\frac{14}{12}$x100 = 116.7
August	9	8	11	16	44	11	$\frac{11}{12}$x100 = 91.7
September	11	11	12.5	13.5	48	12	$\frac{12}{12}$x100 = 100
October	10	9.5	11.5	13	44	11	$\frac{11}{12}$x100 = 91.7
November	11	12.5	10.5	14	48	12	$\frac{12}{12}$x100 = 100
December	12	13	15	16	56	14	$\frac{14}{12}$x100 = 116.7
Total						$\frac{144}{12}$	

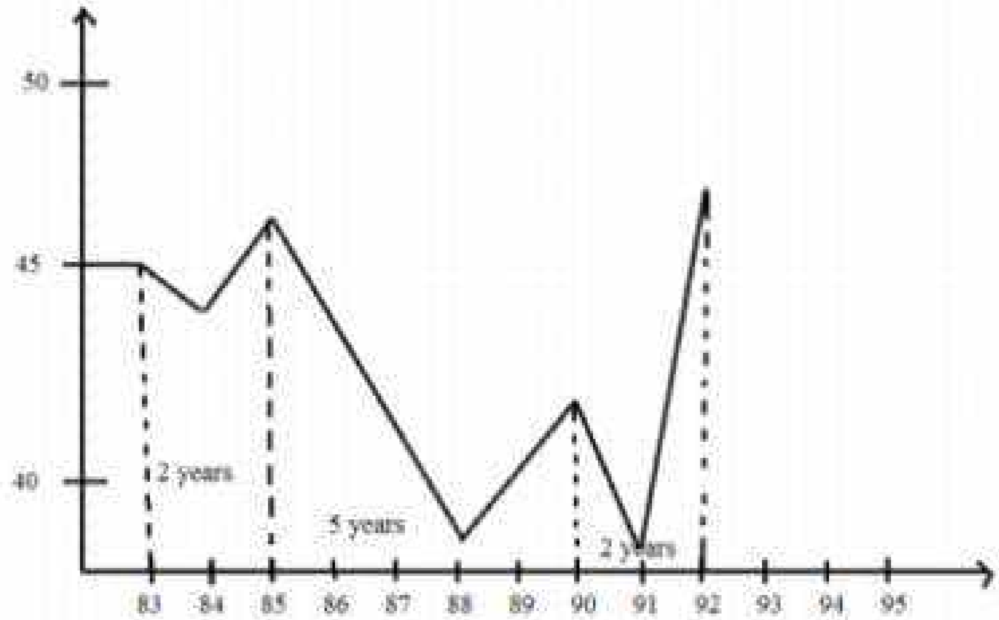
I year	II production in quintals	III 4 yearly moving total	IV 4 yearly moving average	V 2 period moving total	VI trend values (V)/ 2
1982	45	-	-	-	-
1983	46	-	-	-	-
1984	44	182	45.50	90.25	-
1985	47	179	44.75	88.25	45.13
1986	42	174	43.50	85.75	44.13
1987	41	169	42.25	83.25	42.88
1988	39	164	41.00	82.75	41.63
1989	42	167	41.75	83.25	41.38
1990	45	166	41.50	85.85	41.63
1991	40	175	43.75	-	42.93
1992	48	-	-		-

Problem:

Compute the trend values by the method of A yearly moving average for the data given in problem 1.

Problem:

Determine the suitable period of moving average for the data given in problem 1



We observe that the data has peaks at the following years 1983, 1985, 1985, 1990 and 1992.

Thus the data shows 3 cycles with varying periods 2,5,2 respectively.

Hence the suitable period of moving average is taken to be the A.M.

periods.

Hence $\frac{2+5+2}{5} = 3$ is the period of moving average.

Problem:

Compute the seasonal indices for the following data by simple average method

	Season	1990	1991	1992	1993	1994
Princes in different season	Summer	68	70	68	65	60
	Monson	60	58	63	56	55
	Autumn	61	56	68	56	55
	winter	63	60	67	55	58

Solution:

Year	Summer	Monsoon	Autumn	Winter	Total
1990	68	60	61	63	
1991	70	58	56	60	
1992	68	63	68	67	
1993	65	56	56	55	
1994	60	55	55	58	
total	331	292	296	303	
average	66.2	58.4	59.2	60.6	244.4
Seasonal index	$\frac{66.2}{61.1}$x100 = 108.3	$\frac{58.4}{61.1}$x100 = 95.6	$\frac{59.2}{61.1}$x100 = 69.9	$\frac{60.6}{61.1}$x100 = 99.2	$\bar{x} = 61.1$

Exercises:

1. Room the data given below calculate the seasonal indicates assuming that trend is absent

Year	I quarter	II quarter	III quarter	IV quarter
1990	40	35	38	40
1991	42	37	39	38
1992	41	35	38	40
1993	45	36	36	41
1994	44	38	38	42

2. Compute the seasonal index for the following data assuming that there is no need to adjust the data for the trend

Quarter	1989	1990	1991	1992	1993	1994
I	3.5	3.5	3.5	4.0	4.1	4.2
II	3.9	4.1	3.9	4.6	4.4	4.6
III	3.4	3.7	3.7	3.8	4.2	4.3
IV	3.6	4.8	4.0	4.5	4.5	4.7

2. RANDOM VARIABLE

Let S be a sample space associated with a given random experiment. A real valued function defined on S and taking values in $R(-\infty,\infty)$ is called one dimensional random variable.

A random variable X is a rule which associates uniquely a real number with every elementary event $E_i \in S$, $i = 1,2,3,\ldots n$ i.e, a random variable is a real valued function which maps the sample space on to the real line. Discrete Random Variables and Continuous Random Variables are the two types of a random variable.

2.1 DISCRETE RANDOM VARIABLE

A variable which can assume only a countable number of real values and for which the value which the variable takes depends on chance is called discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable. For instance, numbers of members of family, number of students in a class, number of passenger in a bus, tossing a coin and rolling a dice are the example of discrete random variable.

2.1.1 Probability Mass Function

If X is one dimensional discrete random variable taking at most a countable in finite number of values $x_1, x_2, x_3,\ldots$ then it is probabilistic behaviour at each real point described by a function called the probability mass function.

Definition:

If X is a discrete random variable with distinct $x_1, x_2, x_3, \ldots x_n, \ldots$ then the function P(x) defined as $P_X(x) = \begin{cases} P(X = x_i) & \text{if } x = x_i \\ 0 & \text{if } x \neq x_i ; i = 1,2,3,\ldots \end{cases}$ is called the probability mass function of random variable X

Remarks: The numbers $p(x_i)$; $i = 1,2,3,\ldots$ must satisfy the following conditions:

(i) $P(x_i) \geq 0$ and (ii) $\sum_{i=1}^{\infty} P(x_i) = 1$

2.2 CONTINUOUS RANDOM VARIABLE

A random variable which can assume any value from a specified interval of the form [a,b] is known as continuous random variable.

2.2.1 PROBABILITY DENSITY FUNCTION

If X is a continuous random variable, it will have infinite number of values in any interval however small. The probability that this variable lies in the infinitesimal interval $(x, x+dx)$ is expressed as $f(x)\,dx$, where the function $f(x)$ is called probability density function *(p.d.f)*, satisfying the following conditions

(i) $f(x) \geq 0 \quad \forall x$ *(ii)* $\int_{-\infty}^{\infty} f(x)\,dx = 1$

2.3 DISTRIBUTION FUNCTION

Let X be a random variable, the function F defined for all real x by $F(x) = P(X \leq x)$ is called the distribution function*(d.f)* or cumulative distribution function of the random variable X.

If random variable X is discrete then distribution function is $F(x) = P(X \leq x)$

If X is continuous random variable then distribution function is

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

2.3.1 Properties of Distribution Function

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

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

2. If F is the distribution function of random variable X then

(i) $0 \leq F(X) \leq 1$ (ii) $F(x) \leq F(Y)$ if $x < y$

3. If F is the distribution function of random variable X then

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad F(\infty) = \lim_{x \to \infty} F(x) = 1$$

4. $\frac{d}{dx}(F(x)) = f(x)$

Example 2.1 If the random variable X takes the value 1, 2, 3 and 4 such that

2P(X=1)=3P(X=2) = P (X=3)=5P(X=4).Find the probability distribution?

Solution:

$2P(X=1)= k \Rightarrow P(X=1) = k/2$

$3P(X=2) = k \Rightarrow P(X=2) = k/3$

$P(X=3) = k$

$5P(X=4)= k \Rightarrow P(X=4) = k/5$

$$\sum_{x=1}^{4} P(x_i) = 1$$

$$\frac{k}{2}+\frac{k}{3}+k+\frac{k}{5}=1$$

$$k=\frac{30}{61}$$

The probability distribution is

x	1	2	3	4
$P(X=x)$	15/61	10/61	30/61	6/61

Example 2.2 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 $P(X<6)$,$P(X\geq 6)$ and $P(o<X<5)$ (iii) Determine the distribution function of X and (iv) $P(X\leq a)>1/2$ find the minimum value of a,

Solution:

$$\sum_{x=0}^{7} P(x_i) = 1$$

$$k+2k+2k+2k+3k+k^2+2k^2+7k^2+k=1$$

$\Rightarrow 10k^2+9k-1=0 \quad \Rightarrow (10k-1)(k+1) = 0 \quad \Rightarrow k = \frac{1}{10}$ or $k = -1$(negative)

Hence $k = \frac{1}{10}$

(ii) P(X<6) = P(X=0)+ P(X=1)+ P(X=2)+ P(X=3)+ P(X=4)+ P(X=5)

$$= k + 2k + 2k + 2k + 3k + k2$$

$$P(X < 6) = \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100}$$

$P(X \geq 6) = 1 - P(X<6) = 1 - \frac{81}{100} = \frac{19}{100}$

$P(X \geq 6) = \frac{19}{100}$

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

$$= k+2k+2k+2k+3k = 8k = \frac{8}{10}$$

$P(0<X<5) = \frac{8}{10}$

(iii) *Distribution function of X*

$F(x) = P(X \leq x)$

x	$F(x) = P(X \leq x)$
0	0
1	$k = \frac{1}{10}$
2	$k + 2k = 3k = \frac{3}{10}$
3	$k + 2k + 2k = 5k = \frac{5}{10}$
4	$k + 2k + 2k + 3k = 8k = \frac{8}{10}$
5	$k + 2k + 2k + 3k + k^2 = 8k + k^2 = \frac{8}{10} + \frac{1}{100} = \frac{81}{100}$
6	$k + 2k + 2k + 3k + k^2 + 2k^2 = 8k + 3k^2 = \frac{8}{10} + \frac{3}{100} = \frac{83}{100}$
7	$k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 9k + 10k^2 = \frac{9}{10} + \frac{10}{100} = 1$

(iv) *P(X≤a)>1/2 find the minimum value of a*

From the distribution function $P(X \leq 4) = \frac{8}{10} = \frac{4}{5} > \frac{1}{2}$

a = 4

Example 2.3 A discrete random variable X has the following probability distribution

x :	0	1	2	3	4	5	6	7	8
p(x):	a	3a	5a	7a	9a	11a	13a	15a	17a

(i) Find the value of 'a'

(ii) $P(0 < X < 3)$

(iii) $P(X \geq 3)$

(iv) Find the distribution function of X

Solution: We have $\sum_{i=1}^{n} P(X = x) = 1$

$$a+3a+5a+7a+9a+11a+13a+15a+17a = 1$$

$$\therefore 81a = 1 \Rightarrow a = \frac{1}{81}$$

$\therefore$ The actual probability distribution is

x	0	1	2	3	4	5	6	7	8
P(X=x)	$\frac{1}{81}$	$\frac{3}{81}$	$\frac{5}{81}$	$\frac{7}{81}$	$\frac{9}{81}$	$\frac{11}{81}$	$\frac{13}{81}$	$\frac{15}{81}$	$\frac{17}{81}$

$$P(0 < X < 3) = P(X = 1) + P(X = 2) = \frac{3}{81} + \frac{5}{81} = \frac{8}{81}$$

$$P(0 < X < 3) = \frac{8}{81}$$

$$P(X \geq 3) = 1 - P(X<3) = 1 - \left\{\frac{1}{81} + \frac{3}{81} + \frac{5}{81}\right\} = \frac{72}{81}$$

The distribution function of X is

x	0	1	2	3	4	5	6	7	8
F(x)	0	$1/81$	$4/81$	$9/81$	$16/81$	$25/81$	$36/81$	$49/81$	1

Example 2.4 For the following density function, $f(x) = ae^{-|x|}, -\infty < x < \infty$,

find the value of 'a'

Solution:

Given f(x) is a pdf.

$$\therefore \int_{-\infty}^{\infty} f(x)dx = 1$$

$$a\int_{-\infty}^{\infty} e^{-|x|}dx = 1$$

$$2a\int_{0}^{\infty} e^{-x}dx = 1$$

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

$$2a\left(\frac{e^{-\infty}}{-1} - \frac{e^{-0}}{-1}\right) = 1$$

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

Example 2.5 The diameter of an electric cable, say X, is assumed to be a continuous random variable with *p.d.f* : $f(x) = 6x(1-x)$, $0 \le x \le 1$.

(i) Determine a number b such that P(X<b)=P(X>b).

(ii) Compute $P(X \le \frac{1}{2} / \frac{1}{3} \le X \le \frac{2}{3})$

Solution (i)

$$P(X < b) = P(X > b)$$

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

$$\Rightarrow \int_0^b 6x(1-x)dx = \int_b^1 6x(1-x)dx$$

$$\Rightarrow 6\int_0^b (x-x^2)dx = 6\int_b^1 (x-x^2)dx$$

$$\Rightarrow\left(\frac{x^2}{2}+\frac{x^3}{3}\right)_0^b=\left(\frac{x^2}{2}+\frac{x^3}{3}\right)_b^1$$

$$\Rightarrow\left[\left(\frac{b^2}{2}+\frac{b^3}{3}\right)-\left(\frac{0^2}{2}+\frac{0^3}{3}\right)\right]=\left[\left(\frac{1^2}{2}+\frac{1^3}{3}\right)-\left(\frac{b^2}{2}+\frac{b^3}{3}\right)\right]$$

$$\Rightarrow 3b^2-2b^3=(1-3b^2+2b^3)$$

$$\Rightarrow 4b^3-6b^2+1=0$$

$$\Rightarrow(2b-1)(2b^2-2b-1)=0$$

$$\therefore 2b-1=0 \ \Rightarrow\ b=\frac{1}{2}\,or$$

$$2b^2-2b-1=0\Rightarrow b=\frac{2\pm\sqrt{4+8}}{4}=\frac{1\pm\sqrt{3}}{2}$$

Hence $b=\frac{1}{2}$, is the only real value lying between 0 and 1

$$(ii)\ P(X\le\frac{1}{2}/\frac{1}{3}\le X\le\frac{2}{3})=\frac{P\left(X\le\frac{1}{2}\cap\frac{1}{3}\le X\le\frac{2}{3}\right)}{P\left(\frac{1}{3}\le X\le\frac{2}{3}\right)}$$

$$=\ \frac{P\left(\frac{1}{3}\le X\le\frac{1}{2}\right)}{P\left(\frac{1}{3}\le X\le\frac{2}{3}\right)}\qquad=\ \frac{\int_{1/3}^{1/2}6x(1-x)dx}{\int_{1/3}^{2/3}6x(1-x)dx}$$

$$=\ \frac{13/54}{13/27}=\frac{11}{26}$$

$$P(X\le\frac{1}{2}/\frac{1}{3}\le X\le\frac{2}{3})=\frac{11}{26}$$

Example 2.6 Let X be a continuous random variable with *p.d.f* given by

$$f(x)=\begin{cases} kx & ,0\le x<1\\ k & ,1\le x<2\\ -kx+3k & ,2\le x<3\\ 0 & ,otherwise\end{cases}$$

(i) find the value of k (ii)Determine the c.d.f

Solution:

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

$$\int_0^1 kx\,dx + \int_1^2 k\,dx + \int_2^3 (-kx+3k)\,dx = 1$$

$$k\left(\frac{x^2}{2}\right)_0^1 + k(x)_1^2 + \left(-k\frac{x^2}{2}+3kx\right)_2^3 = 1$$

$$k\left(\frac{1^2}{2}-\frac{0^2}{2}\right)+k(2-1)+\left(\left(-k\frac{3^2}{2}+3k3\right)-\left(-k\frac{2^2}{2}+3k2\right)\right)=1$$

$$k\left(\frac{1}{2}\right)+k+\left(\left(-k\frac{9}{2}+9k\right)-\left(-k\frac{4}{2}+6k\right)\right)=1$$

$$\frac{k}{2}+k+\left((k)\left(-\frac{9}{2}+9\right)-(k)(-2+6)\right)=1$$

$$\frac{k}{2}+k+\left((k)\left(\frac{-9+18}{2}-4\right)\right)=1$$

$$\frac{k}{2}+k+\left((k)\left(\frac{-9+18-8}{2}\right)\right)=1$$

$$\frac{k}{2}+k+\frac{k}{2}=1$$

$$\Rightarrow \frac{k+2k+k}{2}=1$$

$$\Rightarrow \frac{4k}{2}=1 \quad \Rightarrow 2k=1 \quad k=\frac{1}{2}$$

(ii) The $c.d.f$

For any x, such that $-\infty<x<0$;

$$F(x)=\int_{-\infty}^{x} f(x)dx \;=0$$

For any x, where $0\le x<1$;

$$F(x)=\int_{-\infty}^{0} 0\,dx+\int_0^x kx\,dx = k\int_0^x x\,dx = \frac{1}{2}\left(\frac{x^2}{2}\right)_0^x = \frac{1}{2}\left(\frac{x^2}{2}-\frac{0}{2}\right)=\frac{x^2}{4}$$

For any x, where $1 \le x < 2$;

$$F(x) = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} kx\,dx + \int_{1}^{x} k\,dx = k\int_{0}^{1} x\,dx + k\int_{1}^{x} dx$$

$$= \frac{1}{2}\int_{0}^{1} x\,dx + \frac{1}{2}\int_{1}^{x} dx = \frac{1}{2}\left(\frac{x^2}{2}\right)_0^1 + \frac{1}{2}(x)_1^x = \frac{1}{2}\left(\frac{1^2}{2} - \frac{0^2}{2}\right) + \frac{1}{2}(x-1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \frac{1}{2}(x-1)$$

$$= \frac{1}{4} + \frac{x-1}{2} = \frac{1+2(x-1)}{4} = \frac{1+2x-2}{4}$$

$$F(x) = \frac{2x-1}{4}$$

For any x, where $2 \le x < 3$;

$$F(x) = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} kx\,dx + \int_{1}^{2} k\,dx + \int_{2}^{x} -kx + 3k\,dx = k\int_{0}^{1} x\,dx + k\int_{1}^{2} dx + k\int_{2}^{x} -x + 3\,dx$$

$$= \frac{1}{2}\int_{0}^{1} x\,dx + \frac{1}{2}\int_{1}^{2} dx + \frac{1}{2}\int_{2}^{x} -x + 3\,dx$$

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

$$= \frac{1}{2}\left(\frac{1^2}{2} - \frac{0^2}{2}\right) + \frac{1}{2}(2-1) + \frac{1}{2}\left[\left(-\frac{x^2}{2} + 3x\right) - \left(-\frac{2^2}{2} + 3(2)\right)\right]$$

$$= \frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}(1) + \frac{1}{2}\left[\left(-\frac{x^2}{2} + 3x\right) - (-2+6)\right]$$

$$= \frac{1}{4} + \frac{1}{2} + \frac{1}{2}\left(-\frac{x^2}{2} + 3x - 4\right) = \frac{1}{4} + \frac{1}{2} + \left(-\frac{x^2}{4} + \frac{3}{2}x - \frac{4}{2}\right) = \frac{1+2-x^2+6x-8}{4}$$

$$F(x) = \frac{-x^2+6x-5}{4}$$

For any x, $x \ge 3$;

$$F(x)=\int_{-\infty}^{0} 0\,dx+\int_{0}^{1} kx\,dx+\int_{1}^{2} k\,dx+\int_{2}^{1} -kx+3k\,dx+\int_{3}^{x} 0\,dx=k\int_{0}^{1} x\,dx+k\int_{1}^{2} dx+k\int_{2}^{3} -x+3\,dx$$

$$=\frac{1}{2}\int_{0}^{1} x\,dx+\frac{1}{2}\int_{1}^{2} dx+\frac{1}{2}\int_{2}^{3} -x+3\,dx$$

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

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

$$=\frac{1}{2}\left(\frac{1}{2}\right)+\frac{1}{2}(1)+\frac{1}{2}\left(\left(-\frac{9}{2}+9\right)-(-2+6)\right)$$

$$=\frac{1}{4}+\frac{1}{2}+\frac{1}{2}\left(-\frac{9}{2}+9-4\right)=\frac{1}{4}+\frac{1}{2}+\frac{1}{2}\left(\frac{-9}{2}+5\right)=\frac{1}{4}+\frac{1}{2}+\frac{1}{2}\left(\frac{-9+10}{2}\right)=\frac{1}{4}+\frac{1}{2}+\frac{1}{4}=1$$

$$F(x)=1$$

Hence the distribution function F(x) is given by

$$F(x)=\begin{cases} 0 & for\ -\infty\le x<0 \\ \dfrac{x^2}{4} & for\ 0\le x<1 \\ \dfrac{2x-1}{4} & for\ 1\le x<2 \\ \dfrac{-x^2+6x-5}{4} & for\ 2\le x<3 \\ 1 & for\ 3\le x<\infty \end{cases}$$

Example 2.7 The cumulative distribution of continuous random variable X is given by

$$F(x)=\begin{cases} 0, & x<0 \\ x^2, & 0\le x<\frac{1}{2} \\ 1-\frac{3}{25}(3-x), & \frac{1}{2}\le x<3 \\ 0, & x\ge 0 \end{cases}$$

Find (i) Probability density function of X (ii) $P(|X|\le 1)$ and (iii) $P(\frac{1}{3}\le X<4)$

Solution:

We know that $f(x)=\frac{d}{dx}F(x)$

The points x = 0, ½ , 3 are points of continuity

$$\therefore f(x) = \begin{cases} 0, & x < 0 \\ 2x, & 0 \le x < \frac{1}{2} \\ \frac{6}{25}(3-x), & \frac{1}{2} \le x < 3 \\ 0, & x \ge 3 \end{cases}$$

$$P(|X| \le 1) = P(-1 \le X \le 1) = F(1) - F(-1) = \frac{3}{25}$$

$$P(\tfrac{1}{3} \le X < 4) = F(4) - F(\tfrac{1}{3}) = 1 - \frac{1}{9} = \frac{8}{9}$$

The 'average' value of a random phenomenon is also termed as its mathematical expectation or expected value. Once we have constructed the probability distribution for a random variable, to compute a mean or expected value of the random variables, where the weights are probabilities associated with the corresponding values. The mathematical expression for computing the expected value of a discrete random variable X with the probability mass function and computing the expected value of a continuous as random variable X with the probability density function are denoted byE(X)

$$E(X) = \begin{cases} \sum_{i=1}^{n} x_i \, P(X = x_i) & \textit{for discrete random variable} \\ \\ \int_{-\infty}^{\infty} x \, f(x)\, dx & \textit{for continuous random variable} \end{cases}$$

4.1.1 Properties of Expectation

Property 1. Addition Theorem of Expectation

If X and Y are random variables then E(X + Y) = E (X) + E(Y), provided all the expectation exists.

Proof

Let X and Y be a continuous random variables with joint p.d.f $f_{XY}(x, y)$ and marginal probability density functions of $f_X(x)$ and $f_Y(y)$ respectively.

$$E(X) = \int_{-\infty}^{\infty} x \, f(x)\, dx \qquad E(Y) = \int_{-\infty}^{\infty} y \, f(y)\, dy$$

$$E(X+Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)\, f_{XY}(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{XY}(x,y) dx. dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{XY}(x,y) dx\, dy$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{XY}(x,y) dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_{XY}(x,y) dx \right] dy = \int_{-\infty}^{\infty} x\, f_X(x) dx + \int_{-\infty}^{\infty} y\, f_Y(y) dy$$

$$\mathbf{E\,(X + Y) = E\,(X) + E\,(Y)}$$

Property 2: Multiplication theorem of Expectation

If X and Y are independent random variables, then E(XY) = E(X). E(Y).

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f_{XY}(x, y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\ dx.\ dy \qquad \text{X, Y are independent}$$

$$= \int_{-\infty}^{\infty} x f_X(x)dx\ \int_{-\infty}^{\infty} y f_Y(y)dy$$

E(XY)= E (X) . E (Y)

Property 3 If X is a random variable and 'a' is constant.

(i) $E[a\ \psi(X)] = a\ E[\psi(X)]$ (ii) $E[\psi(X) + a] = E[\psi(X)] + a$

Where ψ (X) is a function of X, is a r.v and all the expectation are exists.

Proof (i)

$$E[a\ \psi(X)] = \int_{-\infty}^{\infty} a\ \psi(x) f(x)\ dx \quad = a\int_{-\infty}^{\infty} \psi(x) f(x)\ dx$$

$$E[a\ \psi(X)] = a\ E[\psi(X)]$$

(ii)

(i) $$E[\psi(X) + a] = \int_{-\infty}^{\infty} [\psi(x) + a] f(x)\ dx \quad = \int_{-\infty}^{\infty} \psi(x) f(x)\ dx + \int_{-\infty}^{\infty} a\ f(x)\ dx$$

$$= E[\psi(X)] + a\int_{-\infty}^{\infty} f(x)\ dx \qquad \left(\because \int_{-\infty}^{\infty} f(x)\ dx = 1\right)$$

$$= E[\psi(X)] + a$$

Property 4. If X is a random variable and a and b are constants then E(aX + b) = a E(X) + b provided all the Expectations exists.

Proof

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

$$= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \qquad \left(\because \int_{-\infty}^{\infty} f(x)dx = 1\right)$$

$$E(aX + b) = \text{a E (X) + b}$$

Property 5 If $X \geq 0$ then $E(X) \geq 0$.

Proof

If x is continuous random variable such that $X \geq 0$ then

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} xf(x) > 0$$

[If $X \geq 0$ $f(X) = 0$ for $n < 0$] provided the expectation exists.

Property 6

If X and Y are two random variables such that $Y \leq X$, then $E(Y) \leq E(X)$, provided all expectations exists.

Proof:

Since $Y \leq X$

We have r.v $Y - X \leq 0 \Rightarrow X - Y \geq 0$.

Hence $E(X-Y) \geq 0$

$E(X) - E(Y) \geq 0$

$E(X) \geq E(Y)$

$\Rightarrow E(Y) \leq E(X)$.

$x = 3$

$y = 2$

4.2 Variance

The variance of a random variable X is defines as

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

4.2.1 Property

Let X is a random variable then $V(aX+b) = a^2V(X)$ where a and b are constants

If $Y=aX+b$ then

$E[Y] = E(aX+b) = aE[X]+b$

$Y-E[Y] = Y-(aE[X]+b)$

$= (aX+b)-(aE[X]+b)$

$= (aX+b-aE[X]-b)$

$= aX-aE[X]+b-b$

$= aX-aE[X]$

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

Taking expectation and squaring on both sides we get

$$
\begin{aligned}
E[Y-E(Y)]^2 &= E[a(X-E(X))]^2 \\
&= a^2 \{E[X-E[X]]^2\} \\
&= a^2 \{E[X^2-2XE[X]+(E[X])^2]\} \\
&= a^2 \{E[X^2]-2E[X]E[X]+(E[X])^2\} \\
&= a^2 \{E[X^2]-2(E[X])^2+(E[X])^2\} \\
&= a^2 \{E[X^2]-(E[X])^2\}
\end{aligned}
$$

$$V(aX+b)=a^2\, V(X)$$

Example: 4.1 Find the expectation ~~and variance~~ of the number on a die when thrown

Solution

Let X be a random variable representing the number on a die when thrown. Then X can take any one of the values 1,2,3,4,5,6 each with equal probability 1/6

x	1	2	3	4	5	6
P(X=x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$E(X)=\sum_{i=1}^{6} x_i P(X=x_i)=1\frac{1}{6}+2\frac{1}{6}+3\frac{1}{6}+4\frac{1}{6}+5\frac{1}{6}+6\frac{1}{6}$$

$$=\frac{1+2+3+4+5+6}{6}$$

$$E(X)=\frac{21}{6}$$

Example 4.2 If a pair of fair dice is tossed and X denotes the sum of the numbers on them, find the expectation of X.

Solution: Clearly X may be at least 2 and at most 12

X	2	3	4	5	6	7	8	9	10	11	12
P(X)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$E(X)=\sum_{i=2}^{12} x_i P(X=x_i)=2\frac{1}{36}+3\frac{2}{36}+4\frac{3}{36}+5\frac{4}{36}+6\frac{5}{36}+7\frac{6}{36}+8\frac{5}{36}$$

$$+9\frac{4}{36}+10\frac{3}{36}+11\frac{2}{36}+12\frac{1}{36}$$

$$=\frac{1}{36}[2+6+12+20+30+42+48+36+30+22+12]$$

$$E(X)=\frac{252}{36}=7$$

Example 4.3 If X be a random variable with the following probability distribution

X	-3	6	9
P(x)	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Find $E(X)$,$E(X^2)$ and $E(2X+1)^2$

Solution

$$E(X)=\sum x_i P(X=x_i)=-3\frac{1}{6}+6\frac{1}{2}+x\frac{1}{3}=\frac{-3+18+18}{6}=\frac{33}{6}=\frac{11}{2}$$

$$E(X)=\frac{11}{2}$$

$$E(X^2)=\sum x_i^2 P(X=x_i)=(-3)^2\frac{1}{6}+6^2\frac{1}{2}+9^2\frac{1}{3}=\frac{93}{2}$$

$$E(X^2)=\frac{93}{2}$$

$$\mathrm{E}(2\mathrm{X}+1)^2=\mathrm{E}\left[4\mathrm{X}^2+4\mathrm{X}+1\right]=\mathrm{E}\left[4\mathrm{X}^2\right]+\mathrm{E}[4\mathrm{X}]+\mathrm{E}[1]$$

$$=4\mathrm{E}\left[\mathrm{X}^2\right]+4\mathrm{E}[\mathrm{X}]+1$$

$$=4.\frac{93}{2}+4.\frac{11}{2}+1=209$$

$$\mathrm{E}(2\mathrm{X}+1)^2=209$$

Example: 4.4 In a continuous distribution the probability density function of X is $f(x)=\begin{cases}\frac{3}{4}x(2-x) & ,0<x<2\\ 0 & ,otherwise\end{cases}$ Find the expectation of the distribution.

Solution.

$$E(X) = \int_0^2 x\, f(x)\, dx = \int_0^2 x \cdot \frac{3}{4} x(2-x)\, dx$$

$$= \frac{3}{4}\int_0^2 x^2(2-x)\, dx = \frac{3}{4}\int_0^2 2x^2 - x^3\, dx$$

$$= \frac{3}{4}\left[2\frac{x^3}{3} - \frac{x^4}{4}\right]_0^2 = \frac{3}{4}\left[\left(2\frac{2^3}{3} - \frac{2^4}{4}\right) - \left(2\frac{0^3}{3} - \frac{0^4}{4}\right)\right]$$

$$= \frac{3}{4}\left[2\frac{8}{3} - \frac{16}{4}\right] = \frac{3}{4}\left[\frac{16}{3} - \frac{16}{4}\right]$$

$$= \frac{3}{4}\left[\frac{16}{3} - 4\right] = \frac{3}{4}\left[\frac{16-12}{3}\right] = \frac{3}{4}\left[\frac{4}{3}\right] = 1$$

$$E(X) = 1$$

4.3 Cauchy-Schwartz Inequality

If X and Y are random variables taking real values, then $[E(XY)]^2 \leq E(X^2)\, E(Y^2)$

Proof

Consider the expression $(X+tY)^2$ which is a function of real variable t. Since it is always non-negative for all real values of X,Y and t, it follows that

$$E(X+tY)^2 \geq 0 \ \forall t$$

$E(X^2+2XYt+t^2Y^2) \geq 0 \ \forall t$

$E(X^2)+2t\, E(XY)+t^2 E(Y^2) \geq 0 \ \forall t$

i.e., $\varphi(t) = At^2+Bt+C \geq 0 \ \forall t$

Treating as a quadratic in t , its roots will be real i.e., $t \geq 0$

where $A = E(Y^2)$, $B = 2\, E(XY)$ $C = E(X^2) \geq 0 \ \forall t$

Now $\varphi(t) \geq 0$ implies $B^2 - 4AC \leq 0$

$$\therefore 4E[(XY)] - 4E(X^2)\, E(Y^2) \leq 0$$

$$\Rightarrow [E(XY)]^2 \leq E(X^2)\, E(Y^2)$$

4.4. Conditional Expectation and Conditional Variance

Discrete Case: The conditional expectation of mean value of a continuous function g(X,Y) is given that $Y = y_j$ is defined by,

$$E\{g(X,Y)/Y=y_j\} = \sum_{i=1}^{\infty} \sum g(x_i, y_j) P(X = x_i / Y = y_j)$$

$$= \sum \frac{g(x_i, y_j) P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

(ie) $E\{g(X,Y)/Y=y_j\}$ is nothing but the expectation of function $g(x_i, y_j)$ of X with respect to the conditional distribution of X when $y = y_j$. In particular, the conditional expectation of a discrete random variable X is given $Y = y_j$

$$E\{X/Y=y_j\} = \sum x_i \quad P(X = x_i / Y = y_j)$$

The conditional variance of X given $y = y_j$ is given by

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

Continuous case

The conditional expectation of g(X, Y) on hypothesis Y = y is given by

$$E\{g(X,Y)/Y=y)\} = \int_{-\infty}^{\infty} g(x,y) f_{X/Y}(x/y)dx$$

$$= \int_{-\infty}^{\infty} g(x,y) \frac{f(x,y)}{f_Y(y)} dx$$

In particular, the conditional mean of x given y = y is defined as

$$E\{X/Y=y)\} = \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} dx$$

Similarly,

$$E\{Y/X=x)\} = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_X(x)} dy$$

The conditional variance of X defined as

$$V(X/Y=y) = E\left[(X - E(X/Y=Y))^2 / Y = y\right]$$

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

Theorem 4.1 The expected value of X is equal to the expectation of the conditional expectation of X given that is symbolically,

$$E(X) = E\{E(X/Y)\}$$

$$E\{E(X/Y)\} = E\left\{\sum_i x_i \; P(X = x_i / Y = y_j)\right\}$$

$$= E\left\{\sum_i x_i \; \frac{P(X = x_i \cap Y = y_j)}{P(y = y_j)}\right\}$$

$$= \sum_j \left\{\sum_i x_i \; \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right\} P(Y = y_j)$$

$$= \sum_i x_i \sum_j P(X = x_i \cap Y = y_j)$$

$$= \sum_i x_i \sum_j P(X = x_i \cap y = y_j)$$

$$= \sum_i x_i \; P(X = x_i) = E(X) = \cancel{E(X)}$$

$\Rightarrow$ E{E (X/Y)} = E(X)

Hence proved.

Theorem 4.2

The variance of X can be regarded as consisting of two parts the expectation of conditional variance and variance of conditional expectation symbolically

$$\text{Var}(X) = E[V(X/Y)] + V[E(X/Y)]$$

$$= E[V(X/Y)] + V[E(X/Y)]$$

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

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

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

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

[thm : 4.1]

$$= E\left\{\sum_i x_i^2 P(X = x_i / Y = y_j)\right\} - \{E(X)\}^2$$

$$= E\left\{\sum_i x_i^2 \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}\right\} - \{E(X)\}^2$$

$$= \sum_j \left[\left\{ \sum_i x_i^2 \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)} \right\} P(Y = y_j) \right] - [E(X)]^2$$

$$= \sum_i x_i^2 \sum_j P(X = x_i \cap Y = y_j) - [E(X)]^2$$

$$= \sum_i x_i^2 P(X = x_i) - [E(X)]^2$$

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

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

$\Rightarrow \text{Var}(X) = E[V(X/Y)] + V[E(X/Y)]$

Hence the theorem

EXAMPLE : 4.5 Let X and y be a two random variable each taking three values -1, 0, 1 having joint probability function of x and y

Y \ X	-1	0	1
-1	0	0.1	0.1
0	0.2	0.2	0.2
1	0	0.1	0.14

(i) Show that X and Y having different expectation.

(ii) Find the Variance of X and Y

(iii) Given that Y = 0 what is the conditional probability distribution of X.

(iv) Find the Var (Y/X = -1)

Solution

Y \ X	-1	0	1	P(Y=y)
-1	0	0.1	0.1	0.2
0	0.2	0.2	0.2	0.6
1	0	0.1	0.14	0.2
P(X = x)	0.2	0.4	0.4	1

(i) Expectation of X and Y are

$$E(X)=\sum x_i p_i=(-1)(0.2)+(0)(0.4)+(1)(0.4)=0.2$$

$$E(Y)=\sum y_j p_i=(-1)(0.2)+(0)(0.6)+(1)(0.2)=0$$

$$E(X)\neq E(Y)$$

$\therefore$ X and Y are having different expectation.

(ii) Variance of X and Y

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

$$E(X^2)=\sum x_i^2 P(X=x_i)=(-1)^2(0.2)+(0)^2(0.4)+(1)^2(0.4)$$

$$=0.2+0+0.4=0.6$$

$$E(X^2)=0.6$$

$$Var(X)=E(X^2)-(E(X))^2=0.6-(0.2)^2=0.6-0.04=0.56$$

$$Var(X)=0.56$$

$$Var(Y)=E(Y^2)-(E(Y))^2$$

$$E(Y^2)=\sum y_j^2 P(Y=y_j)=(-1)^2(0.2)+(0)^2(0.6)+(1)^2(0.2)$$

$$=0.2+0+0.2=0.4$$

$$E(Y^2)=0.4$$

$$Var(Y)=E(Y^2)-(E(Y))^2=0.4-(0)^2=0.4-0=0.4$$

$$Var(Y)=0.4$$

(iii) Conditional probability of X when Y = 0

$$P(X=-1/Y=0)=\frac{P(X=-1\cap Y=0)}{P(Y=0)}=\frac{0.2}{0.6}=\frac{1}{3}$$

$$P(X=0/Y=0)=\frac{P(X=0\cap Y=0)}{P(Y=0)}=\frac{0.2}{0.6}=\frac{1}{3}$$

$$P(X=1/Y=0)=\frac{P(X=1\cap Y=0)}{P(Y=0)}=\frac{0.2}{0.6}=\frac{1}{3}$$

(iv) V (Y|X = -1)

$$Var(Y/X=-1)=E(Y/X=-1)^2-\left[E(Y/X=-1)\right]^2$$

$$E(Y/X=-1)=\sum_y y\,P(Y=y/X=-1)$$

$$=(-1)(0)+(0)(0.2)+(1)(0)$$

$$E(Y/X=-1)\ =0$$

$$E(Y/X=-1)^2=\sum_y y^2\,P(Y=y/X=-1)$$

$$=(-1)^2(0)+(0)^2(0.2)+(1)^2(0)$$

$$E(Y/X=-1)^2\ =0$$

$$\therefore\ Var(Y/X=-1)=E(Y/X=-1)^2-\left[E(Y/X=-1)\right]^2$$

$$Var(Y/X=-1)=0-0=0$$

Example 4.6 Let $f(x,y)\ =\ \begin{cases}8xy,\ 0<x<y<1\\ 0,\quad elsewhere\end{cases}$.

Find (a) E(Y|X= x) Var (Y|X=x)

Solution : (a)

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)\,dy=\int_x^1 8xy\,dy=8x\int_x^1 y\,dy=8x\left[\frac{y^2}{2}\right]_x^1$$

$$=8x\left[\frac{1^2}{2}-\frac{x^2}{2}\right]=8x\left[\frac{1^2-x^2}{2}\right]$$

$$f_X(x)=4x(1-x^2),\ o<x<1$$

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,y)\,dx=\int_0^y 8xy\,dx=8y\int_0^y x\,dx=8y\left[\frac{x^2}{2}\right]_0^y$$

$$=8y\left[\frac{y^2}{2}-\frac{0^2}{2}\right]=8y\left[\frac{y^2}{2}\right]$$

$$f_Y(y)=4y^3,\ o<y<1$$

$$f_{X/Y}(x/y) = \frac{f(x,y)}{f_Y(y)} = \frac{8xy}{4y^3}$$

$$f_{X/Y}(x/y) = \frac{2x}{y^2}$$

$$f_{Y/X}(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{8xy}{4x(1-x^2)}$$

$$f_{Y/X}(y/x) = \frac{2y}{(1-x^2)}$$

(b) Var(Y/X=x) = E(Y^2/X=x)-{E(Y/X=x}^{2}

$$E(Y/X=x) = \int_x^1 y\, f_{Y/X}(y/x)dy = \int_x^1 y\, \frac{2y}{(1-x^2)}dy$$

$$= \frac{2}{(1-x^2)}\int_x^1 y^2\, dy = \frac{2}{(1-x^2)}\left[\frac{y^3}{3}\right]_x^1$$

$$= \frac{2}{(1-x^2)}\left[\frac{1^3}{3} - \frac{x^3}{3}\right] = \frac{2}{(1-x^2)}\left[\frac{1^3 - x^3}{3}\right]$$

$$E(Y/X=x) = \frac{2}{3}\left[\frac{1-x^3}{1-x^2}\right]$$

$$E(Y^2/X=x) = \int_x^1 y^2\, f_{Y/X}(y/x)dy = \int_x^1 y^2\, \frac{2y}{(1-x^2)}dy$$

$$= \frac{2}{(1-x^2)}\int_x^1 y^3\, dy = \frac{2}{(1-x^2)}\left[\frac{y^4}{4}\right]_x^1$$

$$= \frac{2}{(1-x^2)}\left[\frac{1^4}{4} - \frac{x^4}{4}\right] = \frac{2}{(1-x^2)}\left[\frac{1^4 - x^4}{4}\right]$$

$$E(Y^2/X=x) = \frac{1+x^2}{2}$$

$$Var(Y/X=x) = E(Y^2/X=x) - (E(Y/X=x))^2$$

$$= \left[\frac{1+x^2}{2}\right] - \left(\frac{2}{3}\left[\frac{1-x^3}{1-x^2}\right]\right)^2$$

$$Var(Y/X=x) = \frac{1+x^2}{2} - 9\left(\frac{1-x^3}{1-x^2}\right)^2$$

4.5 MOMENT GENERATING FUNCTION

The Moment Generating Function (M.G.F) of a random variable X defined as

$$M_X(t) = E(e^{tX}) = \begin{cases} \int e^{tx} f(x)dx & \text{for continuous probability distributions} \\ \sum_x e^{tx} p(x=x) & \text{for discrete probability distributions} \end{cases}$$

$$M_X(t) = E(e^{tX}) = \int e^{tx} f(x)dx$$

$$\therefore M_X(t) = E(e^{tX}) = E\left(1 + tX + \frac{t^2 X^2}{2!} + \dots + \frac{t^r X^r}{r!} + \dots\right)$$

$$= 1 + t\,E(X) + \frac{t^2}{2!}E(X^2) + \dots + \boxed{\frac{t^r}{r!}E(X^r)} + \dots$$

$$= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \dots + \frac{t^r}{r!}\mu_r' + \dots$$

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r' \qquad 0! = 1$$

Where $\mu_r' = E(X^r) = \begin{cases} \int x^r f(x)\, dx & \textit{for continuous distribution} \\ \sum_x x^r p(x) & \textit{for discrete distribution} \end{cases}$

is the rth moment of X about origin. Thus the coefficient of $\frac{t^r}{r!}$ in $M_X(t)$ gives μ_r' (about origin). Since $M_X(t)$ generates moments, it is known as moment generating function. Differentiating moment generating function w.r. to 't' 'r' time and put t = 0 we get,

$$\left[\frac{d^r}{dt^r}M_X(t)\right]_{t=0} = \mu_r'$$

put r = 1

$$\mu_1' = \left[\frac{d}{dt}M_X(t)\right]_{t=0} = E(X) = Mean$$

put r = 2

$$\mu_2' = \left[\frac{d^2}{dt^2}M_X(t)\right]_{t=0} = E(X^2)$$

$$Variance = \mu_2' - (\mu_1')^2 = E(X^2) - (E(X))^2$$

4.5.1 Properties of Moment generating function:

Property 1

$M_{cX}(t) = E[e^{tcX}]$, c is a constant.

By definition

L.H.S. $M_{cx}(t) = E[e^{tcx}]$

R.H.S. $M_x(ct) = E[e^{ctx}]$ $\quad = \quad$ L.H.S

$$\therefore M_{cX}(t) = E[e^{tcX}]$$

Property 2

The moment generating function of the sum of a number of random variables is equal to the product of their respective moment generating function.

$$M_{(X_1+X_2+X_3+X_4+\cdots+X_n)}(t) = M_{X_1}(t)M_{X_2}(t)M_{X_3}(t)\cdots M_{X_n}(t)$$

Proof

$$M_{(X_1+X_2+X_3+X_4+\cdots X_n)}(t) = E\left[e^{t(X_1+X_2+\cdots X_n)}\right]$$

$$= E\left[e^{tX_1}e^{tX_2}\ldots e^{tX_n}\right]$$

$$= E\left[e^{tX_1}\right]E\left[e^{tX_2}\right]\ldots E\left[e^{tX_n}\right]$$

$$= M_{X_1}(t)M_{X_2}(t)M_{X_3}(t)\ldots M_{X_n}(t)$$

Property 3 Effect of change of origin and scale on MGF.

Let us transform X to the new variable U by changing both the origin and scale in X as follows $U = \dfrac{X-a}{h}$ where a and h are constants

Moment generating function about U about origin is given by

$$M_U(t) = E(e^{tU}) = E\left(e^{t\left(\frac{X-a}{h}\right)}\right)$$

$$= E\left(e^{\left(\frac{tX-at}{h}\right)}\right) = E\left(e^{\left(\frac{tX}{h}-\frac{at}{h}\right)}\right)$$

$$= E\left(e^{\frac{tX}{h}}e^{\frac{-at}{h}}\right) = e^{\frac{-at}{h}}E\left(e^{\frac{tX}{h}}\right)$$

$$M_U(t) = e^{\frac{-at}{h}}M_X(t/h)$$

Where $M_X(t)$ is the M.G.F of X about orgin.

4.5.2 Limitations of Moment Generating Function

1. ***A random variable X may not have moments although its moment generating function exists.***

Consider a discrete random variable X with probability density function is

$$f(x)=\frac{1}{x(x+1)} \text{ for } x=1,2,3,\ldots \text{and '0' otherwise}$$

$$E(X)=\sum_{x=1}^{\infty} x\, f(x)=\sum_{x=1}^{\infty}\frac{x}{x(x+1)}$$

$$=\sum_{x=1}^{\infty}\frac{1}{(x+1)}=\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots$$

$$=\left\{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots\right\}-1$$

$$E(X)=\sum_{x=1}^{\infty}\frac{1}{x}-1$$

Since $\sum_{1}^{\infty}\frac{1}{x}$ is divergent series, E(X) does not exists and consequently no moment of X exists,how ever , the mgf of X is given by

$$M_X(t)=\sum_{x=1}^{\infty} e^{tx} f(x)=\sum_{x=1}^{\infty} e^{tx}\frac{1}{x(x+1)}$$

Let $z=e^t$

$$M_X(t)=\sum_{x=1}^{\infty}\frac{z^x}{x(x+1)}=\frac{z^1}{1.2}+\frac{z^2}{2.3}+\frac{z^3}{3.4}+\cdots$$

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

$$=\left(z-\frac{z}{2}\right)+\left(\frac{z^2}{2}-\frac{z^2}{3}\right)+\left(\frac{z^3}{3}-\frac{z^3}{4}\right)+\cdots$$

$$=\left(z+\frac{z^2}{2}+\frac{z^3}{3}+\cdots\right)-\left(\frac{z}{2}+\frac{z^2}{3}+\frac{z^3}{4}+\cdots\right)$$

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

$$=-\log(1-z)-\left(\left(1+\frac{z}{2}+\frac{z^2}{3}+\frac{z^3}{4}+\cdots\right)-1\right)$$

$$= -\log(1-z) - \left(\frac{z}{z}\left(1+\frac{z}{2}+\frac{z^2}{3}+\frac{z^3}{4}+\cdots\right)-1\right)$$

$$= -\log(1-z) - \left(\frac{1}{z}\left(z+\frac{z^2}{2}+\frac{z^3}{3}+\frac{z^4}{4}+\cdots\right)-1\right)$$

$$= -\log(1-z) - \frac{1}{z}\left(z+\frac{z^2}{2}+\frac{z^3}{3}+\frac{z^4}{4}+\cdots\right)+1$$

$$= -\log(1-z) - \frac{1}{z}(-\log(1-z))+1$$

$$= -\log(1-z) + \frac{1}{z}\log(1-z)+1$$

$$\cancel{\text{for } |z|<1 \Rightarrow |e^t|<1 \Rightarrow t<0} \qquad e^0 = 1$$

$$= 1+\left(\frac{1}{z}-1\right)\log(1-z)$$

$$= 1+\left(\frac{1}{e^t}-1\right)\log(1-e^t) = 1+\left(e^{-t}-1\right)\log(1-e^t), t<0$$

So that $M_X(t) = 1$ for $t=0$, Hence $M_X(t)$ exists for $t\le 0$.

2. ***A random variable X can have moment generating function along with some or all moments, yet the but m.g.f does not generate the moments.***

Let consider a discrete random variable X with probability functions

$$P(X=2^x) = \frac{e^{-1}}{x!} \text{ for } x = 0,1,2,\ldots \quad \text{Then}$$

$$E(X^r) = \sum_{x=0}^{\infty}(2^x)^r P(X=2^x) = \sum_{x=0}^{\infty}(2^x)^r \frac{e^{-1}}{x!}$$

$$= e^{-1}\sum_{x=0}^{\infty}\frac{(2^r)^x}{x!} = e^{-1}\left[1+\frac{2^r}{1!}+\frac{(2^r)^2}{2!}+\cdots\right] = e^{-1}e^{2^r}$$

$$E(X^r) = e^{2^r-1}$$

Hence all the moments of X exists. The m.g.f of X, if it exists, is given by

$$M_X(t) = \sum_{x=0}^{\infty} e^{t\cdot 2^x}\left(\frac{e^{-1}}{x!}\right) = e^{-1}\sum_{x=0}^{\infty} e^{t\cdot 2^x}\left(\frac{1}{x!}\right)$$

By D' Alembert's ratio test the series on the RHS is convergent for $t\le 0$ and diverges for $t>0$. Hence $M_X(t)$ cannot be differentiated at $t=0$ and has no Maclurin's expansion and consequently it does not generate moments.

3. *A random variable X can have some or all moments, but m.g.f does not exist except perhaps at one point.*

Let consider X be a random variable with probability function

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

The distribution being symmetric, moments of odd order about origin vanish

i.e., $\mu_{2r+1}=0 \Rightarrow E(X^{2r+1})=0$

Now, $E(X^{2r})=\sum_{x=0}^{\infty}(\pm 2^{x})^{2r}\dfrac{e^{-1}}{2x!}=e^{-1}\sum_{x=0}^{\infty}\dfrac{\left(2^{x}\right)^{2r}}{x!}=e^{(2^{2r}-1)}$

Thus all the moments of X exists. The m.g.f of X, if it exists, is given by

$$M_{X}(t)=\sum_{x=0}^{\infty}\left\{\left(e^{t.2^{X}}+e^{-t.2^{X}}\right)\frac{1}{2ex!}\right\}=e^{-1}\sum_{x=0}^{\infty}\left\{\frac{Cosh(t2^{x})}{x!}\right\}$$

Which is only convergent for t = 0. Hence m.g.f of X does not exists at t=0.

Example 4.7 Let the random variable X assume the value of r with probability law $P(X=r)=q^{r-1}.p$. r = 1, 2, 3. Find the moment generating function and hence find its mean and variance.

Solution

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

$$=\sum_{r=1}^{\infty}e^{tr}\,p(x=r)$$

$$=\sum_{r=1}^{\infty}e^{tr}\,q^{r-1}.p$$

$$=\sum_{r=1}^{\infty}e^{tr}\,q^{r}q^{-1}.p$$

$$=\frac{p}{q}\sum_{r=1}^{\infty}(qe^{t})^{r}$$

$$=\frac{p}{q}\sum_{r=1}^{\infty}(qe^{t})^{r}$$

$$=\frac{p}{q}(qe^{t})\left[1+(qe^{t})+(qe^{t})^{2}+\ldots\right]$$

$$=p\,e^{t}(1-qe^{t})^{-1}$$

$$M_X(t) = \frac{Pe^t}{(1-qe^t)}$$

Mean $E(X) = \left[\frac{d}{dt} M_X^{(t)}\right]_{t=0}$

$$\frac{d}{dt} M_X(t) = \frac{d}{dt} \quad \frac{pe^t}{(1-qe^t)}$$

$$= p\frac{d}{dt} e^t (1-qe^t)^{-1}$$

$$= p\left[e^t(-1)(1-qe^t)^{-2}(-qe^t) + e^t(1-qe^t)^{-1}\right]$$

$$= p\left[\frac{qe^{2t}}{(1-qe^t)^2} + \frac{e^t}{(1-qe^t)}\right]$$

$$= p\left[\frac{qe^{2t} + e^t(1-qe^t)}{(1-qe^t)^2}\right]$$

$$= p\left[\frac{qe^{2t} + e^t - qe^t e^t}{(1-qe^t)^2}\right] = p\left[\frac{qe^{2t} + e^t - qe^{2t}}{(1-qe^t)^2}\right]$$

$$= \frac{pe^t}{(1-qe^t)^2}$$

$$E(X) = \left[\frac{d}{dt} M_X(t)\right]_{t=0} = \frac{pe^0}{(1-qe^0)^2} = \frac{p}{(1-q)^2} \frac{p}{p^2}$$

$$E(X) = \frac{1}{p}$$

$$E(X^2) = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$\left[\frac{d^2}{dt^2} M_X(t)\right] = \frac{d^2}{dt^2}\left(\frac{pe^t}{(1-qe^t)}\right)$$

$$= \frac{d}{dt}\left[\frac{pe^t}{(1-qe^t)^2}\right]$$

$$= p\frac{d}{dt} \quad e^t(1-qe^t)^{-2}$$

$$= p\left[e^t(-2)(1-qe^t)^{-2-1}(-qe^t) + e^t(1-qe^t)^{-2}\right]$$

$$= p\left[2qe^t . e^t(1-qe^t)^{-3} + e^t(1-qe^t)^{-2}\right]$$

$$= P\left[\frac{2qe^{2t}}{(1-qe^{t})^{3}}+\frac{e^{t}}{(1-qe^{t})^{2}}\right]$$

$$= P\left[\frac{2qe^{2t}+e^{t}(1-qe^{t})}{(1-qe^{t})^{3}}\right]$$

$$E(X^{2}) = \left[\frac{d^{2}}{dt^{2}}M_{x}(t)\right]_{t=0} \quad P\left[\frac{2qe^{0}+e^{0}(1-qe^{0})}{(1-qe^{0})^{3}}\right]$$

$$= P\left[\frac{2q+1-q}{(1-q)^{3}}\right]$$

$$= P\left[\frac{q+1}{(1-q)^{3}}\right] = P\left[\frac{q+1}{p^{3}}\right]$$

$$E(X^{2}) = \frac{(q+1)}{p^{2}}$$

$$Var(X) = E(X^{2}) - (E(X))^{2}$$

$$= \left(\frac{(q+1)}{p^{2}}\right) - \left(\frac{1}{p}\right)^{2} = \frac{q+1}{p^{2}} - \frac{1}{p^{2}} = \frac{q+1-1}{p^{2}}$$

$$Var(x) = \frac{q}{p^{2}}$$

Example 4.8 A random variable X has probability function $p(x) = \frac{1}{2^{x}}$ $x = 1,2,3,\ldots$ Find the moment generating function, mean and variance.

Solution:

$$M_{X}(t) = E(e^{tx}) = \sum_{x=1}^{\infty} e^{tx}p(x) = \sum_{x=1}^{\infty} e^{tx}\frac{1}{2^{x}} = \sum_{x=1}^{\infty}\frac{e^{tx}}{2^{x}} = \sum_{x=1}^{\infty}\left(\frac{e^{t}}{2}\right)^{x}$$

$$= \left(\frac{e^{t}}{2}\right)^{1} + \left(\frac{e^{t}}{2}\right)^{2} + \left(\frac{e^{t}}{2}\right)^{3} + \left(\frac{e^{t}}{2}\right)^{4} + \cdots$$

$$= \frac{e^{t}}{2}\left[1 + \left(\frac{e^{t}}{2}\right) + \left(\frac{e^{t}}{2}\right)^{2} + \left(\frac{e^{t}}{2}\right)^{3} + \left(\frac{e^{t}}{2}\right)^{4} + \cdots\right]$$

$$=\frac{e^t}{2}\left[1+\left(\frac{e^t}{2}\right)+\left(\frac{e^t}{2}\right)^2+\left(\frac{e^t}{2}\right)^3+\left(\frac{e^t}{2}\right)^4+\cdots\right]$$

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

$$M_X(t)=\left[\frac{e^t}{2-e^t}\right]$$

Mean

$$E(X)=\left[\frac{d}{dt}M_X(t)\right]_{t=0}$$

$$\frac{d}{dt}M_X(t)=\frac{d}{dt}\frac{e^t}{\left(2-e^t\right)}=\left[\frac{\left(2-e^t\right)e^t-e^t\left(-e^t\right)}{\left(2-e^t\right)^2}\right]=\left[\frac{2e^t-e^te^t+e^te^t}{\left(2-e^t\right)^2}\right]=\left[\frac{2e^t}{\left(2-e^t\right)^2}\right]$$

$$E(X)=\left[\frac{d}{dt}M_X(t)\right]_{t=0}=\left[\frac{2e^t}{\left(2-e^t\right)^2}\right]_{t=0}=\left[\frac{2e^0}{\left(2-e^0\right)^2}\right]=\left[\frac{2}{(2-1)}\right]=2$$

$$E(X)=2$$

$$Variance\ =E(X^2)-\left(E(X)\right)^2$$

$$E(X^2)=\left[\frac{d^2}{dt^2}M_X(t)\right]_{t=0}$$

$$\left[\frac{d^2}{dt^2}M_X(t)\right]=\left[\frac{d^2}{dt^2}\frac{e^t}{2-e^t}\right]=\left[\frac{d}{dt}\frac{2e^t}{\left(2-e^t\right)^2}\right]=\left[\frac{\left(2-e^t\right)^2(2e^t)-4e^t\left(2-e^t\right)\left(-e^t\right)}{\left(2-e^t\right)^4}\right]$$

$$E(X^2)=\left[\frac{d^2}{dt^2}M_X(t)\right]_{t=0}=\left[\frac{\left(2-e^t\right)^2(2e^t)-4e^t\left(2-e^t\right)\left(-e^t\right)}{\left(2-e^t\right)^4}\right]_{t=0}$$

$$=\left[\frac{\left(2-e^0\right)^2(2e^0)-4e^0\left(2-e^0\right)\left(-e^0\right)}{\left(2-e^0\right)^4}\right]=\left[\frac{(2-1)2+4(2-1)(1)}{(2-1)^4}\right]=\frac{2+4}{1}$$

$$E(X^2)=6$$

$$Variance\ =E(X^2)-\left(E(X)\right)^2=6-(2)^2=6-4=2$$

Example 4.9 Find the m.g.f of the random variable X having p.d.f is defined as

$$f(x)=\begin{cases} x & for\ 0\le x\le 1 \\ 2-x & for\ 1\le x\le 2 \\ 0 & otherwise \end{cases}$$

Solution:

$$M_X(t)=\int_{-\infty}^{\infty} e^{tx} f(x)\,dx=\int_0^1 e^{tx}\,x\,dx+\int_1^2 e^{tx}\,(2-x)\,dx$$

$$=\int_0^1 x\,e^{tx}\,dx+\int_1^2 (2-x)\,e^{tx}\,dx$$

$$=\left[x\left(\frac{e^{tx}}{t}\right)-\left(\frac{e^{tx}}{t^2}\right)\right]_0^1+\left[(2-x)\left(\frac{e^{tx}}{t}\right)-(-1)\left(\frac{e^{tx}}{t^2}\right)\right]_1^2$$

$$=\left[\left((1)\left(\frac{e^{t(1)}}{t}\right)-\left(\frac{e^{t(1)}}{t^2}\right)\right)-\left((0)\left(\frac{e^{t(0)}}{t}\right)-\left(\frac{e^{t(0)}}{t^2}\right)\right)\right]$$

$$+\left[\left((2-2)\left(\frac{e^{t(2)}}{t}\right)-(-1)\left(\frac{e^{t(2)}}{t^2}\right)\right)-\left((2-1)\left(\frac{e^{t(1)}}{t}\right)-(-1)\left(\frac{e^{t(1)}}{t^2}\right)\right)\right]$$

$$=\left[\left(\frac{e^t}{t}-\frac{e^t}{t^2}\right)+\left(\frac{e^0}{t^2}\right)\right]+\left[\left(0+\frac{e^{2t}}{t^2}\right)-\left(\frac{e^t}{t}+\frac{e^t}{t^2}\right)\right]=\left[\frac{e^t}{t}-\frac{e^t}{t^2}+\frac{1}{t^2}\right]+\left[\frac{e^{2t}}{t^2}-\frac{e^t}{t}-\frac{e^t}{t^2}\right]$$

$$=\frac{e^t}{t}-\frac{e^t}{t^2}+\frac{1}{t^2}+\frac{e^{2t}}{t^2}-\frac{e^t}{t}-\frac{e^t}{t^2}$$

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

$$M_X(t)=\left(\frac{1-e^t}{t}\right)^2$$

4.6 CUMULANTS

Cummlants generating function K(t) is defined as $K_X(t) = \log_e M_X(t)$

Provided the right hand side can be exoanded as a convergent series in power of t or If the logarithm of the m.g.f of a distribution can be expanded as a convergent series in powers of t viz.,

$$K_X(t) = k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + \cdots + k_r \frac{t^r}{r!} + \cdots = \log M_X(t)$$

$$= \log\left(1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots + \frac{t^r}{r!}\mu_r' + \ldots\right)$$

Then the coefficients $k_1, k_2, \ldots$ Are called the first, second cumulant of the distribution and $K_X(t)$ is called the cumulative function.

Differentiating r times both sides with respect to t and putting t = 0 and we have

$$k_r = \left[\frac{d^r}{dt^r} \log M_X(t)\right]_{t=0} = \left[\frac{d^r}{dt^r} K_X(t)\right]_{t=0}$$

4.6.1 Properties of Cumulants

Property 1 : Additive Property

The r^{th} cumulant of the sum of the independent random variables is equal to the sum of the r^{th} cumulants of the individual variables. Symbolically

$$k_r(X_1+X_2+X_3+\ldots+X_n) = k_r(X_1)+kr(X_2)+kr(X3)+\ldots+kr(Xn)$$

where X_i, i=1,2,…,n are independent random variables.

Proof

Since X_i, i=1,2,…,n are independent,

$$M_{X_1+X_2+X_3+\cdots+X_n}(t) = M_{X_1}(t) M_{X_2}(t) M_{X_3}(t) \cdots M_{X_n}(t)$$

Taking logarithm of each side

$$K_{X_1+X_2+X_3+\cdots+X_n}(t) = K_{X_1}(t) + K_{X_2}(t) + K_{X_3}(t) + \cdots + K_{X_n}(t)$$

Differentiating with respect to 'r' times and put t= 0 we get

$$\left[\frac{d^r}{dt^r} K_{X_1+X_2+X_3+\cdots+X_n}(t)\right]_{t=0} = \left[\frac{d^r}{dt^r} K_{X_1}(t)\right]_{t=0} + \left[\frac{d^r}{dt^r} K_{X_2}(t)\right]_{t=0} + \cdots + \left[\frac{d^r}{dt^r} K_{X_n}(t)\right]_{t=0}$$

$$\therefore k_r(X_1+X_2+X_3+\ldots+X_n) = k_r(X_1)+kr(X_2)+kr(X_3)+\ldots+kr(X_n)$$

Property 2: Effect of change of Origin and scale on Cumulants

$$\text{Let } U = \frac{X-a}{h} \text{ then}$$

$$M_U(t) = e^{\frac{-at}{h}} M_X(t/h)$$

Taking logarithm on both sides

$$\log[M_U(t)] = \log\left[e^{\frac{-at}{h}} M_X(t/h) \right]$$

$$K_U(t) = \log M_U(t) = \frac{-at}{h} + K_X(t/h)$$

$$k_1' t + k_2' \frac{t^2}{2!} + k_3' \frac{t^3}{3!} + \cdots + k_r' \frac{t^r}{r!} + \cdots = \frac{-at}{h} + k_1(t/h) + k_2 \frac{(t/h)^2}{2!} + \cdots + k_r \frac{(t/h)^r}{r!}$$

Where k_r' and k_r are the r^{th} cumulants of U and X respectively. Comparing coefficients, we get $k_1' = \frac{k_1 - a}{h}$ and $k_r' = \frac{k_r}{h^r}; r = 2,3,\ldots..$

Thus except the first cumulant, all the cumulants are independent of change of origin. But the cumulants are not invariant of change of scale as the r^{th} cumulant of U is $(1/h^r)$ times the r^{th} cumulant of the distribution of X.

4.7 CHARACTERISTIC FUNCTION

In some case moment generating function does not exists. The characteristic function defined as

$$\phi_x(t) = E(e^{itX}) = \begin{cases} \int e^{itx} f(x)\, dx & \text{for continuous probability distribution} \\ \sum_x e^{itx} p(x) & \text{for discrete probability distribution} \end{cases}$$

4.7.1 Properties of characteristic function

Property 1

For all real t, we have

(i) $\phi(0) = \int_{-\infty}^{\infty} dF(x) = 1$

(ii) $|\phi(t)| \leq 1 = \phi(0)$

Property 2

ϕ (t) is continuous everywhere, i.e., ϕ (t) is continuous function of 't' in $(-\infty,\infty)$. Rather ϕ (t) is uniformly continuous in 't'.

Proof

$$\text{For } h \neq 0 \, |\phi_X(t+h) - \phi_X(t)| = \left| \int_{-\infty}^{\infty} [e^{i(t+h)x} - e^{itx}] dF(x) \right|$$

$$\leq \int_{-\infty}^{\infty} |e^{itx}(e^{ihx} - 1)| \, dF(x) = \int_{-\infty}^{\infty} |e^{ihx} - 1| \, dF(x)$$

The last integral does not depend on 't'. If it tends to zero as h → 0 then ϕ_x (t) is uniformly continuous in 't'

Now $|e^{ihx} - 1| \leq |e^{ihx}| + |1| \leq 1 + 1 = 2$

$$\therefore \int_{-\infty}^{\infty} |e^{ihx} - 1| \, dF(x) \leq 2 \int_{-\infty}^{\infty} |dF(x)| = 2$$

Hence by Dominated convergence theorem (D.C.T) taking the limit inside the integral sign.

$$\lim_{h \to 0} |\phi_X(t+h) - \phi_X(t)| \leq \int_{-\infty}^{\infty} \lim_{h \to 0} |e^{ihx} - 1| \, dF(x) = 0$$

$$\Rightarrow \lim_{h \to 0} \phi_X(t+h) = \phi_X(t), \forall t$$

Hence ϕ_x(t) is uniformly continuous in 't'.

Property 3

ϕ_X(-t) and ϕ_x(t) are conjugate functions.

$\phi_X(-t) = \overline{\phi_X}(t)$, where a is the complex conjugate of 'a'.

Proof

$\phi_x(t) = E(e^{itx}) = E\,[\text{Cos } t_x + i \text{ Sin} t_x]$

$\overline{\phi_X(t)} = E(\text{Cos } tX - i \text{ Sin} t X)$

$= E\{\text{Cos } (-t) X + i \text{ Sin } (-t) X\}$

$= E(e^{-itx}) = \phi_x(-t)$

Property 4

If the distribution function of a r.v.x is symmetrical about zero, ie if

$$1 - F(x) = F(-x)$$

$$\Rightarrow F(-x) = f(x)$$

F = φ

Proof

By the definition the $\phi_x(t)$ is real valued and even function of t

$$\phi_x(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx \qquad \text{put } x = -y$$

$$= \int_{-\infty}^{\infty} e^{-ity} f(-y)dy$$

$$= \int_{-\infty}^{\infty} e^{-ity} f(-y)dy \qquad (f(-y) = f(y))$$

$$= \phi_x(-t)$$

$\Rightarrow$ $\phi_x(-t)$ is an even function of 't'

Property 5

If X is some r.v with characteristic function $\phi_x(t)$ and u $\mu_r' = E(X^r)$ exists.

$$\mu_r' = (-i)^r \left|\frac{\partial^r}{\partial t^r}\phi_x(t)\right|_{t=0}$$

Proof

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx$$

Differentiating (under the integral sign)'r' times w.r. to t, we get

$$\frac{\partial^r}{\partial t^r}\phi(t) = \int_{-\infty}^{\infty} (ix)^r \; e^{itx} f(x)dx$$

$$= \int_{-\infty}^{\infty} i^r x^r \; e^{itx} f(x)dx$$

$$= (i)^r \int_{-\infty}^{\infty} x^r \; e^{itx} f(x)dx$$

$$\therefore \left|\frac{\partial^r}{\partial t^r}\phi_x(t)\right|_{t=0} = (i)^r \left|\int_{-\infty}^{\infty} x^r \; e^{itx} f(x)dx\right|_{t=0}$$

$$= (i)^r \int_{-\infty}^{\infty} x^r f(x)dx$$

$$= (i)^r E(x^r) = i^r \mu_r'$$

Hence

$$\mu_r' = \left(\frac{1}{i}\right)^r \left|\frac{\partial^r}{\partial t^r}\phi_X(t)\right|_{t=0} = (-i)^r \left|\frac{\partial^r}{\partial t^r}\phi(t)\right|_{t=0}$$

Property 6

$\phi_{cx}(t) = \phi_x(ct)$ c is constant.

Property 7

If X_1 and X_2 are independent random variables, then,

$\phi_{X_1+X_2}(t) = \phi_{X_1}(t) + \phi_{X_2}(t)$

Property 8 Effect of change of origin and scale on characteristic Function.

If $U = \frac{x-a}{h}$, a and h being constants, then

$$\phi_u(t) = e^{-iat/h}\, \phi_X\left(t/h\right)$$

In particular we take a = E(x) = μ(say) and h = σ_x = σ, then the characteristic function of the standard variate.

$Z = \frac{X - E(X)}{\delta_X} = \frac{X-\mu}{\delta}$ is given by

$\phi_z(t) = e^{-i\mu t/\sigma}\phi(t/\sigma)$

Example: Find the characteristic function of the Poisson distribution

Solution:

The probability mass function of a Poisson distribution is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}; x = 0,1,2,3,.....$$

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

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{it})^x}{x!} = e^{-\lambda}\left[\frac{(\lambda e^{it})^0}{0!} + \frac{(\lambda e^{it})^1}{1!} + \frac{(\lambda e^{it})^2}{2!} + \cdots\right]$$

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

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

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

Example 4.10 Find the characteristic function of a pdf $f(x) = \frac{\alpha}{2} e^{-\alpha|x|}, -\infty < x < \infty$

Solution Let

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx = \int_{-\infty}^{\infty} e^{itx} \frac{\alpha}{2} e^{-\alpha|x|}\,dx$$

$$= \frac{\alpha}{2}\int_{-\infty}^{\infty} e^{itx} e^{-\alpha|x|}\,dx = \frac{\alpha}{2}\left[\int_{-\infty}^{0} e^{itx} e^{-\alpha(-x)}\,dx + \int_{0}^{\infty} e^{itx} e^{-\alpha(x)}\,dx\right]$$

$$= \frac{\alpha}{2}\left[\int_{-\infty}^{0} e^{itx} e^{\alpha x}\,dx + \int_{0}^{\infty} e^{itx} e^{-\alpha x}\,dx\right] = \frac{\alpha}{2}\left[\int_{-\infty}^{0} e^{itx+\alpha x}\,dx + \int_{0}^{\infty} e^{itx-\alpha x}\,dx\right]$$

$$= \frac{\alpha}{2}\left[\int_{-\infty}^{0} e^{(\alpha+it)x}\,dx + \int_{0}^{\infty} e^{-(\alpha-it)x}\,dx\right] = \frac{\alpha}{2}\left[\left(\frac{e^{(\alpha+it)x}}{(\alpha+it)}\right)_{-\infty}^{0} + \left(\frac{e^{-(\alpha-it)x}}{-(\alpha-it)}\right)_{0}^{\infty}\right]$$

$$= \frac{\alpha}{2}\left[\left(\left(\frac{e^{(\alpha+it)(0)}}{(\alpha+it)}\right) - \left(\frac{e^{(\alpha+it)(-\infty)}}{(\alpha+it)}\right)\right) + \left(\left(\frac{e^{-(\alpha-it)\infty}}{-(\alpha-it)}\right) - \left(\frac{e^{-(\alpha-it)0}}{-(\alpha-it)}\right)\right)\right]$$

$$= \frac{\alpha}{2}\left[\int_{-\infty}^{0} e^{(\alpha+it)x}\,dx + \int_{0}^{\infty} e^{-(\alpha-it)x}\,dx\right] = \frac{\alpha}{2}\left[\left(\frac{e^{(\alpha+it)x}}{(\alpha+it)}\right)_{-\infty}^{0} + \left(\frac{e^{-(\alpha-it)x}}{-(\alpha-it)}\right)_{0}^{\infty}\right]$$

$$= \frac{\alpha}{2}\left[\left(\left(\frac{e^{(\alpha+it)(0)}}{(\alpha+it)}\right) - \left(\frac{e^{(\alpha+it)(-\infty)}}{(\alpha+it)}\right)\right) + \left(\left(\frac{e^{-(\alpha-it)\infty}}{-(\alpha-it)}\right) - \left(\frac{e^{-(\alpha-it)0}}{-(\alpha-it)}\right)\right)\right]$$

$$= \frac{\alpha}{2}\left[\left(\left(\frac{e^{0}}{(\alpha+it)}\right)-(0)\right)+\left((0)-\left(\frac{e^{0}}{-(\alpha-it)}\right)\right)\right] = \frac{\alpha}{2}\left[\frac{1}{(\alpha+it)}+\frac{1}{(\alpha-it)}\right]$$

$$= \frac{\alpha}{2}\left[\frac{(\alpha-it)+(\alpha+it)}{(\alpha+it)(\alpha-it)}\right] = \frac{\alpha}{2}\left[\frac{\alpha-it+\alpha+it}{(\alpha^2-(it)^2)}\right] = \frac{\alpha}{2}\left[\frac{2\alpha}{\left(\alpha^2-\left(i^2t^2\right)\right)}\right] = \frac{\alpha}{2}\left[\frac{2\alpha}{\left(\alpha^2-(-1)t^2\right)}\right]$$

$$\varphi_X(t) = \left[\frac{\alpha^2}{\left(\alpha^2+t^2\right)}\right]$$

Example 4.11 Show that the distribution which the characteristic function $e^{-|t|}$ has the density function is $f(x)=\frac{1}{\pi}\frac{dx}{1+x^2}\quad -\infty \le x \le \infty$

Solution

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\phi_X(t)e^{-itx}dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-|t|}e^{-itx}dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-|t|}(\cos tx - i\sin tx)dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-|t|}(\cos tx)dt - i\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-|t|}(\sin tx)dt$$

$$= \frac{1}{2\pi}\int_{0}^{\infty}e^{-|t|}(\cos tx)dt = \frac{1}{2\pi}\int_{0}^{\infty}e^{-(t)}(\cos tx)dt = \frac{1}{2\pi}\int_{0}^{\infty}e^{-t}(\cos tx)dt$$

$$= \frac{2}{2\pi}\int_{0}^{\infty}e^{-t}(\cos tx)dt = \frac{1}{\pi}\left[\frac{e^{-t}}{1+x^2}(-\cos xt + x\sin xt)\right]_0^{\infty}$$

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

2.4 Binomial Distribution

Definition

A r.v X which takes two values 0 and 1 with probabilities q and p respectively. i.e., P(X=1)p; P(X=0)=q is called a Bernoulli variate and its said have a Bernoulli distribution.

If the experiment is repeated n-times independently with two possible outcome,then they are called Bernoulli trials.

An experiment consisting of a repeated n number of Bernoulli trails is called Bernoulli experiment.

Binomial Experiment

A binomial distribution can be used under the following condition:

(i) Any trail with two possible outcomes that is any trail result in a success or failure.

(ii) The number of trials n is finite and independent,when n is number of trial.

(iii) a probability of success is the same in each trial. i.e., p is the constant.

Definition

A random variable X is said to have a binomial distribution, if its pmf is given by

$$P(X=x)=\begin{cases} nC_x P^x q^{n-x}, & x=0,1,2,\ldots n \\ 0 & ,\text{otherwise} \end{cases} \quad \text{where } q=1-p$$

It is denoted by B(n, p), where n and p are parameters

Applications of Binomial Distribution

1. The quality control measures and sampling process in industries to classify the items are defective or non-defective.
2. Medical applications as a success or failure of a surgery and cure or non cure of a patient.
3. Military application as a hit a target or miss a target

Derivation of mean and variance of B (n, p):

By the definition of mathematical expectation,

$$E(X)=\sum_{x=0}^{n} xP(x)=\sum_{x=0}^{n} x\,nC_x p^x q^{n-x}$$

$$= np\sum_{x=1}^{n} n-1C_x\, p^{x-1} q^{n-x}$$

$$= np\,(q+p)^{n-1} \quad \text{(by binomial expansion)}$$

$$= np(1) \quad (q+p=1)$$

$$\text{Mean} = E(x) = np \qquad (1)$$

$$Var(x) = E(x^2) - [E(x)]^2$$

$$E(x^2) = \sum_{x=0}^{n} x^2 P(x)$$

$$= \sum_{x=0}^{n} [x(x-1)+x]p(x)$$

$$= \sum_{x=0}^{n} x(x-1)p(x) + \sum_{x=0}^{n} xp(x)$$

$$= \sum_{x=0}^{n} x(x-1).nC_x p^x q^{n-x} + np \text{ (From (1))}$$

$$= \sum_{x=0}^{n} x(x-1).\frac{n(n-1)}{x(x-1)} n-2C_{x-2} p^2 . p^{x-2} q^{n-x} + np.$$

$$= n(n-1)p^2 \sum_{x=0}^{n} n-2C_{x-2} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p(q+p)^{n-2} + np$$

$$= n(n-1)p^2 + np$$

$$E(x^2) = np(np+q)$$

$$Var(x) = E(x^2) - [E(x)]^2$$

$$= np(np+q) - (np)^2$$

$$= n^2p^2 + npq - n^2p^2$$

$$Var(x) = npq$$

MGF and hence mean and variance

By the definition of MGF,

$$M_x(t) = E[e^{tx}]$$

$$= \sum_{x=0}^{n} e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx} nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} nC_x (pe^t)^x q^{n-x}$$

$$= nC_0 (pe^t)^0 q^n + nC_1 (pe^t)^1 q^{n-1} + \dots + nC_n (pe^t)^n q^{n-n}$$

$$= q^n + nC_1 (pe^t) q^{n-1} + \dots + (pe^t)^n$$

$$M_x(t) = (q + pe^t)^n$$

Differentiate with respect to t, we get

$$\frac{d}{dt}M_x(t) = n(q + pe^t)^{n-1}.pe^t$$

$$\text{Put } t = 0,\ \frac{d}{dt}M_x(t) = n(q + p)^{n-1}.pe^0$$

$$\text{Mean} = np = \mu_1'$$

$$\frac{d}{dt}M_x(t) = n(q + pe^t)^{n-1}.pe^t$$

$$= np(q + pe^t)^{n-1}e^t$$

$$\frac{d^2}{dt^2}M_x(t) = np\left\{(q + pe^t)^{n-1}e^t + e^t(n-1).(q + pe^t)^{n-2}.pe^t\right\}$$

$$\left.\frac{d^2}{dt^2}M_x(t)\right|_{t=0} = np\{1 + (n-1)p\}$$

$$np + n^2p^2 - np^2 = \mu_2'$$

$$\therefore \text{var}(x) = \mu_2' - (\mu_1')^2$$

$$= np + n^2p^2 - np^2 - (np)^2$$

$$\text{Var}(x) = npq$$

Definition of Moments

Moments about origin μ_r' is defined as the expectations of the powers of the r.v X. That is $\mu_r' = E(x^r)$. Similarly, the central momentsabout mean is defined as $\mu_r = E(x-\mu)^r$.

Recurrence relation for the central moments of a B(n, p)

By the definition of k^{th} order central moment μ_k is given by

$$\mu_k = E(x-\mu)^k = E(x-np)^k$$

$$= \sum_{x=0}^{n}(x-np)^k nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n}(x-np)^k nC_x p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} nC_x (x-np)^k p^x (1-p)^{n-x}$$

Differentiate with respect to p, we get

$$\frac{d}{dp}\mu_k = \sum_{x=0}^{n} nC_x \left\{(x-np)^k (p^x(n-x)(1-p)^{n-x-1}(-1) + (1-p)^{n-x}.(xp^{x-1}) + p^x(1-p)^{n-x}.k(x-np)^{k-1}(-n)\right\}$$

After simplification, we get,

$$\frac{d\mu_k}{dp} = -nk\mu_{k-1} + \frac{1}{pq}\mu_{k+1}$$

$$\mu_{k+1} = pq\left[\frac{d\mu_k}{dp} + nk\mu_{k-1}\right] \ldots\ldots (1)$$

Central moments of B(n, p)

Using the above recurrence relation we may compute the moments of higher order, provided the moments of lower order, that is $\mu_0 = 1 \, and \, \mu_1 = 0$.

$$\therefore \mu_{k+1} = pq\left[\frac{d\mu_k}{dp} + nk\mu_{k-1}\right]$$

Put k = 1,

$$\mu_2 = pq\left[\frac{d}{dp}\mu_1 + n\mu_0\right]$$

$$= pq[0 + n]$$

$$= npq, \text{ which is variance of X}$$

$$\therefore \mu_2 = npq$$

Put k = 2,

$$\mu_3 = pq\left[\frac{d}{dp}\mu_2 + 2n\mu_1\right]$$

$$= pq\left[\frac{d}{dp}(npq) + 0\right]$$

$$= npq(1-2p)$$

Put k = 3,

$$\mu_4 = pq\left[\frac{d}{dp}\mu_3 + 3n\mu_2\right]$$

$$= pq\left\{\frac{d}{dp}[npq(1-2p)] + 3n(npq)\right\}$$

$$= pq\left\{n\frac{d}{dp}p(1-p)(1-2p) + 3n^2pq\right\}$$

$$= npq\{1 + 3pq(n-2)\}$$

These are the first four binomial central moments.

The first four raw moments (or) moment about origin of B(n, P)

By the definition of moments about origin $\mu_r' = E(x^r)$

To find the first four raw moments:

Put r = 1

$\mu_1' = E(x^1)$

$$= \sum_{x=0}^{n} xp(x)$$

$$= \sum_{x=0}^{n} x\, nC_x p^x q^{n-x}$$

$$= np\sum_{x=0}^{n} n-1C_x p^{x-1} q^{n-x}$$

$$= np\,(q+p)^{n-1}$$

$\mu_1' = np$

$\mu_2' = E(x^2)$

$$= \sum_{x=0}^{n} x^2 p(x)$$

$$= \sum_{x=0}^{n} x(x-1)p(x) + \sum_{x=0}^{n} x\; p(x)$$

$$= n(n-1)p^2 \sum_{x=2}^{n} n-2C_{x-2} p^{x-2} q^{n-x} + np$$

$$= n\,(n-1)P^2\,(q+p)^{n-2} + np$$

$\mu_2' = np(np+q)$

$\mu_3' = E(x^3)$

$$= \sum_{x=0}^{n} x^3 p(x)$$

$$= \sum_{x=0}^{n} [x(x-1)(x-2) + 3x(x-1) + x] nC_x p^x q^{n-x}$$

$$= n(n-1)(n-2)p^3 \sum_{x=0}^{n} n-3C_{x-3} p^{x-3} q^{n-x} + 3n(n-1)p^2 \sum_{x=0}^{n} n-2C_{x-2} p^{x-2} q^{n-x} + np$$

$$\mu_3' = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

$$\mu_4' = E(x^4)$$

$$= \sum_{x=0}^{n} x^4 p(x)$$

$$= \sum_{x=0}^{n} [x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x] \; nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)(x-2)(x-3) nC_x p^x q^{n-x} + 6\sum_{x=0}^{n} x(x-1)(x-2) nC_x p^x q^{n-x}$$

$$+ 7\sum_{x=0}^{n} x(x-1) nC_x p^x q^{n-x} + \sum_{x=0}^{n} x \; nC_x p^x q^{n-x}$$

$$= n(n-1)(n-2)(n-3)p^4(p+q)^{n-4} + 6n(n-1)(n-2)p^3(p+q)^{n-3} + 7n(n-1)p^2(p+q)^{n-2} + np$$

$$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np.$$

Additive property of B(n, p) or Reproductive property

Statement

If $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$, then $X+Y \sim B(n_1+n_2, p)$ where X and Y are independent.

Proof

We know that, the MGF of $B(n, p) = (q+pe^t)^n$.

$\therefore$ The MGF of $X \sim B(n_1, p) = (q + pe^t)^{n_1}$.

Also the MGF of $Y \sim B(n_2, P) = (q + pe^t)^{n_2}$,

We know that, If X and Y are independent r.vs, then

$$M_{X+Y}(t) = M_x(t) \cdot M_x(t)$$

$$= (q + pe^t)^{n_1} \cdot (q + pe^t)^{n_2}$$

$$= (q + pe^t)^{n_1 + n_2}$$

$\therefore M_{X+Y}(t) = (q + pe^t)^{n_1 + n_2}$

Which is the MGF of $B(n_1+n_2, p)$

$\therefore (X+Y) \sim$ Binomial distribution

Note

If $X_1, X_2, \ldots, X_k$ are independent binomial variates with parameters $(n_1,p), (n_2,p), \ldots, (n_k,p)$ respectively,then $X_1+X_2+\ldots+X_k$ is also a binomial variate with parameter $(n_1+n_2+\ldots+n_k, p)$.

Mode of Binomial distribution

Definition

The value of x at which p(x) obtains maximum is called mode of the distribution.

Let X be a binomial r.v. Then $P(X=x)=p(x)=nC_x p^x q^{n-x}$; $x = 0, 1, 2, \ldots n$

Themode of the binomial distribution is defined by m_0 and it is given by

$$p(m_0-1) \leq p(m_0) \geq p(m_0+1)$$

Consider,

$$p(m_0-1) \leq p(m_0)$$

$$nC_{m_0-1}p^{m_0-1}q^{n-(m_0-1)} \le nC_{m_0}p^{m_0}q^{n-m_0}$$

$$\Rightarrow \frac{(n-m_0)!m_0!}{(n-m_0+1)!(m_0-1)!}\cdot\frac{q}{p} \le 1$$

$$\frac{m_0}{n-m_0+1} \le \frac{p}{q}$$

$m_0 \le p(n+1)$ .... (1)

Consider,

$P(m_0) \ge p\,(m_0+1)$

$$nC_{m_0}p^{m_0}q^{n-m_0} \ge nC_{m_0+1}p^{m_0+1}q^{n-(m_0+1)}$$

$$\Rightarrow \frac{(n-m_0-1)!(m_0+1)!}{(n-m_0)!(m_0)!} \ge \frac{p}{q}$$

$$\frac{m_0+1}{n-m_0} \ge \frac{p}{q}$$

$m_0 \ge np - q$ ....(2)

from (1) and (2)

$np - q \le m_0 \le p\,(n+1)$

For checking:

when $n = 10$, $p=1/2$, $q = ½$

$4.5 \le m_0 \le 5.5$.

Characteristic function and Cumulative function or cumulative generating function

The characteristic function is defined

$$\varphi_x(t) = E[e^{itx}]$$

Cumulative generating function is defined by

$$\kappa_x(t) = \log M_x(t)$$

Characteristic function of B(n,p)

By the definition of characteristic function,

$$\varphi_x(t) = E[e^{itx}]$$

$$= \sum_{x=0}^{n} e^{itx} p(x)$$

$$= \sum_{x=0}^{n} e^{itx} nC_x p^x q^{n-x}$$

$$\varphi_x(t) = (q + pe^{it})^n$$

2.5 Poisson distribution

- Simen Denis Poisson

Definition

A random variable X is said to follow the Poisson distribution if its probability mass function is given by,

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

Here the λ is the parameter and $\lambda > 0$

Poisson distribution as a limiting case of Binomial distribution:

Poisson distribution as a limiting case of Binomial distribution under the following condition:

i) The number of trial n is infinitely large. i.e., $n \to \infty$.

ii) The constant probability of success p in each trail is vary small. i.e., $p \to 0$

iii) $np = \lambda$ is finite, where λ is a positive real number.

Proof:

In the case of Binomial distribution,the probability of x success is given by,

$$p(X = x) = p(x) = nC_x p^x q^{n-x}$$

$$= \frac{n(n-1)(n-2)...[n-(x-1)]}{x!} p^x q^{n-x}$$

Put $np = \lambda$; $p = \lambda/n$

$$q = 1 - \frac{\lambda}{n}$$

$$\Rightarrow p(x) = \frac{n(n-1)(n-2)...[n-(x-1)]}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!}.\frac{n}{n}.\frac{n-1}{n}.\frac{n-2}{n}...\frac{n-(x-1)}{n}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{x!}\left[1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\left(1-\frac{3}{n}\right)...\left(1-\frac{x-1}{n}\right)\right]\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

Taking limit $n \to \infty$, we get

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

which is the pmf of Poisson distribution.

$\therefore$ Poisson distribution is the limiting case of binomial distribution.

Aliter

The MGF of B(n ,p) is

$$M_x(t)=\left(q+pe^t\right)^n$$

Put np = λ; p = λ/n

$$q=1-\frac{\lambda}{n}$$

$$\therefore M_x(t)=\left(1-\frac{\lambda}{n}+\frac{\lambda}{n}e^t\right)^n$$

$$=\left(1+\frac{\lambda(e^t-1)}{n}\right)^n$$

Taking limit $n\to\infty$ we get

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

which is the MGF of Poisson distribution.

$\therefore$ Poisson distribution is limiting case of Binomial distribution.

Mean and variance of Poisson distribution

$$\text{Mean, } E(x)=\sum_{x=0}^{\infty}x\,p(x)$$

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

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

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

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

$\therefore$ Mean $E(x) = \lambda$

$$\text{Variance }(x) = E(x^2) - [E(x)]^2$$

$$E(x^2) = \sum_{x=0}^{\infty} x^2\, p(x)$$

$$= \sum_{x=0}^{\infty} [x(x-1) + x] \frac{e^{-\lambda}\lambda^x}{x!}$$

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

$$E(x^2) = \lambda^2 + \lambda$$

$$\text{Var}(x) = E(x^2) - [E(x)]^2$$

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

$$\text{Var}(x) = \lambda$$

$\therefore$ Mean = Variance = λ.

MGF and hence mean and variance of Poisson distribution

By the definition of MGF,

$$M_x(t) = E[e^{tx}]$$

$$= \sum_{x=0}^{\infty} e^{tx} p(x)$$

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

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

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

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

To find mean and variance

By the property of MGF,

$$M_x'(t) = e^{\lambda(e^t-1)} \lambda(e^t)$$

$$M_x'(t)\big|_{t=0} = e^{\lambda(1-1)} \lambda(e^0) = \lambda$$

$$M_x'(t) = \lambda$$

$$\because M_x'(t) = \lambda e^t e^{\lambda(e^t-1)}$$

$$M_x''(t) = \lambda\left[e^t e^{\lambda(e^t-1)} \lambda e^t + e^{\lambda(e^t-1)} e^t\right]$$

$$M_x''(t)\big|_{t=0} = \lambda[\lambda+1] = \lambda^2 + \lambda = \mu_2'$$

$$\text{Var}(x) = \mu_2 = \mu_2' - \left(\mu_1'\right)^2$$

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

$$\text{Var}(x) = \lambda$$

$\therefore$ Mean = Variance = λ.

Recurrence formula for the central moments of the Poisson distribution:

For Poisson distribution with parameter λ; the recurrence formula is,

$$\mu_{r+1} = \lambda\left[\frac{d\mu_r}{d\lambda} + r.\mu_{r-1}\right]$$

Proof

By definition of r^{th} order central moment is given by

$$\mu_r = E(x-\mu)^r$$

$$= E(x-\lambda)^r \quad (\because E(x) = \lambda)$$

$$= \sum_{x=0}^{\infty}(x-\lambda)^r.p(x)$$

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

Differentiate with respect to λ, we get,

$$\frac{d}{d\lambda}\mu_r = \sum_{x=0}^{\infty}\frac{1}{x!}\left[(x-\lambda)^r.(e^{-\lambda}x\lambda^{x-1} + \lambda^x e^{-\lambda}(-1)) + (e^{-\lambda}\lambda^x).r(x-\lambda)^{r-1}(-1)\right]$$

$$\Rightarrow \lambda\frac{d\mu_r}{d\lambda} = \mu_{r+1} - \lambda r\ \mu_{r-1}$$

$$\Rightarrow \mu_{r+1} = \lambda \frac{d\mu_r}{d\lambda} + \lambda r\ \mu_{r-1}$$

$$\Rightarrow \mu_{r+1} = \lambda\left[\frac{d\mu_r}{d\lambda} + r\mu_{r-1}\right].$$

The central moments μ_1, μ_2, μ_3 and μ_4:

The recurrence formula for central moments of Poisson distribution is,

$$\mu_{r+1} = \lambda \frac{d\mu_r}{d\lambda} + \lambda r\ \mu_{r-1} \qquad \ldots\ldots\ldots\ldots\ldots(*)$$

Also, we know that, $\mu_0 = 1$

$$\mu_1 = 0.$$

In order to get μ_2, put r=1 in (*),

$$\therefore \mu_2 = \lambda \frac{d\mu_1}{d\lambda} + \lambda \mu_0$$

$$= \lambda x_0 + \lambda x_1$$

$$\mu_2 = \lambda.$$

In order to get μ_3.Put r = 2 in (*),

$$\therefore \mu_3 = \lambda \frac{d\mu_2}{d\lambda} + 2\lambda \mu_{2-1}$$

$$= \lambda.1 + 2\lambda(0)$$

$$\mu_3 = \lambda$$

In order to get μ_4.Put r = 3 in (*),

$$\therefore \mu_4 = \lambda \frac{d\mu_3}{d\lambda} + 3\lambda \mu_2$$

$$= \lambda.1 + 3\lambda.\lambda$$

$$\mu_4 = \lambda + 3\lambda^2$$

$\therefore \quad \mu_1 = 0,\ \mu_2 = \lambda,\ \mu_3 = \lambda,\ \mu_4 = \lambda + 3\lambda^2$ are the first four central moments.

The first four moments about origin:

By the definition of r^{th} order raw moments,

$$\mu_r' = E\left[x^r\right]$$

$$\therefore \mu_1' = E(x) = E(x)$$

$$= \sum_{x=0}^{\infty} x \cdot p(x)$$

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

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

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

$$\mu_1' = \lambda$$

Also, $\mu_2' = E(x^2)$

$$\mu_2' = \sum_{x=0}^{\infty} x^2\, p(x)$$

$$= \sum_{x=0}^{\infty}[x(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}$$

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

$$\mu_2' = \lambda^2 + \lambda$$

Also, $\mu_3' = E(x^3)$

$$\mu_3' = \sum_{x=0}^{\infty}x^3\,p(x)$$

$$= \sum_{x=0}^{\infty}x^3\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}[x(x-1)(x-2)+3x(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}x(x-1)(x-2)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty}3x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty}x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^3 e^{\lambda} + 3e^{-\lambda}\lambda^2 e^{\lambda} + \lambda$$

$$\mu_3' = \lambda^3 + 3\lambda^2 + \lambda$$

Also $\mu_4' = E(x^4)$

$$\mu_4' = \sum_{x=0}^{\infty}x^4\,p(x)$$

$$= \sum_{x=0}^{\infty}x^4\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} [x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x] \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1)(x-2)(x-3) \frac{e^{-\lambda}\lambda^4\lambda^{x-4}}{x(x-1)(x-2)(x-3)(x-4)!}$$

$$+ 6\sum_{x=0}^{\infty} x(x-1)(x-2) \frac{e^{-\lambda}\lambda^3\lambda^{x-3}}{x(x-1)(x-2)(x-3)!}$$

$$+ 7\sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$

$\mu_4{}' = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$.

Additive property:

The sum of independent Poisson variates is also a Poisson variate.

i.e., $X_1, X_2, \ldots, X_n$ are n independent Poisson variates with parameter $\lambda_1, \lambda_2, \ldots \lambda_n$. Then $X_1 + X_2 + \ldots + X_n$ is also a Poisson variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_n$.

Proof:

We know that the MGF of Poisson distribution is,

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

Also we know that,

$$M_{x_1+x_2+\ldots+x_n}(t) = M_{x_1}(t).M_{x2}(t)..M_{x_n}(t)$$

$$= e^{\lambda_1(e^t-1)} + e^{\lambda_2(e^t-1)} + \ldots\ldots + e^{\lambda_n(e^t-1)}$$

$\therefore M_{X_1+X_2+\ldots+X_n}(t) = e^{(\lambda_1+\lambda_2+\ldots+\lambda_n)(e^t-1)}$ which is the MGF of $X_1 + X_2 + \ldots + X_n$ with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_n$.

$\therefore X_1 + X_2 + \ldots + X_n$ is also Poisson variate.

Examples of a Poisson distribution (Real life Problems)

1. Number of printing mistakes at each page of a book.
2. The number of road accident reported in a city per day
3. The number of death in a district due to rare disease.
4. The number of defective articles in a pocket of 200
5. The number of cars passing through a time interval t.

Theorem 1

If X and Y are two independent Poisson variates with parameters λ_1, λ_2,then the conditional distribution of $(X|X+Y)$ is Binomial.

Proof

Given X and Y are independent Poisson variates with parameter λ_1 and λ_2 respectively.

$$\therefore P(X = m) = \frac{e^{-\lambda_1}\lambda_1^{m}}{m!};\ X = 0, 1, 2, \ldots, m, \ldots$$

$$\therefore P(Y = n) = \frac{e^{-\lambda_2}\lambda_2^{n}}{n!};\ Y = 0, 1, 2, \ldots, n, \ldots$$

$$\therefore P(X|X+Y) = P(X = m|X+Y = n)$$

$$= \frac{P(X = m, X+Y = n)}{P(X+Y = n)}$$

$$= \frac{P(X = m, Y = n-m)}{P(X+Y = n)}$$

$$= \frac{P(X = m)P(Y = n-m)}{P(X+Y = n)}$$

$\therefore$ X and Y are independent.

$$= \frac{\dfrac{e^{-\lambda_1}\lambda_1^{\,m}}{m!} \cdot \dfrac{e^{-\lambda_2}\lambda_2^{\,n-m}}{(n-m)!}}{\dfrac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}}$$

Multiply and divide by $\left(\dfrac{n!}{\lambda_1+\lambda_2}\right)^m$

$$= \frac{n!}{(n-m)!m!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^m\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-m}$$

$$= nC_m p^m q^{n-m} \text{ where } p = \frac{\lambda_1}{\lambda_1+\lambda_2} \text{ and } q = \frac{\lambda_2}{\lambda_1+\lambda_2}$$

Which is the pmf of binomial distribution.

$\therefore$ If X and Y are two independent Poisson variate, then the condition probability of $X|X+Y$ is Binomial.

Theorem 2

If X is a Poisson variate with parameter λ and conditional distribution of $y \mid x$ follows binomial with parameters n and p, then the distribution of Y follows the Poisson distribution with parameter λp.

Proof

Given X is a Poisson variate with parameter λ.

$$\therefore P(X=x) = p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,...\infty$$

For a Binomial distribution $P(X=x) = p(x) = nC_x p^x q^{n-x}; x = 0, 1, 2, ...n$

Then we prove that, Y~Poisson (λp)

$$\therefore P[Y = m|X = n] = \frac{P(Y = m, X = n)}{P(X = n)}$$

$$\Rightarrow P(X = n, Y = m) = P(Y = m|X = n).P(X = n)$$

$$= nC_m p^m q^{n-m}.\frac{e^{-\lambda}\lambda^n}{n!} \qquad (1)$$

$$\therefore P[Y = m] = \sum_{n=0}^{\infty} P(X = n, Y = m)$$

$$= \sum_{n=m}^{\infty} nC_m p^m q^{n-m}.\frac{e^{-\lambda}\lambda^n}{n!} \text{ (from (1))}$$

$$= \frac{e^{-\lambda}p^m\lambda^m}{m!}\sum_{n=m}^{\infty}\frac{(\lambda q)^{n-m}}{(n-m)!}$$

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

which is the pmf of Poisson distribution with parameter is λp.

∴ If X~Poisson (λ) and $Y|X \sim B(n, p)$, then Y ~ Poisson(λp).

Theorem 3

If X and Y are two independent Poisson variates then X-Y is not a Poisson variate.

Proof

Given,

$$M_x(t) = e^{\lambda_1(e^t - 1)}$$

$$M_y(t) = e^{\lambda_2(e^t - 1)}$$

$$M_{x-y}(t) = M_x(t).M_{(-y)}(t)$$

$$= M_x(t).M_y(-t)$$

$$= e^{\lambda_1(e^t - 1)}.e^{\lambda_2(e^{-t} - 1)} \text{ which is not in the form of } e^{\lambda(e^t - 1)}.$$

So difference X-Y is not a Poisson variate.

Example:

8 coins are tossed at a time, 256 times. Find the expected frequencies of success (getting a head) and tabulate the result obtained

Solution:

$$p = \frac{1}{2}; q = \frac{1}{2}; n = 8; N = 256$$

The probability of success r times in n trials is given by $^{n}C_{r}q^{n-r}p^{r}$.

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

$$= {}^{8}C_{r}\left(\frac{1}{2}\right)^{8-r}\left(\frac{1}{2}\right)^{r}$$

$$= {}^{8}C_{r}\left(\frac{1}{2}\right)^{8}$$

Frequencies of 0, 1, 2, 3,…, 8 successes are:

Success	N P(r)	Expected frequency
0	$256\left(\frac{1}{256}\times {}^{8}C_{0}\right)$	1
1	$256\left(\frac{1}{256}\times {}^{8}C_{1}\right)$	8
2	$256\left(\frac{1}{256}\times {}^{8}C_{2}\right)$	28
3	$256\left(\frac{1}{256}\times {}^{8}C_{3}\right)$	56
4	$256\left(\frac{1}{256}\times {}^{8}C_{4}\right)$	70
5	$256\left(\frac{1}{256}\times {}^{8}C_{5}\right)$	56
6	$256\left(\frac{1}{256}\times {}^{8}C_{6}\right)$	28
7	$256\left(\frac{1}{256}\times {}^{8}C_{7}\right)$	8
8	$256\left(\frac{1}{256}\times {}^{8}C_{8}\right)$	1

Fitting a Poisson Distribution

When we want to fit a Poisson Distribution to a given frequency distribution, first we have to find out the arithmetic mean of the given data i.e., $\overline{X} = m$ when m is known the other values can be found out easily.

$$N\,P(X = x) = N \times \frac{e^{-\lambda}\lambda^{x}}{x!}; \qquad x = 0,1,2,\ldots,\infty;$$

$$NP(X = 0) = Ne^{-\lambda}$$

$$NP(X = 1) = NP(X = 0) \times \frac{m}{1}$$

$$NP(X = 2) = NP(X = 1) \times \frac{m}{2}$$

$$NP(X = 3) = NP(X = 2) \times \frac{m}{3}$$

$$NP(X = 4) = NP(X = 3) \times \frac{m}{4} \text{ and so on.}$$

Example 1

100 Car Radios are inspected as they come of the production line and number of defects per set is recorded below:

No. of Defects	0	1	2	3	4
No. of sets	79	18	2	1	0

Fit a Poisson distribution to the above data and calculate the frequency of 0, 1, 2, 3 and 4 defects.

$\left(e^{-0.25} = 0.779\right)$

Solution

Fitting Poisson distribution

No. of Defectives (x)	No. of Sets (f)	(fx)
0	79	0
1	18	18
2	2	4
3	1	3
4	0	0
	N = 100	**$\sum$fx = 25**

$$\overline{X} = \frac{25}{100} = 0.25 = \lambda$$

$$e^{-.25} = 0.779$$

$$\mathrm{NP}(0) = \mathrm{N}e^{-m} = 100 \times 0.779 = 77.90$$

$$\mathrm{NP}(1) = \mathrm{NP}(0) \times \frac{m}{1} = 77.90 \times 0.25 = 19.48$$

$$\mathrm{NP}(2) = \mathrm{NP}(1) \times \frac{m}{2} = 19.48 \times \frac{0.25}{2} = 2.44$$

$$\mathrm{NP}(3) = \mathrm{NP}(2) \times \frac{m}{3} = 2.44 \times \frac{0.25}{3} = 0.20$$

$$\mathrm{NP}(4) = \mathrm{NP}(3) \times \frac{m}{4} = 0.20 \times \frac{0.25}{4} = 0.10$$

Example 2

Fit a Poisson distribution to the following data and calculate the theoretical frequencies:

x:	0	1	2	3	4
f:	123	59	14	3	1

Solution

x	0	1	2	3	4	
f	123	59	14	3	1	$\sum f=200$
fx	0	59	28	9	4	$\sum fx = 100$

$$\text{Mean} = \frac{100}{200} 0.25$$

$$NP_{(0)} = Ne^{-m}$$

$$= 200 \times e^{-0.5}$$

$$= 200 \times 6065 = 121.3$$

Conclusion of expected frequencies:

x	Frequency N P(X=x)	
0	NP(0) =121.3	121
1	$NP(0) \times \frac{m}{1} = 121.3 \times 5 = 60.65$	61
2	$NP(1) \times \frac{m}{2} = \frac{60.65 \times 5}{2} = 15.16$	15
3	$NP(2) \times \frac{m}{3} = \frac{15.16 \times 5}{3} = 2.53$	3
4	$NP(3) \times \frac{m}{4} = \frac{2.53 \times 5}{4} = 0.29$	0
	Total	200

3.3 Normal Distribution or Gaussian Distribution

A random variable X is said to follow a normal distribution if its pdf is given by,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \qquad -\infty < x < \infty$$

$$-\infty < \mu < \infty$$

$$\sigma > 0$$

Here, f(x) is a legitimate density function as the total area under the normal curve is unity.

To prove that total probability is one,

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

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2} \text{dx}$$

$$\text{put } t = \frac{x-\mu}{\sqrt{2}\sigma}$$

$$dt = \frac{1}{\sqrt{2}\sigma} dx$$

$$\Rightarrow dx = \sqrt{2}\sigma\, dt$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} \sqrt{2}\,\sigma\, dt$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-t^2}\, dt$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2}\, dt$$

$$= \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2}\, dt$$

$$\text{put } t^2 = y$$

$$\Rightarrow \quad t = \sqrt{y}$$

$$\therefore \int_{-\infty}^{\infty} f(x)dx = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y} \frac{1}{2\sqrt{y}}\, dy$$

$$= \frac{2}{\sqrt{\pi}} \frac{1}{2} \int_{-\infty}^{\infty} e^{-y} y^{-1/2}\, dy$$

$$= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} y^{\frac{1}{2}-1} e^{-y}\, dy$$

We know that $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, dx$

$$= \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right)$$

$$=\frac{1}{\sqrt{\pi}}\sqrt{\pi}$$

$$=1$$

$$\therefore \int_{-\infty}^{\infty} f(x)dx = 1$$

f(x) is a legitimate density function.

Mean and Variance of $N(\mu,\sigma^2)$

If $X \sim N(\mu,\sigma^2)$, then $E(X) = \mu$ and $V(X) = \sigma^2$.

Proof

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

$$= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Put

$$t = \frac{x-\mu}{\sqrt{2}\sigma} \Rightarrow x = \mu + \sqrt{2}\sigma t$$

$$dt = \frac{1}{\sqrt{2}\sigma} dx$$

$$dx = \sqrt{2}\sigma dt$$

$$= \int_{-\infty}^{\infty} \left(\mu + \sqrt{2}\sigma t\right) \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} \sqrt{2}\sigma dt$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(\mu + \sqrt{2}\sigma t\right) e^{-t^2} dt$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \mu e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \sqrt{2}\sigma t e^{-t^2} dt$$

$$=\frac{\mu}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2}\,dt+\frac{\sqrt{2}\sigma}{\sqrt{\pi}}\int_{-\infty}^{\infty} t\,e^{-t^2}\,dt$$

$$=\frac{\mu}{\sqrt{\pi}}\times\sqrt{\pi}+\frac{\sqrt{2}\sigma}{\sqrt{\pi}}\times(0)=\mu=\mu_1'$$

$\therefore$ **Mean = E(X) = μ**

To find variance,

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

$$=\int_{-\infty}^{\infty} x^2\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

Put

$$t=\frac{x-\mu}{\sqrt{2}\sigma}\Rightarrow x=\mu+\sqrt{2}\,\sigma t$$

$$dt=\frac{1}{\sqrt{2}\sigma}dx$$

$$\Rightarrow dx=\sqrt{2}\,\sigma dt$$

$$=\int_{-\infty}^{\infty}\left(\mu+\sqrt{2}\,\sigma t\right)^2\frac{1}{\sigma\sqrt{2\pi}}e^{-t^2}\sqrt{2}\,\sigma dt$$

$$=\int_{-\infty}^{\infty}\left(\mu^2+2\sigma^2t^2+2\mu\sqrt{2}\,\sigma t\right)\frac{1}{\sigma\sqrt{2\pi}}e^{-t^2}\sqrt{2}\,\sigma dt$$

$$=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\mu^2e^{-t^2}dt+\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}2\sigma^2t^2e^{-t^2}dt+\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}2\mu\sqrt{2}\,\sigma t\,e^{-t^2}dt$$

$$=\frac{\mu^2}{\sqrt{\pi}}\int_{-\infty}^{\infty}e^{-t^2}dt+\frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty}t^2e^{-t^2}dt+\frac{1}{\sqrt{\pi}}2\sqrt{2}\mu\sigma\int_{-\infty}^{\infty}t\,e^{-t^2}dt$$

$$=\frac{\mu^2}{\sqrt{\pi}}\times\sqrt{\pi}+\frac{2\sigma^2}{\sqrt{\pi}}\times2\int_{0}^{\infty}t^2e^{-t^2}dt+0$$

$$\text{Put } t^2 = y$$

$$2t\,dt = dy$$

$$\Rightarrow dt = \frac{dy}{2\sqrt{y}}$$

$$\therefore E\left(X^2\right) = \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \times 2\int_0^\infty e^{-y}\, y \frac{dy}{2\sqrt{y}}$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty e^{-y} \sqrt{y}\, dy$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty y^{1/2} e^{-y} dy$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \int_0^\infty y^{3/2-1} e^{-y} dy$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \Gamma\left(3/2\right)$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \times \frac{\sqrt{\pi}}{2}$$

$$\therefore \mu_2' = \mu^2 + \sigma^2.$$

$$\therefore V(X) = \mu_2' - \left(\mu_1'\right)^2$$

$$= \mu^2 + \sigma^2 - \mu^2$$

$$\therefore Var(X) = \sigma^2.$$

Standard Normal Variate or Standard Norman Distribution

If X follows normal distribution $N(\mu, \sigma^2)$,then $z = \frac{x-\mu}{\sigma}$ is a standard normal variate with mean zero and variance one and is denoted by N(0,1).

The pdf of standard normal variate is given by,

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}; \quad -\infty < x < \infty$$

MGF and Mean and Variance

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

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

$$=\int_{-\infty}^{\infty} e^{tx}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx$$

Put

$$z=\frac{x-\mu}{\sigma} \Rightarrow x=z\sigma+\mu$$

$$dz=\frac{1}{\sigma}dx$$

$$\Rightarrow dx=\sigma dz$$

$$M_X(t)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{t(\mu+z\sigma)}e^{-\frac{z^2}{2}}\sigma\,dz$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{t(\mu+z\sigma)}e^{-\frac{z^2}{2}}dz$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{t\mu}e^{t\sigma z-\frac{z^2}{2}}dz$$

$$=\frac{e^{t\mu}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(z^2-2t\sigma z\right)}dz$$

Add and subtract by σ^2t^2

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(z^2 - 2t\sigma z + \sigma^2 t^2 - \sigma^2 t^2\right)} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[(z-\sigma t)^2 - \sigma^2 t^2\right]} dz$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \frac{\sigma^2 t^2}{e^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-\sigma t)^2} dz$$

Put

$U = z - \sigma t$

$du = dz$

$$= \frac{e^{\mu t} e^{\sigma^2 \frac{t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du$$

$$= e^{\mu t} e^{\sigma^2 \frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du$$

($\because$ the total probability of Standard Normal is one)

$$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}}$$

To find Mean and Variance

$$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}}$$

$$= e^{t\left(\mu + \frac{\sigma^2 t}{2}\right)}$$

$$= 1 + t\left(\mu + \frac{\sigma^2 t}{2}\right) + \frac{t^2}{2!}\left(\mu + \frac{\sigma^2 t}{2}\right)^2 + \frac{t^3}{3!}\left(\mu + \frac{\sigma^2 t}{2}\right)^3 + \ldots$$

$$= 1 + \frac{t}{1!}\mu + \frac{t^2}{2!}\sigma^2 + \frac{t^2}{2!}\mu^2 + \ldots.$$

The coefficient of $\frac{t}{1!} = \mu = \mu_1'$

$\therefore$ Mean $= \mu$.

The coefficient of $\frac{t^2}{2!}$ is $\sigma^2 + \mu^2$.

$\therefore \mu_2' = \sigma^2 + \mu^2$

$\therefore Var(X) = \mu_2' - \left(\mu_1'\right)^2$

$= \sigma^2 + \mu^2 - \mu^2$

$\therefore Var(X) = \sigma^2$

The first four Moments about Origin

$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}}$

$= e^{t\left(\mu + \sigma^2 \frac{t}{2}\right)}$

$$= 1 + t\left(\mu + \frac{\sigma^2 t}{2}\right) + \frac{t^2}{2!}\left(\mu + \frac{\sigma^2 t}{2}\right)^2 + \frac{t^3}{3!}\left(\mu + \frac{\sigma^2 t}{2}\right)^3 + \frac{t^4}{4!}\left(\mu + \frac{\sigma^2 t}{2}\right)^4 + \ldots$$

$$= 1 + \frac{t}{1!}\mu + \frac{t^2}{2!}\sigma^2 + \frac{t^2}{2!}\left(\mu^2 + \mu\sigma^2 t + \sigma^4 \frac{t^2}{4}\right)$$

$$+ \frac{t^3}{3!}\left(\mu^3 + 3\mu^2\sigma^2 \frac{t}{2} + 3\mu \frac{\sigma^4 t^2}{4} + \frac{\sigma^6 t^3}{8}\right)$$

$$+ \frac{t^4}{4!}\left(\mu^4 + 4\mu^3 \frac{\sigma^2 t}{2} + 6\mu^2\left(\frac{\sigma^2 t}{2}\right)^2 + 4\mu\left(\frac{\sigma^2 t}{2}\right)^3 + \left(\frac{\sigma^2 t}{2}\right)^4\right) + \ldots$$

$\therefore \mu_1' = The\ Coefficient\ of\ \frac{t}{1!} = \mu$

$$\mu_2' = \text{The Coefficient of } \frac{t^2}{2!} = \sigma^2 + \mu^2$$

$$\mu_3' = \text{The Coefficient of } \frac{t^3}{3!} = 3\mu\sigma^2 + \mu^3$$

$$\mu_4' = \text{The Coefficient of } \frac{t^4}{4!} = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$$

The First Four Central Moments

We know that, $\mu_0 = 1, \mu_1 = 0$

By the definition of central moments,

$$\mu_r = E(X-\mu)^r$$

$$\therefore \mu_2 = E(X-\mu)^2$$

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

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Put

$$t = \frac{x-\mu}{\sqrt{2}\,\sigma}$$

$$\Rightarrow x-\mu = \sqrt{2}\,\sigma t \quad , x = \sqrt{2}\,\sigma t + \mu$$

$$\Rightarrow dt = \frac{1}{\sqrt{2}\,\sigma} dx$$

$$= \int_{-\infty}^{\infty} \left(\sqrt{2}\,\sigma t\right)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} \sqrt{2}\,\sigma\,dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} t^2\,dt$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}2\int_0^\infty t^2e^{-t^2}\,dt$$

Put $t^2=y \Rightarrow 2t\,dt=dy$

$$\Rightarrow dt=\frac{1}{2\sqrt{y}}dy$$

$$=\frac{4\sigma^2}{\sqrt{\pi}}\int_0^\infty ye^{-y}\frac{1}{2\sqrt{y}}dy$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty y^{1-\frac{1}{2}}e^{-y}\,dy$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty y^{\frac{1}{2}}e^{-y}\,dy$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty y^{3-\frac{1}{2}}e^{-y}\,dy$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right)$$

$$=\frac{2\sigma^2}{\sqrt{\pi}}\frac{\sqrt{\pi}}{2}$$

$$\mu_2=\sigma^2$$

(or)

$$\mu_2=\mu_2'-(\mu_1')^2$$

$$=\sigma^2+\mu^2-\mu^2$$

$$\mu_2=\sigma^2$$

$$\mu_3=\mu_3'-3\mu_2'\mu_1'+2(\mu_1')^3$$

$$=3\mu\sigma^2+\mu^3-3(\sigma^2+\mu^2)\mu+2\mu^3=0$$

$$\therefore \mu_3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$$

$$= 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4 - 4(3\mu\sigma^2 + \mu^3)\mu + 6(\sigma^2 + \mu^2)\mu^2 - 3\mu^4$$

$$= 3\sigma^4 + 12\mu^2\sigma^2 - 12\mu^2\sigma^2 + 7\mu^4 - 7\mu^4$$

$$\therefore \mu_4 = 3\sigma^4$$

The r^{th} Central Moments of Normal Distribution

If X is a normal variate then the all odd order central moments does not exists, but all even order central moments exists.

Proof

By the definition of r^{th} order central moment

$$\mu_r = E(X-\mu)^r$$

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

$$= \int_{-\infty}^{\infty} (x-\mu)^r \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Put } t = \frac{x-\mu}{\sqrt{2}\sigma}$$

$$\Rightarrow x-\mu = \sqrt{2}\sigma t \quad , x = \sqrt{2}\sigma t + \mu$$

$$\Rightarrow dt = \frac{dx}{\sqrt{2}\sigma}$$

$$\Rightarrow dx = dt\sqrt{2}\sigma$$

$$\mu_r = \int_{-\infty}^{\infty} (\sqrt{2}\sigma t)^r \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2} \sqrt{2}\sigma\, dt$$

$$\mu_r = \frac{(2)^{\frac{r}{2}} \sigma^r}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^r e^{-t^2}\, dt \qquad (1)$$

Case (i)

If r is an odd integer, r = 2n+1.

From the equation (1),

$$\mu_{2n+1} = \frac{2^{\frac{2n+1}{2}} \sigma^{2n+1}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^{2n+1} e^{-t^2}\, dt$$

$$\mu_{2n+1} = 0, \quad n = 0,1,2,\ldots \left(\because t^{2n+1} e^{-t^2} \text{ is an odd function}\right)$$

$$\mu_1 = \mu_3 = \mu_5 = \ldots = 0$$

Case (ii)

If r is an even integer, r = 2n.

$$\mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^{2n} e^{-t^2}\, dt$$

$$= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} 2 \int_0^{\infty} t^{2n} e^{-t^2}\, dt$$

$$\text{Put } y = t^2 \quad \Rightarrow t = \sqrt{y} = y^{\frac{1}{2}}$$

$$dy = 2t\, dt$$

$$\Rightarrow dt = \frac{1}{2\sqrt{y}} dy$$

$$\mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} 2 \int_0^{\infty} y^{\frac{2n}{2}} e^{-y} \frac{1}{2\sqrt{y}} dy$$

$$= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} y^{n-\frac{1}{2}} e^{-y}\, dy$$

$$= \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} y^{\left(n+\frac{1}{2}\right)-1} e^{-y} \, dy$$

$$\mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \Gamma\left(n + \frac{1}{2}\right) \qquad (2)$$

After simplification, we get,

$$\mu_{2n} = 1.3.5.7....(2n-1).\sigma^2 n \qquad (3)$$

when $n = 1, \mu_2 = 1.\sigma^{2(1)} = \sigma^2$

when $n = 2, \mu_4 = 3.\sigma^{2(2)} = 3\sigma^4$

and so on.

The Recurrence relations of Central Moments

We consider the equation (2),

$$\mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \Gamma\left(n + \frac{1}{2}\right)$$

Put n = n-1, 2n = 2(n-1) = 2n-2

Also,

$$\mu_{2n-2} = \frac{2^{n-1} \sigma^{2(n-1)}}{\sqrt{\pi}} \Gamma\left(n - 1 + \frac{1}{2}\right)$$

$$\mu_{2n-2} = \frac{2^{n-1} \sigma^{2n-2}}{\sqrt{\pi}} \Gamma\left(n - \frac{1}{2}\right) \qquad (4)$$

From the equations (2) and (4), we get,

$$\frac{\mu_{2n}}{\mu_{2n-2}} = \frac{\dfrac{2^n \sigma^{2n}}{\sqrt{\pi}} \Gamma\left(n + \frac{1}{2}\right)}{\dfrac{2^{n-1} \sigma^{2(n-1)}}{\sqrt{\pi}} \Gamma\left(n + \frac{1}{2}\right)}$$

$$= \frac{2\sigma^2\left(n-\frac{1}{2}\right)\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(n-\frac{1}{2}\right)}$$

$$\frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2\,\frac{2n-1}{2}$$

$$\frac{\mu_{2n}}{\mu_{2n-2}} = (2n-1)\sigma^2$$

$$\Rightarrow \mu_{2n} = (2n-1)\sigma^2\,\mu_{2n-2}$$

which is the recurrence relation of the even order central moment of normal distribution.

Additive Property (or) Reproductive Property:

If $X_1, X_2, \ldots X_n$ are n independent normal variates with mean $\mu_1, \mu_2, \ldots \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively,then $\sum_{i=1}^{n} a_i x_i$ is also a normal variate with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i \sigma_i^2$.

Proof

The mgf of normal distribution is,

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$$\Rightarrow M_{\sum_{i=1}^{n} a_i x_i}(t) = M_{a_1 x_1}(t).M_{a_2 x_2}(t)\ldots\ldots M_{a_n x_n}(t)$$

$$= e^{a_1 \mu_1 t + \frac{a_1^2 \sigma_1^2 t^2}{2}}.e^{a_2 \mu_2 t + \frac{a_2^2 \sigma_2^2 t^2}{2}}\ldots\ldots$$

$$M_{\sum_{i=1}^{n} a_i x_i}(t) = e^{\sum_{i=1}^{n} a_i \mu_i t + \sum_{i=1}^{n} \frac{a_i^2 \sigma_i^2 t^2}{2}}$$

Which is the mgf of normal distribution with mean $\sum a_i x_i$ and variance $\sum a_i \sigma_i^2$.