data given below.

Age (yrs) : Below	20	20–25	25–30	30–35	35–40	40–45	45–50	50 & above
No. of employee	13	29	46	60	112	94	45	21

Solution : This is an open-end series. So. the most suitable measure of Skewness is Bowley's Co-efficient of Skewness.

Age (yrs)	No. employee (f)	c.f.
Below 20	13	13
20 – 25	29	42
25 – 30	46	88
30 – 35	60	148
35 – 40	112	260
40 – 45	94	354
45 – 50	45	399
50 and above	21	420

420 = N

We know that, Bowley's coefficient of Skewness

$$SK_B = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$$

Here N = 420

$\therefore$ Q_1 = Value of $\left(\frac{N}{4}\right)^{th} = \frac{420}{4} = 105^{th}$ item

Median = Value of $\left(\frac{N}{2}\right)^{th} = \frac{420}{2} = 210^{th}$ item

Q_3 = Value of $\left(\frac{3N}{4}\right)^{th} = \frac{3 \times 420}{4} = 315^{th}$ item

$\therefore$ Q_1 class is 30 – 35

Median class is 35 – 40

Q_3 class is 40 – 45

$$\therefore Q_1 = l + \frac{N/4 - cf}{f} \times i$$

$$= 30 + \frac{105-88}{60} \times 5$$

= 30 + 1.42 = 31.42 yrs

Median $= l + \frac{N/2 - cf}{f} \times i$

$$= 35 + \frac{210-148}{112} \times 5$$

= 34 + 2.77 = 37.77 yrs

and $Q_3 = l + \frac{3N/4 - cf}{f} \times i$

$$= 40 + \frac{315-260}{94} \times 5$$

= 42 + 2.93 = 42.93 yrs

$$\therefore SK_B = \frac{42.93 + 31.42 - 2 \times 37.77}{42.93 - 31.42}$$

$$= \frac{-1.19}{11.51} = -0.103$$

4.5 MOMENTS

In statistics, 'Moments' are used to describe the various characteristics of a frequency distribution viz central tendency, dispersion, Skwness and Kurtosis.

4.5.1 Moments About Mean

Let a variable x takes the values $x_1, x_2, \ldots x_n$. Then $\bar{x} = \frac{\Sigma x}{n}$ is its arithmetic mean.

The r^{th} moment of x about the mean $\bar{x}$, is usually denoted by μ_r [where μ is the letter (mu) of the Greek alphabet] is defined as

$$\mu_r = \frac{1}{n}\Sigma(x - \bar{x})^r \qquad r = 1, 2, 3, 4 \qquad (1)$$

Similarly, for a frequency distribution of a variable x

x	x_1	x_2	x_n	
y	f_1	f_2	f_n	$\Sigma f = N$

Then, $\bar{x} = \frac{\Sigma fx}{N}$ is the arithmetic mean

Then r[th] moment about $\bar{x}$ is defined as

$\mu_r = \frac{1}{N}\Sigma f(x-\bar{x})^r$, r = 1, 2, 3, 4 (2)

Putting r = 1 in (1) and (2) we get

$\mu_1 = \frac{1}{n}\Sigma(x-\bar{x})^r = 0$ → For individual series

$\mu_1 = \frac{1}{N}\Sigma f(x-\bar{x})^r = 0$ → For frequency distribution

[Because, the algebric sum of deviations of a given set of observations from their AM is zero]

Thus, the first moment about mean is always zero.

Putting r = 2.

$\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 = \sigma x^2$ → For individual series.

$\mu_2 = \frac{1}{N}\Sigma f(x-\bar{x})^2 = \sigma x^2$ → For frequency distribution.

Thus, the second moment about mean is the variance of the distribution.

Also, $\left.\begin{aligned}\mu_3 &= \frac{1}{n}\Sigma(x-\bar{x})^3 \\ \mu_4 &= \frac{1}{n}\Sigma(x-\bar{x})^4\end{aligned}\right\}$ For individual series

and $\left.\begin{aligned}\mu_3 &= \frac{1}{N}\Sigma f(x-\bar{x})^3 \\ \mu_4 &= \frac{1}{N}\Sigma f(x-\bar{x})^4\end{aligned}\right\}$ For frequency distribution

4.5.2 Moments About Arbitrary Point A

When actual mean is in fractions, moments are first calculated about an assumed mean A (arbitrary point) and then are converted about the actual mean.

The r[th] moment of a variable x about an arbitrary point A is

given by

$$\mu'_r = \frac{\Sigma(x-A)^r}{n} \qquad \rightarrow \text{For individual series}$$

and $$\mu'_r = \frac{\Sigma f(x-A)^r}{N} \qquad \rightarrow \text{For frequency distribution.}$$

Now, $$\mu'_1 = \frac{1}{n}\Sigma(x-A)^2 = \frac{1}{n}\Sigma x - \frac{1}{n}\Sigma A$$

$$= \bar{x} - \frac{1}{n} \times nA = \bar{x} - A$$

Also, $$\mu'_1 = \frac{1}{N}\Sigma f(x-A)$$

$$= \frac{1}{N}\Sigma fz - \frac{1}{N}\Sigma fA$$

$$= \bar{x} - \frac{A}{N}\Sigma f = \bar{x} - \frac{A}{N} \times N = \bar{x} - A$$

$\therefore$ For both individual series and frequency distribution.

$$\boxed{\mu'_1 = \bar{x} - A}$$

Note : 1) μ_r the rth moment about the mean; r = 1, 2, 3, 4, ... are also called **Central** moments and μ'_r the rth moment about any arbitrary point A are called **raw moments** or **non-central moments.**

2) The moments (raw or central) of higher order other than fourth order are seldom used.

4.5.3 Relation Between Central and Raw Moments

Central Moment $= \mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$

Raw Moments about an arbitrary point A

$$\mu'_r = \frac{1}{n}\Sigma(x-A)^r$$

Now, $$\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$$

$$= \frac{1}{n}\Sigma\{x - A - (\bar{x} - A)\}^r$$

$$= \frac{1}{n}\Sigma\left(x - A - \mu'_1\right)^r$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\ \mu'_1 \Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\ \Sigma(x-A)^{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\ \mu'_1 \frac{1}{n}\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\ \Sigma(x-A)^{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \mu'_r - {}^rC_1\ \mu'_1 - \mu'_{r-1} + {}^rC_2(\mu'_1)^2\mu'_{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r$$

From the above relation, for various values of r we have

For r = 1, $\mu_1 = \mu'_1$

For r = 2 $\mu_2 = \mu_2' - {}^2C_1\mu'^2_1 + {}^2C_2\mu_1'^2\mu_0'$

$$\Rightarrow \boxed{\mu_2 = \mu_2' - (\mu_1')^2}$$

$$\mu_0 = \frac{1}{n}\Sigma(x-x)^0$$

$$= \frac{1}{n}\Sigma 1$$

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

$$\mu_0' = \frac{1}{n}\Sigma(x-A)^0$$

$$= \frac{1}{n}\Sigma 1$$

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

For r = 3, $\mu_3 = \mu_3' - {}^3C_1\mu_1'\ \mu_2' + {}^3C_2(\mu_1')^2\mu_1' - {}^3C_3(\mu_1')^3\mu_0'$

$$\Rightarrow \boxed{\mu_3 = \mu_3' - 3\mu_1'\ \mu_2' + 2(\mu_1')^3}$$

Similarly,

For r = 4, $\boxed{\mu_4 = \mu_4' - 4\mu_1'\ \mu_3' + 6\mu_2'(\mu_1')^2 - (\mu_1')^4}$

Example 9 : The first two moments of a distribution about the value 5 of the variable are 2 and 20 respectively. Find the mean and variance of the distribution.

Solution : Given A = 5, $\mu_1' = 2$, $\mu_2' = 20$

We know that mean $(\bar{x}) = A + \mu_1' = 5 + 2 - 7$

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

$= 20 - (2)^2 = 16$

Example 10 : Find the first four moment about the mean for the set of numbers 2, 4, 6 and 8.

Solution : $\mu_r = r^{th}$ Central moment

$$= \frac{\Sigma(x-\bar{x})^r}{n}, \text{ r = 1, 2, 3, 4}$$

Table for Calculation of First Four Moments

x	$x-\bar{x}=x-5$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
2	–3	9	–27	81
4	–1	1	–1	1
6	1	1	1	1
8	3	9	27	81
$20=\Sigma x$	$0=\Sigma(x-\bar{x})$	$20=\Sigma(x-\bar{x})^2$	$0=\Sigma(x-\bar{x})^3$	$164=\Sigma(x-\bar{x})^4$

$$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{20}{4} = 5$$

$$\therefore \mu_1 = \frac{\Sigma(x-\bar{x})}{4} = 0, \quad \mu_2 = \frac{\Sigma(x-\bar{x})^2}{4} = \frac{20}{4} = 5$$

$$\mu_3 = \frac{\Sigma(x-\bar{x})^3}{4} = 0, \quad \mu_4 = \frac{\Sigma(x-\bar{x})^4}{4} = \frac{164}{4} = 41$$

Example 11 : The first four moments of a distribution about the value 4 are – 1.5, 17, – 30 and 108. Find the moments about mean.

Solution : Given A = 4

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

$\therefore$ First moment about mean = μ_1

$$= A + \mu_1'$$

$$= 4 - 1.5 = 2.5$$

Second moment about mean = μ_2

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

$$= 17 - (-1.5)^2 = 14.75$$

Third moment about mean = μ_3

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

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

$$= -30 + 76.5 - 6.75 = 39.75$$

Fourth central moment about mean

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

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

$$= 108 - 180 + 229.5 - 15.1875$$

$$= 142.3125$$

4.6 KARL PEARSON'S BETA (β) AND GAMMA (γ) COEFFICIENT

Prof Karl Pearson defined the following four co-efficients based on the 1st four central moment.

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

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

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \quad \left[\because \mu_2 = \sigma^2\right]$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

It may be stated here that these co-efficients are pure numbers independent of units of measurement and as such can be conveniently used for comparative studies. In practice, they are used as measure of Skewness and Kurtosis as discussed below.

4.6.1 Co-efficient of Skewness Based on Moments

The co-efficient of Skewness, based on the moments is given by–

$$SK = \frac{\beta_2 + 3\sqrt{\beta_1}}{2(5\beta^2 - 6\beta_1 - 9)}$$

$\therefore$ Sk = 0 if $\beta_1 = 0$ or $\beta_2 + 3 = 0 \Rightarrow \beta_2 = -3$

But $\mu_4 = \frac{1}{n}\Sigma(x - \bar{x})^4 > 0$

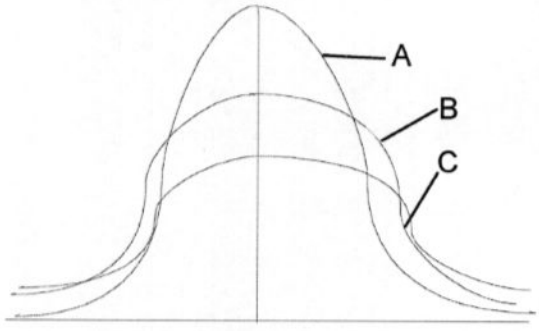
and $\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 > 0$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} > 0$$

Since β_2 can not be negative, SK = 0 if $\beta_1 = 0$.

4.7 KURTOSIS

So far we have studied three measure viz central tendency, dispersion and skewness to describe the characteristics of a frequency distribution. However, even if we know all these three measures, we all not in a position to charcterise a distribution completely. The diagram given below will clarify the situation.



A – Lapto Kurtic
B – Meso Kurtic
C – Platy Kurtic

All these three curves are symmetrical about the mean and have same variation (range). But the shape and nature of the middle part of the curve are different i.e the peakedness of the three curves are not same. So. to identify a distribution completaly, we need one more measure which is called **'Kurtosis'**.

The word **Kurtosis** comes from a Greek word meaning 'humped'. In statistics, **Kurtosis refers to the degree of flatness or peakedness of the frequency curve.**

Curve of type B which is neither flat peaked is known as normal curve and shape of its hump is accepted as a standard one. This curve is termed as **mesokurtic.** The curves of type A, which is more peaked than the normal curve are known as **leptokurtic** and the curves of type C, which are flatter than the normal curve are called **platykurtic**

4.7.1 Measures of Kurtosis

As a measure of Kurtosis, Karl Pearson gave the coefficient Beta two (β_2) or its derivative Gramma two (δ_2) defined as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{6^4} \qquad (1)$$

$$\alpha_2 = \beta_2 - 3 = \frac{\mu_4}{6^4} - 3 \qquad (2)$$

For a normal curve (mesokurtic curve), $\beta_2 = 3$ or $r_2 = 0$. For a lepto kurtic curve, $\beta_2 > 3$ or $r_2 > 0$ and for a platy kurtic curve, $\beta_2 < 3$ or $r_2 < 0$.

Note : A British statistician W. S. Gosset explained the use of the terms **platykurtic** and **leptokurtic** as : **'platykurtic curves like platyplus, are squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping.**

Example 12 : The first four moments of a distribution about the value 5 and 2, 20, 40 and 50. Do you think that the distribution is platykurtic ?

Solution : Here A = 5, $\mu_1' = 2, \mu_2' = 20, \mu_3' = 40, \mu_4' = 50$

$\therefore$ Mean $(\bar{x}) = \mu_1' + A = 2 + 5 = 7$

Now, $\mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$

$\therefore$ Variance $\mu_2 = 16$

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

$= 40 - 3 \times 2 \times 20 + 2\,(2)^3 = -64..7$

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

$= 50 - 4 \times 2 \times 40 + 6 \times 20 \times (2)^2 - 3(2)^4 = 162$

$$= \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = 0.63$$

Since $\beta_2 < 3$, the distribution is platykurtic.

Example 13 : The first four moments of a distribution about the origin are 1, 4, 10 and 46 respectively. Calculate the central moments and indicates the nature of the distribution.

Solution : Given A = 0, $\mu_1' = 1, \mu_2' = 4, \mu_3' = 10, \mu_4' = 46$

$\therefore$ Mean $(\bar{x}) = A + \mu_1' = 0 + 1 = 1$

Variance $(\sigma)^2 = \mu_2 = \mu_2' - (\mu_1')^2 = 4 - (1)^2 = 3$

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

= 10 – 3 x 4 x 1 + 2(1)^3

= 0

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

= 46 – 4 x 10 x 1 + 6 x 4 x 1 – 3 = 27

$\therefore$ Karl Pearson's moment co-efficient of Skewness is given by

$$\delta_1 = \frac{\mu_3}{\mu_2^{3/2}} = 0, \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

Since, $\beta_1 = 0$ the given distribution is symmetrical (normal).

Again, Karl Pearson's measure of Kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{27}{(3)^2} = 3$$

and $\delta_2 = \beta_2 - 3 = 0$

Since, $\beta_2 = 3$, it is mesokurtic (normal) distribution with

Mean $(\bar{x}) = 1$

and $S.D = \sqrt{3} = 1.732$

Example 14 : If $\beta_1 = +1, \ \beta_2 = 4$ and variance = 9, find the values of μ_3 and μ_4 and comment upon the nature of Kurtosis.

Solution: Given $\beta_1 = +1, \ \beta_2 = 4$

Variance $= \sigma^2 = \mu_2 = 9$

Now, $\beta_1 = +1 \Rightarrow \frac{\mu_3^2}{\mu_2^3} = 1 \Rightarrow \mu_3^2 = \mu_2^3 = 9 \times 9 \times 9 = (3 \times 3 \times 3)^2$

$\Rightarrow \mu_3^2 = (27)^2$

$\Rightarrow \mu_3 = \pm 27^2$

Also $\beta_2 = 4$

$\Rightarrow \frac{\mu_4}{\mu_2^2} = 4$

$\Rightarrow \mu_4 = 4 \times 9 \times 9 = 324$

$\therefore \mu_3 = \pm 27$ and $\mu_4 = 324$

$\mu_2 = 4 > 3$

Hence the given distribution is leptokurtic.



CHECK YOUR PROGRESS

Q 1: Explain meaning of skewness using sketches of frequency curves. How does skewness different from dispersion.

Q 2: Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why?

Q 3: How is Kurtosis measured ?

Q 4: Fill in the blanks.

(i) For a symmetrical distribution mean, median and mode are .....

(ii) If $\mu_2 > 3$, the curve is called .....

(iii) If $\mu_1 = 0$, the distribution is .....

(iv) If Mode > Mean, the distribution is ..... skewed.

(v) Kurtosis measures ..... of the frequency curve.

(vi) Bowley's co-efficient of skewness lies between .....

(vii) Relative measure of Kurtosis is .....

Q 5: Say whether the following statements are true or false.

(i) For a symmetrical distribution $\beta_1 = 0$.

(ii) Bowely's co-efficient of skewness lie between ± 3.

(iii) For a negatively skewed distribution mean > mode.

(iv) The first moment about the origin L mean.

(v) Variance = μ_2

(vi) If $\beta_2 > 3$, the curve is platykurtic.

Q 6: Consider the following distribution–

	Distribution A	**Distribution B**
Mean	100	90
Median	90	80
Variance	100	100

(i) Both the distribution have same degree of skewness.

(ii) Distribution A has the same degree variation as B.

– Say whether the following statements are true or false. Justify your answer.

Q 7: For a distribution $\bar{x} = 30$, $\sigma = 8$ and SK_p = 0.4. Find Median and Mode.

Q 8: For a distribution, Mean = 50, CU = 40%, SK_p = –0.4. Find SD, Median and Mode.

Q 9: For a group of 10 items, $\Sigma x = 452$ and Mode = 43.7. Find the pearsonion co-efficient of Skewness.

Q 10: The measure of skewness for a distribution is – 0.8. If the upper and lower quartiles are respectively 56.6 and 44.1, find Median.

Q 11: Calculate Karl Pearson's co-efficient of skewness from the data given below.

Marks	**No. of students**
21 – 25	5
26 – 30	15
31 – 35	28
36 – 40	42
41 – 45	15
46 – 50	12
51 – 55	3

Q 12: Calculate Bowley's co-efficient of skewness

Wt (kg):	0–15	15–30	30–45	45–60	60–75	75–90	90 and above
f :	20	30	30	35	45	15	5

Q 13: (a) Find the first and second moments about the mean for the numbers 3, 4, 5 and 8.

(b) Find the first, second, third moments of the numbers 1, 3, 6, 7, 8 about the value 4.

Q 14: (a) The first three moments of a distribution about the value of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment measure of skewness.

(b) The first four moments of a distribution about x = 2 are 1, 2.5, 5.5 and 16. Calculate the four moments about ($\overline{X}$) and about zero.

Q 15: (a) The central moments of a distribution are given by (), calculate the moment measure of skewness and kurtosis and comment on the shape of the distribution.

(b) From the data given below do you think that the distribution is platykurtic ?

N = 100, $\Sigma fd = 50$, $\Sigma fd^2 = 1970$, $\Sigma fd^3 = 2948$ and $\Sigma fd^4 = 86{,}752$ where d = x – 48.



4.8 LET US SUM UP

In this unit we have learnt about the following–

- Meaning of symmetrical and asymmetrical distribution.
- Meaning of Skewness and types of Skewness.
- Different measures of skewness.
- Meaning of moments, central moments, Raw moments.
- Use of moments for describing various characteristics of frequency distribution.
- Meaning of kurtosts and types of Skewness.
- To measure Skewness and Kurtosis using moments.



4.9 FURTHER READINGS

1) Gupta S. C. *Fundamental of Statistics.* Himalaya Publishing House.

2) Gupta S. C. and Gupta Indira. *Business Statistics.* Himalaya Publishing House.

3) P. N. Arora, Sumeet Arora, S. Arora, *Comprehensive Statistical Methods.* S. Chand and Company.

4) R. K. Ghosh and S. Saha. *Business Mathematics and Statistics.* New Central Book Agency (P) Ltd.

5) Agarwal D. R. *Business Statistics.* Delhi, Vrinda Publication.

6) Sharma J. K. *Business Statistics.* Pearson Education Ltd.



4.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 4: (i) Equal (ii) Leptokurtic (iii) symmetrical. (iv) negatively (v) peakedness/flatness. (vi) ± 1 (vii) β_2 or r_2.

Ans to Q No 5: (i) True (ii) False (iii) False (iv) True (v) True (vi) False.

Ans to Q No 6: (i) True (ii) False

Ans to Q No 7: Median = 28.93, Mode = 26.8

Ans to Q No 8: SD = 20, Mode = 58, Median = 52.67

Ans to Q No 9: $SK_p = 0.077$

Ans to Q No 10: Median = 55.35

Ans to Q No 11: $SK_p = -0.109$

Ans to Q No 12: $SK_B = -0.1126$

Ans to Q No 13: (a) $\mu_1 = 0,\ \mu_2 = 3.5$

(b) $\mu_1' = 1,\ \mu_2' = 7.8,\ \mu_3' = 14.2$

Ans to Q No 14: (a) mean = 3, SD = 4.58

$\delta_1 = \sqrt{\beta_1}$ = – 0.561

(b) Moments about mean.

$\mu_2 = 1.5,\ \mu_3 = 0,\ \mu_4 = 6$

Moments about zero

$\mu_1' = 3,\ \mu_2' = 10.5,\ \mu_3' = 40.5,\ \mu_4' = 168$

Ans to Q No 15: (a) $\delta_1 = 0.0893,\ \beta_2 = 0.3076$

The distribution is approximately symmetrical and platykurtic

(b) $\beta_2 = 2.2.14$, $\beta_2 < 3$, distribution is platykurtic.



4.11 MODEL QUESTIONS

Q 1: Averages dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distribution. Discuss the Statement.

Q 2: Define moments establish the relationship between the moments about mean and moments about any orbitrary point.

Q 3: The first two moments of a distribution about the value 5 are 2 and 20. Find the mean and variance.

Q 4: Calculate Karl Peason's Co-efficients of Skewness for the given distribution.

Marks less than:	20	40	60	80	100
No. of students:	18	40	70	90	100

Q 5: Calculate Bowley's Co-efficient of Skewness.

Marks :	0–10	10–20	20–30	30–40	40–50
No. of students:	5	9	12	8	6

Q 6: The arithmetic mean of a certain distribution is 5. The second and the third moments about the mean are 20 and 140 respectively. Find the third moment of the distribution about zero.

*** ***** ***

UNIT 5: CORRELATION

UNIT STRUCTURE

5.1 Learning Objectives

5.2 Introduction

5.3 Correlation: Meaning and Concept

5.4 Types of correlation

5.5 Methods of Measuring correlation

5.6 Scatter Diagram and Correlation

5.7 Types of Correlation

5.8 Correlation Co-efficient

5.9 Underlying Assumptions

5.10 Interpretation of Correlation Co-efficient

5.11 Rank Correlation Co-efficient

5.12 Let Us Sum Up

5.13 Further Readings

5.14 Answers To Check Your Progress

5.15 Model Questions

5.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- explain the concept of correlation coefficients of correlation and their properties
- discuss different types of correlation
- explain the essential features of correlation analysis
- explain and compare the various aspects of correlation

5.2 INTRODUCTION

Depending on the number of variables that we study, we may have univariate, bivariate or multivariate series or distribution. Univariate series or distribution is one when we deal with only one variable. For example, if we consider, say, only age of different persons, then we will have a univariate series, i.e. values of only age. Bivariate distribution, on the other hand,

involves two variables (note that 'bi' means two). If, for instance, we take both income and expenditure of persons then distribution thus obtained will constitute a bivariate distribution. Multivariate distribution will, similarly consist of more than two variables. The concept - correlation that we will study in this unit are associated with bivariate and/or multivariate series or distributions.



LET US KNOW

A **variable** is an entity that can take any value. Age, income, height, weight, expenditure, marks etc. are examples of variables. Note that for different persons these will take different values. Contrary to it, **constant** is an entity that can assume only fixed value.



ACTIVITY 1

A. Try to list some variables and constants that you find in common day to day life.

.....

.....

.....

5.3 CORRELATION: MEANING AND CONCEPT

In the simplest term, correlation means relation between two variables (note that 'co' simply means two things together). In case of bivariate distributions, many a times, we may be interested in looking at how the two variables are related among themselves. Obviously, when two variables are related, change in any one variable is responded by the other variable showing some changes in it. Correlation analysis tries to measure the 'extent' of such responses among the variables. Thus, correlation is an analysis of co-variation between two (or even more) variables. Two variables are said to be correlated if the change in any one of them results in a corresponding change in the other.

5.4 TYPES OF CORRELATION

Correlation may be classified into the following three types:

(a) Positive correlation, (b) Negative correlation and (c) Zero Correlation.

(a) Positive Correlation: When increase (or decrease) in one variable leads to a corresponding increase (or decrease) in the other variable then it is called positive correlation. The basic idea of positive correlation is that both the variables move in the same direction. For example, there exists positive correlation between the following pairs of variables:

(i) Income and expenditures of a group of families

(ii) Amount of rainfall and yield of crops

(iii) Price and supply of a company

(iv) Temperature and sale of ice-creams on different days of a month in summer.

(b) Negative Correlation: Negative correlation is, on the other hand, is a situation when increase (or decrease) in one variable is accompanied by a decrease (or increase) in the other variable. Unlike the positive correlation, the fundamental element in negative correlation is that the variables move in opposite directions. For example, following are the situations of negative correlations:

(i) Price of a commodity and demand for it.

(ii) Sale of woolen garments and the day temperature.

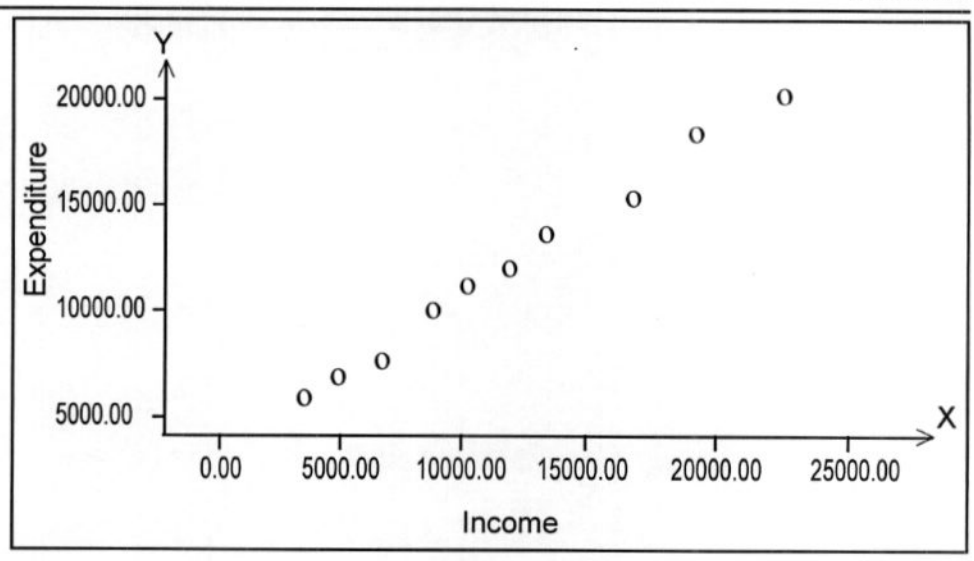
(iii) Number of workers and the time required to complete work.

(c) Zero correlation: In case, variables are found as uncorrelated i.e., change in one of the variables fail to produce any corresponding change in the other, and then it is called zero correlation. For example, one should expect zero correlation between the heights of workers and the income earned by them, or between price of rice and demand for sugar.

5.5 METHODS OF MEASURING CORRELATION

The following are the methods used to measure the correlation between two variables:

(i) Scatter diagram Method, (ii) Karl Pearson's Correlation Method and (iii) Spearman's Rank Correlation Method.



5.6 SCATTER DIAGRAM AND CORRELATION

How can one find whether any two variables are correlated or not? One of the easiest ways for identifying correlation between two variables is scatter diagram. Scatter diagram is a very simple graphical technique of identification of correlation. When we consider respective pairs of values as points and diagram them against X-Y axes then we get the scatter plot of the values. To understand this, let us consider the following bivariate series of income and expenditure of say, ten persons.

Person (No.)	Income (Rs.)	Expenditure (Rs.)
1	9200	6700
2	6860	4256
3	12825	9875
4	17980	12900
5	22125	19700
6	7892	4780
7	4645	3200
8	10925	8645
9	13820	10900
10	24780	21200

Taking income in X-axis and expenditure in Y-axis, the values of the above table can now be plotted in a graph to see whether these values are

correlated. The scatter thus obtained will be somewhat like this

It may be seen from the above scatter that points show an upward-rising trend. If the scatter does indicate some sort of trend or pattern then it implies existence of correlation among the variables. If, for example, pattern is such that there is a visible upward trend from left to right, then variables are said to have positive correlation. In case, there is no such identifiable pattern then the variables are said to be uncorrelated.



ACTIVITY 2

Note down height and weight of 10 of friends of yours. Then try to draw a scatter plot of the values. What can you say about their possible correlation?

5.7 TYPES OF CORRELATION

Depending upon the pattern that the scatter exhibits, we may have two distinct types of correlation - positive and negative.

When increase (or decrease) in one variable leads to a corresponding increase (or decrease) in the other variable then it is called positive correlation. The basic idea of positive correlation is that both the variables move in the same direction. Such a movement will result in scatter-plot to show an upward pattern of points or dots from left to right.

Negative correlation is, on the hand, is a situation when increase (or decrease) in one variable is accompanied by a decrease (or increase) in the other variable. Unlike the positive correlation, the fundamental element in negative correlation is that the variables move in opposite directions. Obviously, scatter-diagram for negatively correlated variables will be a downward slopping pattern from left to right.

In case, variables are found as uncorrelated i.e. change in one of the variables fail to produce any corresponding change in the other, then it is called zero correlation. The scatter exhibits no identifiable pattern in case the variables are uncorrelated.



CHECK YOUR PROGRESS

Q 1: Choose the correct answer from the following:

i) Correlation and regression are associated with

a) Univariate series b) Bivariate series

c) Both bivariate and multivariate series d) Multivariate series

.....

ii) Scatter plot technique is used to detect

a) Correlation b) Regression

.....

iii) When variables move in opposite directions then it is

a) Positive Correlation b) Negative Correlation

.....

5.8 CORRELATION COEFFICIENT

Scatter plots are particularly useful to identify whether the variables are correlated and in case they are then what is the type of such relation. It, however, can't quantify or quantitatively measure the 'amount', 'extent', 'magnitude' or 'degree' of relation that exists between the variables. British statistician Karl Pearson (1867 - 1936) introduced a simple technique for measuring the magnitude of such correlation. His measure is called Pearson's Product-Moment Correlation Coefficient or simply Pearson' Correlation Coefficient and usually denoted by 'r' (standing for 'relation') which is given as

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$$

Here, cov (x, y) means "covariance of x and y" and σ_x and σ_y are standard deviations of x and y respectively. Formulas for obtaining these are as below,

$$Cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum(y-\bar{y})^2}$$

Putting these in the original formula and simplifying, we will have,

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

This can be further simplified and can be written as

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

This formula is relatively easier to calculate and most commonly used to obtain the Pearson's coefficient of correlation; Here 'n' means total number of observations.

Here is procedure of finding the 'r' for a series involving two variables x and y:

- Obtain total of x i.e. $\sum x$
- Obtain total of y i.e. $\sum y$
- Get product of x and y for all pairs and then get the total of xy i.e. $\sum xy$
- Get x^2 and y^2 for all values of x and y and then get their totals i.e. $\sum x^2, \sum y^2$
- Put these values to the formula and get the r.



LET US KNOW

Carl Pearson, later known as Karl Pearson (1857-1936) was born to William Pearson and Fanny Smith, was also an accomplished historian and Germanist. Karl Pearson was educated privately at University College School,



after which he went to King's College, Cambridge in 1876 to study mathematics. He then spent part of 1879 and 1880 studying medieval and 16th century German literature at the universities of Berlin and Heidelberg. He graduated from Cambridge University in 1879 as Third Wrangler in the Mathematical Tripos. In 1911 he founded the world's first university statistics department at University College London. Besides, correlation coefficient, chi-square is another significant tool contributed by Pearson.



EXERCISE 1

A. Given the following pairs of values:

Capital employed (lakhs of Rs)	12	13	15	16	18	29	32
Profits (lakhs of Rs)	6	5	7	8	12	11	17

a) Draw a scatter diagram.

b) Do you see any correlation between capital employed and volume of profits? What can be concluded about their correlation if it exists?

B. Calculate Pearson's coefficient of correlation for following data

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

5.9 UNDERLYING ASSUMPTIONS

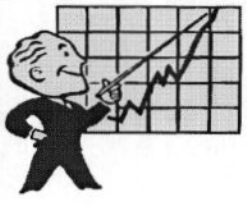
The most fundamental assumption of Pearson's correlation coefficient is that it supposes that the variables are quantitative. Further, it assumes a linear relation between the two variables. Therefore, the value of 'r' indicates extent of linearity that exists between the two variables. In case, the value of 'r' is found to be equal to 'zero', it may be best understood as 'absence of linear relations' among the variables. It may, however, be possible that the variables have other non-linear types of relations.

5.10 INTERPRETATION OF CORRELATION COEFFICIENT

Once the value of 'r' is obtained, it is very important to correctly interpret the result. Interpretation of the value of 'r' depends on two things - sign and magnitude. When 'sign' is positive correlation is said to be positive, and when 'sign' happens to be negative correlation is also said to be negative.

On the other hand, the magnitude of the value of correlation coefficient can not exceed the value of 1. The value of 'r', therefore, will always lie in between -1 and +1. More closer is the value to 1, greater is the extent of linear correlation. Conversely, further away is the value from 1 greater is amount of deviation from (linear) correlation. Many scholars provide ranges of values of 'r' for meaningful interpretation. Most commonly it is held that –

When the value of 'r' is 1, it is called perfect correlation. For value ranging from 0.8 to 1, variables are said to be "highly correlated". When the value lies in between 0.6 to 0.8, variables are called "fairly correlated". For value of 'r' in between 0.3 to 0.6, it is called moderate correlation. For values of 'r' less than 0.3 it is called low correlation. When the value of 'r' is equal to 0, the variables are said to be linearly uncorrelated or linearly independent.



CHECK YOUR PROGRESS

Q 2: Choose the correct answer from the following:

i) Correlation analysis is based on the assumption that

a) Variables are quantitative b) Variables are qualitative

.....

ii) Interpretation of r depends on

a) Only sign b) Only magnitude c) Both sign and magnitude

.....

iii) When value of r falls within the range of 0.6 to 0.8 it is called

a) Fair Correlation b) Moderate Correlation

c) Low Correlation c) Perfect Correlation

.....

iv) In case of perfectly positive correlation, the value of r is

a) -1 b) +1 c) 0.8 d) 0

.....

5.11 RANK CORRELATION COEFFICIENT

Many a times, we come across variables, which can not be quantitatively measured. However, these can be 'ranked' and accordingly 'ordered'. For example, beauty, intelligence, honesty etc. In fact, most of the qualitative attributes are of this type. It is difficult to measure exact amount of honesty that a person has, or exact amount of intelligence that he possesses. It is, nevertheless, very much possible to find out which person is more honest, or intelligent or beautiful so on and so forth. When we try to do such comparisons, we in fact, attach some ranks to individual and then try to arrange in some order depending on the ranks. Such variables which can't be directly measured but can be ordered are called ordinal variables. When we have a bivariate distribution involving two such ordinal variables then how do we find correlation? In this case Pearson's correlation coefficient can not be applied as it is based on the fundamental assumption of quantitative variable.

Charles Edward Spearman, a British psychologist, developed a formula in 1904 for obtaining a coefficient of correlation of ranks of two attributes. To obtain the correlation of ranks the two attributes are to be ordered and then their differences of ranks are to be calculated. If say 'd' is the difference of ranks between the two attributes then the rank correlation coefficient is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$, n is the number of observation

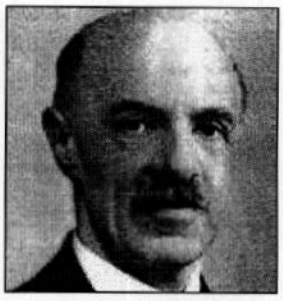
Procedure to obtain the rank correlation coefficient is fairly simple and easy:

- Arrange the two attributed in same order by their ranks
- Subtract each pair of ranks to get their differences, i.e. d and then get d^2
- Get the total of d^2 i.e. $\sum d^2$
- Put the values thus obtained in the formula and get the coefficient.

Just like the value of r, the value of also lies in between -1 and +1. Interpretation is also same as r except the foundational assumption of linearity.



LET US KNOW



Charles Edward Spearman, (September 10, 1863 - September 17, 1945) was an English psychologist known for work in statistics. He was a pioneer of factor analysis. He also did seminal work on models for human intelligence. Although Spearman achieved most recognition in his day for his statistical work, he regarded this work as subordinate to his quest for the fundamental laws of psychology, and he is now similarly renowned for both.



CHECK YOUR PROGRESS

Q 3: Choose the correct answer from the following:

I) Rank Correlation analysis is used when

a) Variables are quantitative b) Variables are qualitative

.....

ii) Spearman's rank correlation does not consider linearity

a) True b) False

.....

iii) Just like the value of r, Rank correlation also falls within +1 and -1

a) True b) False

.....



ACTIVITY 3

Ranking of 10 students by two teachers A and B are shown below. Obtain the rank correlation coefficient for the data and figure out the problem in decision making

A:	3	5	8	4	7	10	2	1	6	9
B:	6	4	9	8	1	2	3	10	5	7



5.12 LET US SUM UP

- Depending on the number of variables, we have univariate, bivariate or multivariate distributions.
- Correlation is with bivariate and/or multivariate distributions
- Correlation is the relation between two variables.
- Scatter diagram is a simple graphical method of identifying correlation. For correlated variables scatter plot exhibits some observable pattern or trend.
- Correlation can be of two types - positive when variables move in the same direction and negative when the variables move in the opposite directions.
- Correlation coefficient is based on assumptions of that variables are quantitative and linearly related.
- Interpretation of the correlation coefficient depends on both sign and magnitude of the value of r.
- Rank correlation is applied for qualitative variables which are not measurable but ordinal.
- Like Pearson's correlation coefficient, Spearman's rank correlation coefficient also lies in between +1 and -1.



5.13 FURTHER READINGS

1) Agarwal, D. R. (2006). Business Statistics. Delhi, Vrinda Publications.

2) Gupta S.C. (1994). Fundamentals of statistics. New Delhi, Himalayan Publishing House.

3) Rajagopalan, S. P. & Sattanthan R. (2009).Business Statistics and Operations Research. New Delhi, Tata McGraw-Hill

4) Sharma, J. K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

5) Verma, A. P. (2007). Business Statistics. Guwahat, Asian Books Private Limited.



5.14 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: i) c) Both bivariate and multivariate series

ii) a) Correlation

iii) b) Negative Correlation

Ans to Q No 2:

i) a) Variables are quantitative

ii) c) both sign and magnitude

iii) a) Fair Correlation

iv) b) +1

Ans to Q No 3:

(i) a) Variables are quantitative

(ii) a) True

(iii) a) True

Ans to Q No 4:

A. a) Average

B. c) Best

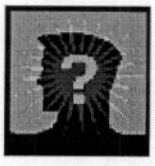
C. a) Minimum

D. a) 90^0

E. c) Mean

F. b) False

G. b) False



5.15 MODEL QUESTIONS

Q 1: What is meant by correlation ?What are the properties o the coefficient of correlation ?

Q 2: What are the methods of calculating coefficient of correlation ?

Q 3: Calculate Karl pearson's coefficient correlation from the data given below

X:	2	4	6	8	10
Y:	20	18	16	14	12

Q 4: Find Karl Pearson's coefficient of correlation between x and y from following data giving test scores of 10 candidates in mathematics and statistics and interpret

Scores in Mathematics :	98	70	40	20	85	75	95
	80	10	5				
Scores in Statistics :	85	65	32	30	80	60	61
	55	54	65				

Q 5: Find the coefficient of correlation from the following data give below

X:	12	20	15	18	33	24	30	12	15	22
Y:	30	35	28	36	29	39	30	25	30	38

*** ***** ***

UNIT 6: REGRESSION

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Meaning and Concept of Regression

6.4 Linear Regression

6.5 Line of Regression and Regression Equation

6.6 Coefficient of Regression

6.7 Relation between Correlation and Regression Coefficients

6.8 Correlation and Regression Analysis: A Comparison

6.9 Let Us Sum Up

6.10 Further Readings

6.11 Answers To Check Your Progress

6.12 Model Questions

6.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- explain the concept of regression and regression lines and equation of regression, and their properties
- discuss different types of regression
- explain the essential features of correlation and regression analysis
- explain and compare the various aspects of regression analysis

6.2 INTRODUCTION

In the previous unit, we introduced the concept correlation between two variables where it was discussed whether the variables move in the same direction or in the opposite direction and the extent of association between the two variables under study. On the other hand regression analysis is a statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable. In business, several

times it becomes necessary to have some forecast so that it can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets.

6.3 MEANING AND CONCEPT OF REGRESSION

Regression is one of the most important and powerful statistical techniques widely used in both science and social science studies. The term regression was introduced by Francis Galton. In a famous scientific paper "Family Likeness in Stature" published in Proceedings of Royal Society, London in 1886, he first talked about "regression to mediocrity".

Today, broadly speaking, regression analysis is referred to the study of the dependence of one variable on one or more other variables, with a view to estimating and/or predicting the average value of the former in terms of fixed or known values of the latter.

To understand consider this - suppose x and y are two variables such that values of y depends on values of x. Then regression analysis will try to estimate or predict the average value of y that it will take for any given value of x.

6.4 LINEAR REGRESSION

We have understood that regression is the technique of estimating average value of the dependent variable corresponding to any given value of the independent variable when the two variables are dependent upon each other i.e. the two variables are 'related' to each other. Now we know that relation may be either linear or non-linear. Linear regression is the technique of regression when variables are linearly related i.e. for each successive change in the independent variable there is a 'constant' change in the dependent variable.



LET US KNOW

A dependent variable is a variable that depends on value of some other variable. Independent variable is a variable that does not depend on values of the other variable. Linearity on the other hand means that when independent variable changes the dependent variable changes by constant rate. Such changes produce a straight line when plotted.



ACTIVITY 4

A. Try to list five dependent variables and five independent variables. Also find out which depends on what.

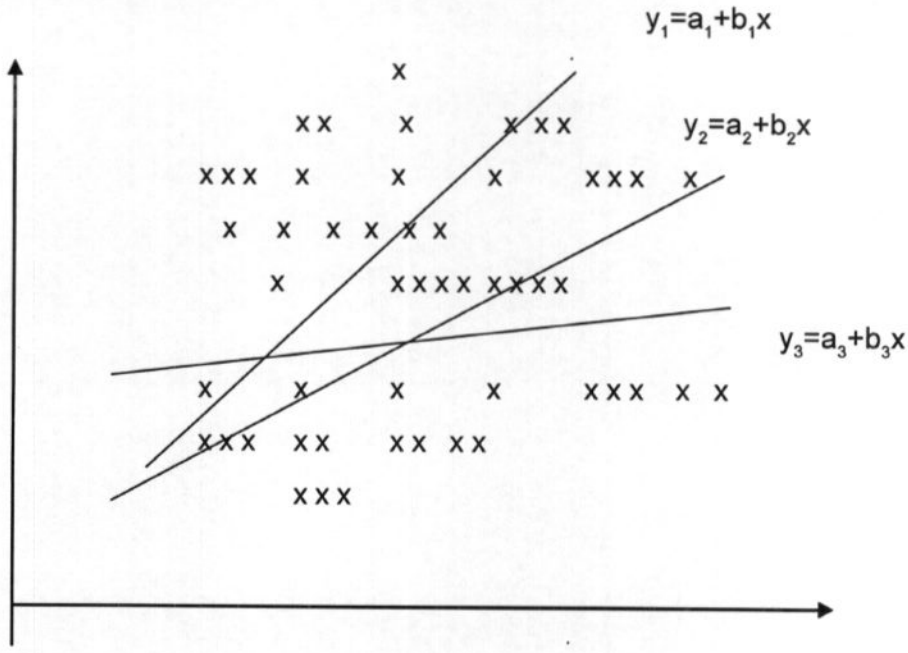
.....

.....

6.5 LINE OF REGRESSION AND REGRESSION EQUATION

Let us consider two variables x and y, which are linearly related. Relation is such that value of y depends on value of x. The relation may be expressed as $y = a + bx$, where a and b are two constants usually called parameters.

Note that value of y now depends on values of a and b given the value of x. For different values of the parameters a and b, we will have different lines that will pass through the points showing pairs x and y. Look at the following figure



In the figure you will notice that three lines passing through the points produces three different values of y depending on three different values a and b. Now since all three values of y are 'estimated' value therefore there will be an 'error' of estimation. Error is simply the difference between the 'actual' and 'estimated' values of y. That is

$e = y - \hat{y}$; where $\hat{y}$ is the estimated value of *y*.

Ideally it would be better if 'error' *r* be zero or minimum. Now since $\hat{y}$ depends on values of *a* and *b* we need to estimate *a* and *b* in such a way that *e* becomes minimum. Suppose, we estimate values of *a* and *b* as $\hat{a}$ and $\hat{b}$ so that we have $\hat{y}$ as

$\hat{y} = \hat{a} + \hat{b}x$; and resultant *e* is minimum.

Then this equation is called the regression equation and the line that this equation produces is called the line of regression or simply regression line. In other words regression equation is the equation that estimates dependent variable in such a way that the difference between the estimated and actual value of y or error in short, is minimum.

The line of regression often interpreted as line of "best fit". Among the all possible lines (look at the figure above showing three such lines) the line of regression fits i.e. touches most of the points (and that is why error is minimum).

When dependent and independent variables are clearly identified then we have single line of regression. When it is not clear will have two lines of regression one showing y as dependent and other showing x as dependent.

It is useful to remember that in case the variables are perfectly related i.e. if correlation coefficient is either -1 or +1 both the lines will coincide. When the variables are linearly independent regression lines will be perpendicular to each other. Also, both the lines of regression pass through the mean values of the variables.

6.6 COEFFICIENT OF REGRESSION

Consider the regression equation

$$\hat{y} = \hat{a} + \hat{b}x$$

In this equation $\hat{a}$ is called 'constant' term of the regression equation and $\hat{b}$ is called the 'coefficient' of regression. How can one obtain the $\hat{b}$? In fact, $\hat{b}$ is simply given by

$$\hat{b} = \frac{Cov(x, y)}{\sigma_x}$$

Formal derivation of $\hat{b}$ involves partial differentiation. However, the following formula is used to estimate the regression coefficient

$$\hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Once the $\hat{b}$ is obtained then the constant can be obtained by solving the following equation

$$\bar{y} = \hat{a} + \hat{b}\bar{x}$$

Where $\bar{y}$ and $\bar{x}$ are means of *y* and *x* respectively. The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.

6.7 RELATION BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

It may be seen that both coefficient of regression and coefficient of correlation are related to each other. We know that

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y} \text{ ; and}$$

$$\hat{b} = \frac{Cov(x, y)}{\sigma_x}$$

It is thus clear that

$$r = \frac{\hat{b}}{\sigma_y}$$

This relation is practically very useful for analysis. Remember that, in case we don't know clearly which one is dependent and which one is independent variable then we will have two lines of regression and hence two coefficients of regression. Let us assume that the two possible coefficients of regression are $\hat{b}_1$ and $\hat{b}_2$ respectively. Then they satisfy the following property

$$r^2 = \hat{b}_1 \times \tilde{b}_2$$

This follows from the above property that both regression coefficients have same sign and if value of any one is greater than unity value of the other must be less than unity.

6.8 CORRELATION AND REGRESSION ANALYSIS: A COMPARISON

Correlation analysis simply detects whether two variables are having any relations or not. Such detection simply looks for possible movements in variables and then some identifiable patterns in the movements. On the other hand, regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable. It also tries to explain the extent of variation caused by a variable. In this sense regression analysis has greater explanatory abilities than correlation analysis.



CHECK YOUR PROGRESS

Q 4: Choose the correct answer from the following:

A. Regression analysis tries to estimate ..... values.

a) Average b) Median c) Mode d) Range

.....

B. The line of regression is called line of ..... fit.

a) Worst b) Highest c) Best d) Lowest

.....

C. Regression equation estimates the values such that error is

a) Minimum b) Maximum c) Normal d) Abnormal

.....

D. When correlation is 0, then the angle between the lines of regression s

a) 90^0 b) 30^0 c) 45^0 d) 0^0

.....

E. Lines of regression pass through ..... values of the variables.

a) Maximum b) Minimum c) Mean d) Median

.....

F. Regression coefficients are of opposite sign.

a) True b) False

.....

G. Both the regression coefficients can be greater than unity.

a) True b) False

.....



6.9 LET US SUM UP

- Regression is a method of estimating average value of dependent variable given the value of independent variable.
- Line of regression gives the line of best fit and regression equation defines the line of regression. The line of regression passes through the means of the variables.
- The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.
- Correlation analysis simply detects whether two variables are having any relations or not. On the hand regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable.



6.10 FURTHER READINGS

1) Agarwal, D. R. (2006). Business Statistics. Delhi, Vrinda Publications.
2) Gupta S.C. (1994). Fundamentals of statistics. New Delhi, Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan R. (2009).Business Statistics and Operations Research. New Delhi, Tata McGraw-Hill
4) Sharma, J. K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.
5) Verma, A. P. (2007). Business Statistics. Guwahat, Asian Books Private Limited.



6.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: i) c) Both bivariate and multivariate series

ii) a) Correlation

iii) b) Negative Correlation

Ans to Q No 2:

i) a) Variables are quantitative

ii) c) both sign and magnitude

iii) a) Fair Correlation

iv) b) +1

Ans to Q No 3:

(i) a) Variables are quantitative

(ii) a) True

(iii) a) True

Ans to Q No 4:

A. a) Average

B. c) Best

C. a) Minimum

D. a) 90^0

E. c) Mean

F. b) False

G. b) False



6.12 MODEL QUESTIONS

Q 1: Explain the concept of regression and bring out its utility in business and commerce.

Q 2: What is a regression line? Why can we have two lines of regression? Explain the relationship between lines of regression and correlation coefficient.

Q 3: Why the regression line is called line of best fit? What do you understand by line of best fit?

Q 4: What is regression coefficient? What does it indicate?

Q 5: Fit a line of regression to following data, taking Y as dependent variable and estimate Y for X is equal to 20.

Y:	1	2	4	5	7	8	9
X:	1	3	4	8	9	11	14

Q 6: If followings are the two lines of regression find out the correlation coefficient and means of x and y

8x-10y+66=0 and 40x-18y=214

Q 7: Distinguish between correlation and regression analyses.

*** ***** ***