Measures of Dispersion

Measures of Dispersion / Spread / Variation

Measures of dispersion quantify how much individual value differ from the mean, focusing on amount not direction. Example: deviation of 6 above or below mean is same variation.

Significance of Measuring Variation

1) Determine Reliability of An Average

- Small variation → average is reliable
- Large variation → average is less reliable unless data is large

2) Control of The Variability

Helps identify causes of variation in quality control and production.

3) Compare Variability of Two or More series

- Less variation → more uniformity
- More variation → less consistency

4) Use Advance Statistical Tools

Needed in correlation, hypothesis testing, ~~fluctions~~ fluctuation analysis, cost control etc.

Properties of A Good Measure of Variation

i. Simple to understand

ii. Easy to compute

iii. Based on each and every observation of the distribution

iv. Amenable to further algebric treatment

v. Sampling stability

vi. Not be unduly affected by extreme observations.

Methods of Measuring Variation

1) **Absolute Measures:** In original units expressed in the same unit as data.

i) **The range**

$$R = L - S$$

S → smallest

L → Largest

·Merits

- Simplest to understand and easiest to compute.
- Takes minimum time to calculate the value of range.

Limitations

- Range is not based on each and every observation of the distribution.
- Range can't be computed in case of open-end distributions.

ii) The quantile deviation / interquartile range

- Interquartile range : $Q_3 - Q_1$

- $QD = \frac{Q_3 - Q_1}{2}$ → Quartile

Merits:

- Has a special utility in measuring variation in case of open-end distribution.
- Not affected by the presence of extreme values

Limitations:

- Ignores 50% items (Ex: the first 25% and the last 25%)
- Not capable of mathematical manipulation.

iii) The average (mean) deviation

$$AD_{(Med)} = \frac{\sum|x - Median|}{N}$$

$$AD_{(\bar{x})} = \frac{\sum|x - \bar{x}|}{N}$$

Merits:

- It is simple to understand and easy to compute
- Based on each and every observation of the data

Limitations:

- Algebric signs are ignored while taking the deviation
- Not capable of further algebric treatment.

iv) The standard deviation

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}}$$

v) Variance

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{N}$$

Mathematical properties of standard deviation

i) Combined standard deviation : Used to find SD of two groups together

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$d_1 = |\bar{x}_1 - \bar{x}_{12}| \qquad d_2 = |\bar{x}_2 - \bar{x}_{12}|$$

ii) SD of first n natural numbers

$$\sigma = \sqrt{\frac{N^2-1}{12}}$$

iii) SD is independent of change of origin but not scale

CamScanner

iv) • Mean (μ) $\pm 1\sigma$ $\longrightarrow$ 68.27% observations
• $\mu \pm 2\sigma$ $\longrightarrow$ 95.45% observations
• $\mu \pm 3\sigma$ $\longrightarrow$ 99.73% observations

Three sigma limit property

Merits of SD:

• Best among all measures of variation because of its mathematical characteristics

• Only SD can calculate the combined SD of two or more groups.

Limitations of SD

• Difficult to compute than other

• Gives more weights to extreme values and less to those is near the mean.

2) Relative Measures: Unit free, used when data series have different units.

i) Coefficient of Range

$$\text{Coefficient of Range} = \frac{L-S}{L+S} \times 100$$

L → largest

S → smallest

ii) Coefficient of Quartile Deviation

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

iii) Coefficient of Average Deviation

$$\text{Coefficient of Average Deviation} : AD_{(A)} = \frac{AD}{A} \times 100$$

A → Mean/Median/Mode/ Any arbitrary value

iv) Coefficient of Variation (CV)

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Theorem

For two unequal observation : $MD = SD = \frac{R}{2}$

Proof:

Let x_1, x_2 be two unequal observation $(x_1 > x_2)$

$$\bar{x} = \frac{x_1 + x_2}{2} \; ; \; R = x_1 - x_2 \; ; \; SD = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{2}}$$

$$MD = \frac{\sum_i |x_i - \bar{x}|}{2} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}|}{2}$$

$$= \frac{\left| x_1 - \frac{x_1 + x_2}{2} \right| + \left| x_2 - \frac{x_1 + x_2}{2} \right|}{2}$$

$$= \frac{\left| \frac{x_1 - x_2}{2} \right| + \left| \frac{x_2 - x_1}{2} \right|}{2}$$

$$= \frac{\frac{x_1 - x_2}{2} + \frac{x_1 - x_2}{2}}{2}$$

$$= \frac{x_1 - x_2}{2} = \frac{R}{2} \quad \text{——— (1)}$$

$$SD^2 = \frac{\sum(x-\bar{x})^2}{2} = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2}{2}$$

$$= \frac{\left(x_1 - \frac{x_1+x_2}{2}\right)^2 + \left(x_2 - \frac{x_1+x_2}{2}\right)^2}{2}$$

$$= \frac{\left(\frac{x_1-x_2}{2}\right)^2 + \left(\frac{x_2-x_1}{2}\right)^2}{2}$$

$$= \frac{\frac{(x_1-x_2)^2}{4} + \frac{(x_2-x_1)^2}{4}}{2}$$

$$= \left(\frac{x_1-x_2}{2}\right)^2 = \left(\frac{R}{2}\right)^2$$

$\therefore SD = \frac{R}{2}$ ——— (ii)

from (i) & (ii)

$\therefore MD = SD = \frac{R}{2}$ [proved]

Q-1: The following are the prices of shares of a company from Monday to saturday. Calculate range and coefficient of range.

Day	Price (Rs)	Day	Price (Rs)
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	saturday	250

Sol:

$L = 250$

$S = 160$

$$\text{Range} = L - S$$

$$= 250 - 160 = \text{Rs}\cdot 90$$

$$\text{Coefficient of Range} = \frac{L-S}{L+S} \times 100$$

$$= \frac{250-160}{250+160} \times 100$$

$$= 0\cdot219 \times 100$$

$$= 21\cdot9$$

Q-2: Calculate the coefficient of range from the following data:

Profits (Rs Lakhs)	No. of Cos	Profits (Rs Lakhs)	No. of Cos
10-20	8	40-50	8
20-30	10	50-60	4
30-40	12		

Sol:

$L = 60$

$S = 10$

$$\text{Coefficient of Range} = \frac{L-S}{L+S} \times 100$$

$$= \frac{60-10}{60+10} \times 100$$

$$= 0.714 \times 100$$

$$= 71.4$$

Ans

Q-4: Calculate the average deviation and coefficient of average deviation of the two income groups of five and seven workers working in two different branches of a firm.

Sol:

Branch-1		Branch - 2	
Income (Rs)	$\lvert x - Med \rvert$ (Med = 4400)	Income (Rs)	$\lvert x - Med \rvert$ (Med = 4400)
4000	400	3000	1400
4200	200	4000	400
4400	0	4200	200
4600	200	4400	0
4800	400	4600	200
		4800	400
		5800	1400
N = 5	$\sum \lvert x - Med \rvert = 1200$	N = 7	$\sum \lvert x - Med \rvert = 4000$

Branch-1:

$$AD = \frac{\sum \lvert x - Med \rvert}{N} = \frac{1200}{5} = 240$$

$$\text{Coefficient of AD} = \frac{240}{4400} \times 100 = 0.054 \times 100$$
$$= 5.4$$

Branch-2:

$$AD = \frac{\sum |x - Med|}{N} = \frac{4000}{7} = 571.43$$

$$\text{Coefficient of } AD = \frac{\sum |x - Med|}{Med} \times 100 = \frac{571.43}{4400} \times 100$$

$$= 0.13 \times 100$$

$$= 13$$

Q-5: Calculate average deviation from mean from the following data:

Sales (thous.)	mp (x)	f	$\frac{(x-35)}{10}$ (d)	fd	$|x-\bar{x}|$	$f|x-\bar{x}|$
10-20	15	3	-2	-6	18	54
20-30	25	6	-1	-6	8	48
30-40	35	11	0	0	2	22
40-50	45	3	+1	+3	12	36
50-60	55	2	+2	+4	22	44
		N=25		$\sum fd = -5$		$\sum f|x-\bar{x}| = 204$

$$\bar{x} = A + \frac{\sum fd}{N} \times i$$

$$= 35 - \frac{5}{25} \times 10$$

$$= 35 - 2 = 33$$

$$AD = \frac{\sum f|x-\bar{x}|}{N}$$

$$= \frac{204}{25} = 8.16$$

Thus the average sales are Rs 33 thousand per day and the average deviation of sales is Rs 8.16 thousand.

Q-6: Find the standard deviation from the weekly wages of ten workers working in a factory;

Sol:

Workers	Weekly Wages (Rs) (x)	$(x-\bar{x})$	$(x-\bar{x})^2$
A	1320	-3	9
B	1310	-13	169
C	1315	-8	64
D	1322	-1	1
E	1326	+3	9
F	1340	+17	289
G	1325	+2	4
H	1321	-2	4
I	1320	-3	9
J	1331	+8	64
$N=10$	$\sum x = 13230$	$\sum(x-\bar{x}) = 0$	$\sum(x-\bar{x})^2 = 622$

$$\bar{x} = \frac{\sum x}{N} = \frac{13230}{10} = 1323$$

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{N}} = \sqrt{\frac{622}{10}} = 7.89$$

Assumed Mean Method

Workers	Weekly Wages (x)	$(x-A)$ $A=1320$ (d)	d^2
A	1320	0	0
B	1310	-10	100
C	1315	-5	25
D	1322	+2	4
E	1326	+6	36
F	1340	+20	400
G	1325	+5	25
H	1321	+1	1
I	1320	0	0
J	1331	+11	121
$N=10$		$\sum d = 30$	$\sum d^2 = 712$

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{712}{10} - \left(\frac{30}{10}\right)^2}$$

$$= \sqrt{71.2 - 9}$$

$$= 7.89$$

Ans

Q-7: An analysis of production rejects resulted in the following figures. Calculate mean and standard deviation.

Sol:

No. of rejects per operator	mp (X)	No. of operator	$\frac{X-38}{5}$ (d)	fd	fd^2
20.5 - 25.5	23	5	-3	-15	45
25.5 - 30.5	28	15	-2	-30	60
30.5 - 35.5	33	28	-1	-28	28
35.5 - 40.5	38 → A	42	0	0	0
40.5 - 45.5	43	15	+1	15	15
45.5 - 50.5	48	12	+2	24	48
50.5 - 55.5	53	3	+3	9	27
		N = 120		$\sum fd = -25$	$\sum fd^2 = 223$

Mean: $\bar{x} = A + \frac{\sum fd}{N} \times i = 38 - \frac{25}{120} \times 5 = 38 - 1.04 = 36.96$

SD: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$

$= \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5 = 1.347 \times 5 = 6.375$ Ans

Q-9: The numbers of workers employed, the mean wage (Rs) per week and the standard deviation (Rs) in each branch of a company are given below. Calculate mean wages and standard deviation of all the workers taken together for company.

Branch	No. of workers employed	Weekly mean wage	SD
A	50	1413	60
B	60	1420	70
C	90	1415	80

sol:

$$\bar{x}_{123} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$= \frac{(50\times1413) + (60\times1420) + (90\times1415)}{50+60+90}$$

$$= 1416$$

$$d_1 = |\bar{x}_1 - \bar{x}_{123}| = |1413 - 1416| = 3$$

$$d_2 = |\bar{x}_2 - \bar{x}_{123}| = |1420 - 1416| = 4$$

$$d_3 = |\bar{x}_3 - \bar{x}_{123}| = |1415 - 1416| = 1$$

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2}{N_1 + N_2 + N_3}}$$

$$= \sqrt{\frac{50\times(60)^2 + 60\times(70)^2 + 90\times(50)^2 + 50\times 3^2 + 60\times 4^2 + 90\times 1^2}{50 + 60 + 90}}$$

$$= \underline{59.182}$$

Q-27: The price of a Tea company shares in Mumbai and Kolkata markets during the last ten months are recorded below. Determine the arithmetic mean and standard deviation of the prices of shares. In which market are the share prices stable?

Sol:

Mumbai			Kolkata		
x	$(x-\bar{x})$ x	x^2	y	$(y-\bar{y})$ (y)	y^2
105	-10	100	108	-11	121
120	+5	25	117	-2	4
115	0	0	120	+1	1
118	+3	9	130	+11	121
130	+15	225	100	-19	361
127	+12	144	125	+6	36
109	-6	36	125	+6	36
110	-5	25	120	+1	1
104	-11	121	110	-9	81
112	-3	9	135	+16	256
$\sum x=1150$	$\sum x=0$	$\sum x^2=694$	$\sum y=1190$	$\sum y=0$	$\sum y^2=1018$

Mumbai:

$$\bar{x} = \frac{\sum x}{N} = \frac{1150}{10} = 115$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{694}{10}} = 8.33$$

$$CV = \frac{8.33}{115} \times 100 = 7.24$$

Kolkata:

$$\bar{y} = \frac{\sum y}{N} = \frac{1190}{10} = 119$$

$$\sigma = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{1018}{10}} = 10.09$$

$$CV = \frac{\sigma}{y} \times 100 = \frac{10.09}{119} \times 100 = 8.48$$

Hence, $CV_{(Mumbai)} < CV_{(Kolkata)}$, Mumbai market shows greater stability.

Skewness, Moments and Kurtosis

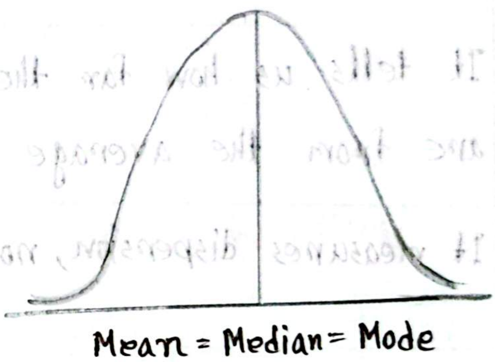
Skewness

Skewness means how much a distribution is tilted to one side instead of being perfectly symmetric.

Types of Skewness

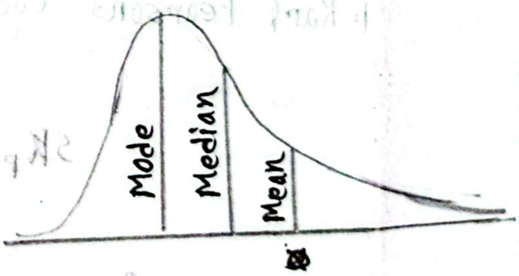
1. Zero Skewness (Symmetric)

- Both side equally balanced
- Mean = Median = Mode



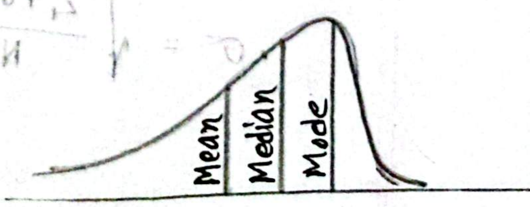
2. Positive Skewness (Right-skewed)

- Tail is longer on right side
- Mean > Median > Mode
- Few very large values; More small values



3. Negative Skewness (Left-skewed)

- Tail is longer on left side
- Mean < Median < Mode
- Few small values ; More ~~very~~ large values



Variation and skewness

Variation	Skewness
Variation means how spread out the data is	Skewness means how tilted or symmetric the data is
It tells us how far the values are from the average	It tells whether the data is leaning left or right
It measures dispersion, not shape	It measures shape, not spread

Measures of skewness

1. Karl Pearson's Coefficient of skewness:

$$SK_p = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

σ → standard Deviation

Or,

$$SK_p = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$\cdot\ \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

2. Bowley's Coefficient of Skewness :

$$Sk_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

$\bullet\ Q_1 = L + \frac{N/4 - p.cf}{f} \times i$ $\qquad \bullet\ Q_3 = L + \frac{3N/4 - pcf}{f} \times i$

$Q_2 \rightarrow$ Median $\qquad \bullet\ Q_2 = L + \frac{N/2 - pcf}{f} \times i$

3. Kelly's Coefficient of Skewness :

$$Sk_K = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$ (Based on percentiles)

$$Sk_K = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$$ (Based on deciles)

Q1: Calculate the coefficient of skewness from the data relate to the profits of 1000 companies

Sol:

Profits	Midpoint (x)	f	(x-170)/20 (d)	fd	fd^2
100-120	110	17	-3	-51	153
120-140	130	53	-2	-106	212
140-160	150	199	-1	-199	199
160-180	170	194	0	0	0
180-200	190	327	+1	+327	327
200-220	210	208	+2	+416	832
220-240	230	2	+3	+6	18
		N=1000		$\sum fd = 393$	$\sum fd^2 = 1741$

$$\bar{x} = A + \frac{\sum fd}{N} \times i = 170 + \frac{393}{1000} \times 20 = 170 + 7.86 = 177.86$$

Mode lies in the class 180-200

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 180 + \frac{133}{133+119} \times 20 = 190.56$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i = \sqrt{\frac{1741}{1000} - \left(\frac{393}{1000}\right)^2} \times 20$$

$$= \sqrt{1.741 - 0.15} \times 20 = 25.2$$

$$\therefore Sk_p = \frac{177.86 - 190.56}{25.2} = -0.504$$

It is a negatively skewed distribution.

Q-2: The following table gives the distribution of daily wages of 500 skilled workers in a factory

i) Obtain the limits of daily wages of central 50 percent of the observed workers.

ii) Calculate Bowley's Coefficient of Skewness.

Sol: i)

Daily Wages	f	cf
Below 200	10	10
200 – 250	25	35
250 - 300	145	180
300 – 350	220	400
350 – 400	70	470
400 and above	30	500

Q_1 = size of $\frac{N}{4}$ th observation $= \frac{500}{4} = 125^{th}$ observation

Q_1 lies in the class 250-300

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i$$

$$= 250 + \frac{125 - 35}{145} \times 50 = 250 + 31.03 = 281.03$$

Q_3 = size of $\frac{3N}{4}$ th observation $= \frac{3 \times 500}{4} = 375$ th observation

Q_3 lies in the class 300-350

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i$$

$$= 300 + \frac{375 - 180}{220} \times 50 = 300 + 44.32 = 344.32$$

Hence the daily wages of central 50% of workers lies between Rs 281.03 and Rs. 344.32

ii) Q_2 = size of $\frac{N}{2}$ th observation $= \frac{500}{2} = 250$ th observation

Q_2 = Median lies in the class 300-350

$$Q_2 = L + \frac{N/2 - pcf}{f} \times i = 300 + \frac{250-180}{220} \times 50 = 300 + 15.9 = 315.9$$

$$Sk_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

$$= \frac{344.32 + 281.03 - 2 \times 315.9}{344.32 - 281.03}$$

$$= \frac{-6.45}{63.29}$$

$$= -0.102$$

The distribution is skewed to the left.

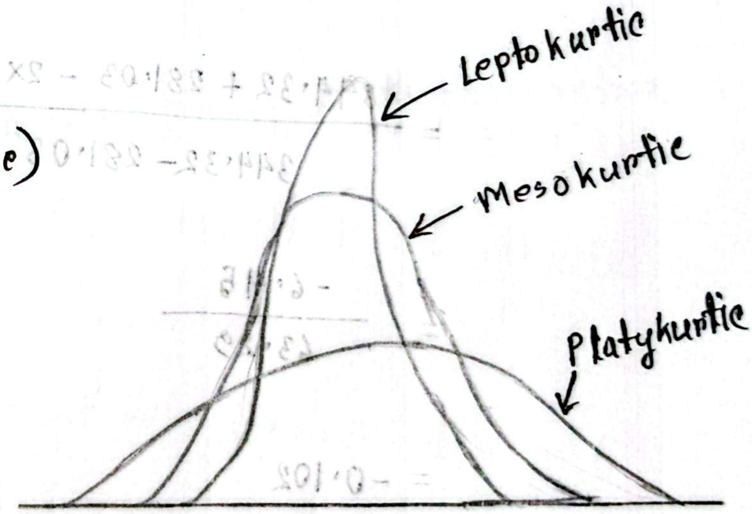
Kurtosis

Kurtosis shows how peaked or flat a distribution is compared to normal distribution.

Types of Kurtosis

1. Leptokurtic

- High and sharp peak.
- Heavy tails (More extreme value)
- $\beta_2 > 3$



2. Mesokurtic

- Normal shape
- Moderate peak
- $\beta_2 = 3$

3. **Platykurtic**

- Flat peak
- Light tails (less extreme values)
- $\beta_2 < 3$

Measures of Kurtosis

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

- $\gamma_1 = \frac{\mu_3}{\sigma^3}$ • $\sigma = \sqrt{\mu_2}$

Q-6: The first central moments of a distribution are 0, 16, -36 and 120. Comment on the skewness and kurtosis of the distribution.

Sol: $\mu_1 = 0$, $\mu_2 = 16$, $\mu_3 = -36$, $\mu_4 = 120$

$$\sigma = \sqrt{\mu_2} = \sqrt{16} = 4$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-36}{4^3} = -0.5625$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{120}{(16)^2} = 0.469 \text{ which is less than } 3$$

so, the distribution is negatively skewed and platykurtic

Q-7: An analysis of production rejects resulted in the following figures. Calculate mean, standard deviation and coefficient of skewness and comment on the results.

Sol:

No. of rejects per operator	Midpoints (x)	f	$(x-38)/5$ d	fd	fd^2
20·5 - 25·5	23	5	-3	-15	45
25·5 - 30·5	28	15	-2	-30	60
30·5 - 35·5	33	28	-1	-28	28
35·5 - 40·5	38	42	0	0	0
40·5 - 45·5	43	15	1	15	15
45·5 - 50·5	48	12	2	24	48
50·5 - 55·5	53	3	3	9	27
		$N_i = 120$		$\sum fd = -25$	$\sum fd^2 = 223$

Mean: $\bar{x} = A + \frac{\sum fd}{N} \times i = 38 - \frac{25}{120} \times 5 = 36.96 \approx 37$

standard deviation: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$

$= \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5$

$= 6.736 \approx 7$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 35.5 + \frac{14}{14+27} \times 5$$

$$= 35.5 + 1.71 = 37.21$$

$$Sk = \frac{36.96 - 37.21}{6.736} = -0.037$$

The mean value indicates that on ~~an~~ the average, rejects per operator were 37 in number.

The value of standard deviation says that the variation in the data from central value is approximately 7.

The coefficient of skewness indicates it is slightly left skewed.

Q-8: Distinguish between Karl Pearson's and Bowley's coefficient of skewness. Compute an appropriate measure of skewness for the data.

Sol:

Sales	f	cf
Below 50	12	12
50 - 60	30	42
60 - 70	65	107
70 - 80	78	185
80 - 90	80	265
90 - 100	55	320
100 - 110	45	365
110 - 120	25	390
Above 120	10	400

Since it is an open-ended distribution, therefore Bowley's method of calculating skewness should be more appropriate.

Q_1 = size of $\frac{N}{4}$th observation $= \frac{400}{4} = 100^{th}$ observation

Q_1 lies in 60-70

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 60 + \frac{100-42}{65} \times 10$$

$$= 60 + 8.92 = 68.92$$

Q_3 = size of $\frac{3N}{4}$ th observation $= \frac{3 \times 400}{4} = 300^{th}$ observation

Q_3 lies in the class 90-100

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i = 90 + \frac{300-265}{55} \times 10$$

$$= 90 + 6.36 = 96.36$$

Q_2 = size of $N/2$ th observation $= \frac{400}{2} = 200^{th}$ observation

Q_2 lies in the class 80-90

$$Q_2 = L + \frac{N/2 - pcf}{f} \times i = 80 + \frac{200-185}{80} \times 10$$

$$= 80 + 1.875 = 81.875$$

$$\text{Coefficient of Skewness} = \frac{96.36 + 68.92 - 2 \times 81.875}{96.36 - 68.92}$$

$$= \frac{165.28 - 163.75}{27.44}$$

$$= 0.056$$

Correlation Analysis

Correlation

- Correlation studies relationships between two or more variables where changes in one are associated with changes in others.
- Defined as analysis of covariation between variables.
- Coefficient of correlation (r) measures direction and degree of relationship.
- Steps in analysis:
  - → Determine and measure relation
  - → Test significance
  - → Establish cause-effect
- Requires paired observation.

Significance of Correlation Study

- Measures degree of relation in one figure
- Estimates one variable from another
- Aids economic understanding
- Reduces guesswork in business decisions (estimate costs from related variables)
- Advances science by revealing inter-related forces.
- Caution: Measures only linear strength ; does not imply causation.

Types of Correlation

i. Positive vs Negative Correlation

- Both variables move same direction ⟶ Positive
- Variables move opposite direction ⟶ Negative

ii. Simple, Partial, Multiple Correlation

- Simple : Involves one independent and one dependent variable
- Partial ; Involves three or more variables ; effect of one variable is controlled.
- Multiple: Involves three or more variables ; shows combined effect.

iii. Linear vs Non-linear Correlation

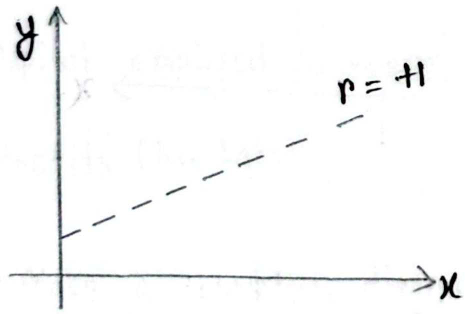
- Linear if constant ratio of change → plots straight linear
- Non-linear if varying ratio → plots curvilinear

Methods of Correlation

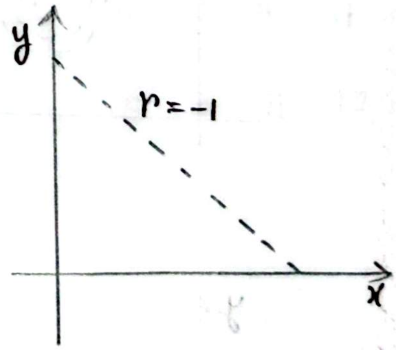
i. Scatter Diagram

- Plots pairs as dots on graph ; observe pattern

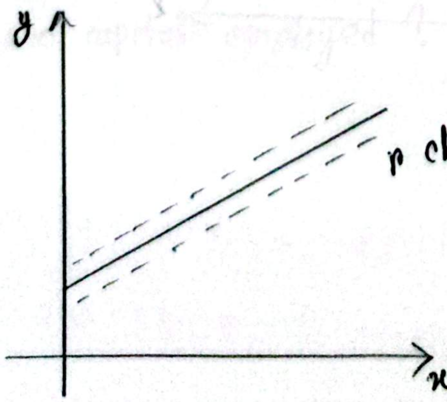
a) Perfect Positive Correlation



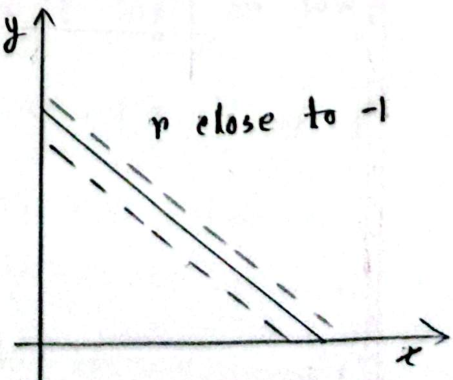
b) Perfect Negative Correlation



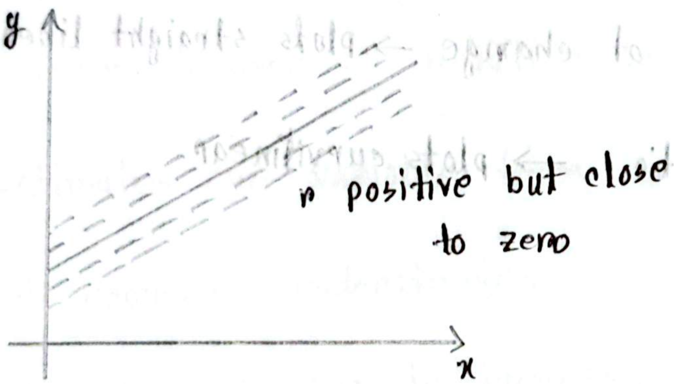
c) High Degree of Positive Correlation



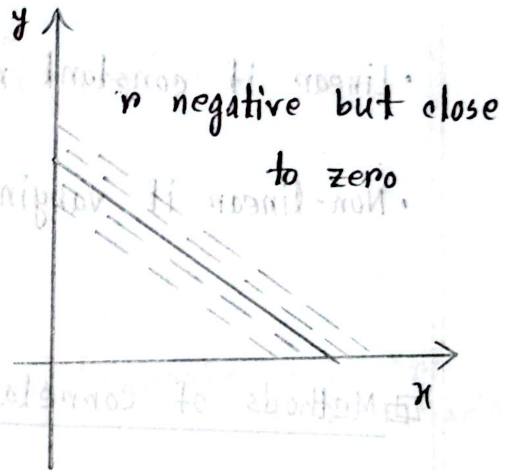
d) High Degree of Negative Correlation



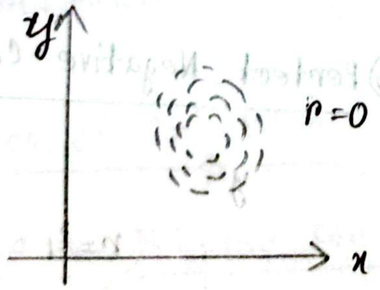
e) Low Degree of Positive Correlation

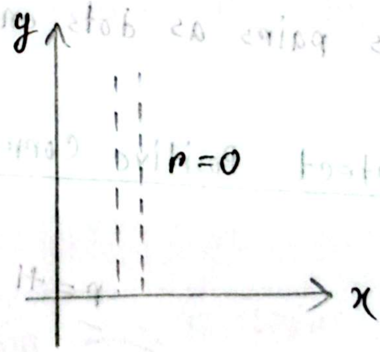


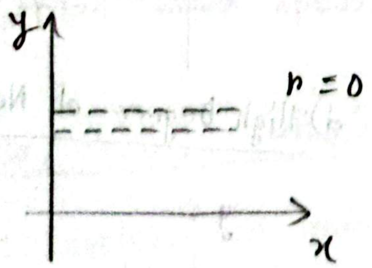
f) Low Degree of Negative Correlation

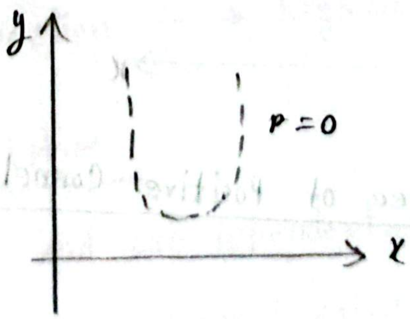


g) Zero Correlation









Merits:

- It is a simple and non-mathematical method of studying correlation. Easy to understand and ideas can be formed quickly.

- Not affected by extreme values.

- First step of studying relationship between variables

Limitations:

- Cannot establish exact degree → qualitative only

Q-1: Given the following pairs of values

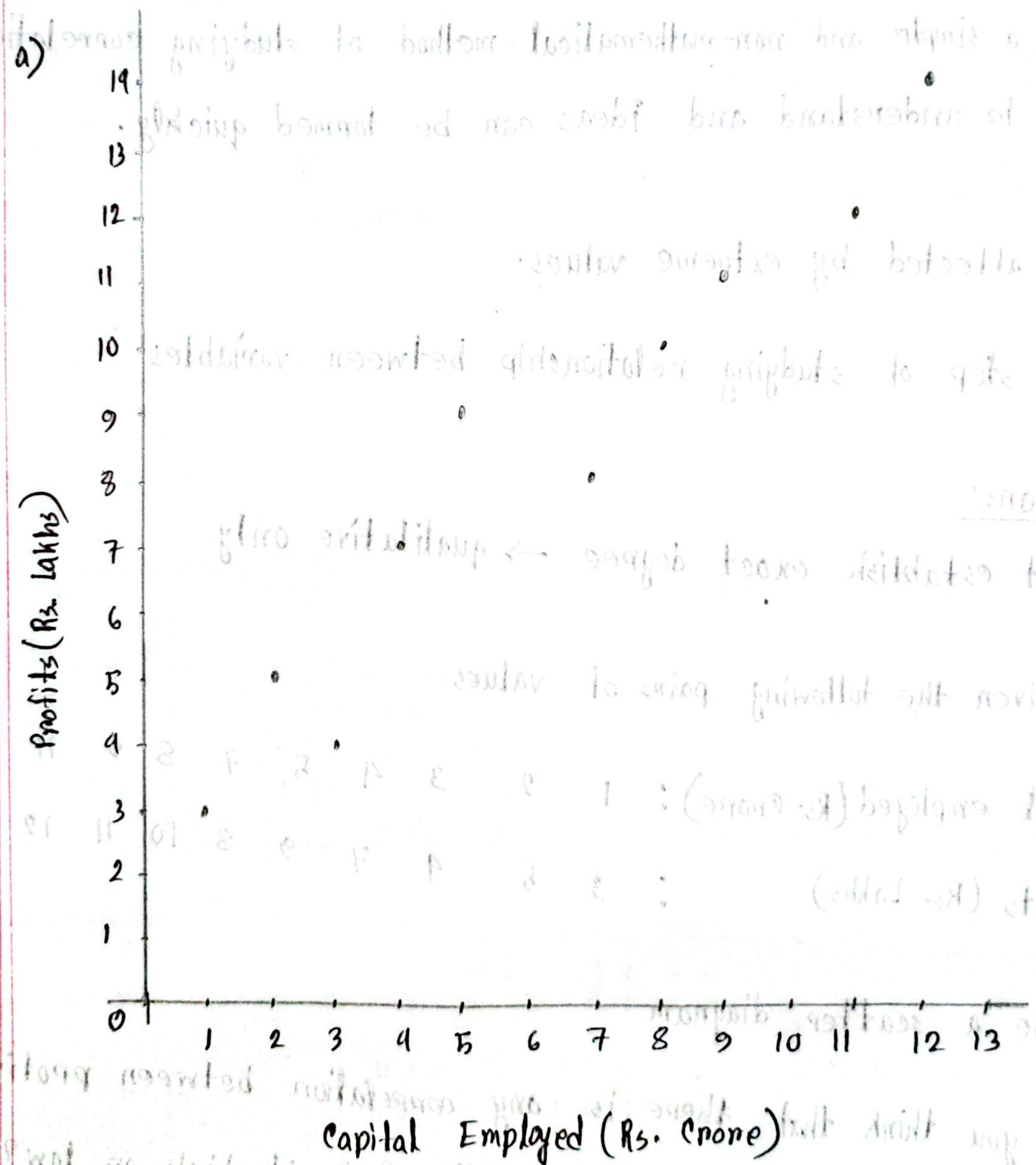
Capital employed (Rs Crore) :	1	2	3	4	5	7	8	9	11	12
Profits (Rs. Lakhs) :	3	5	4	7	9	8	10	11	12	14

a) Make a scatter diagram

b) Do you think that there is any correlation between profits and capital employed? Is it positive? Is it high or low?

Sol:

a)



b) ~~Vari~~ From the scatter diagram I can say —

- Variables are correlated
- Points are upward rising from left to right so positive
- The degree of relationship is high because the points are in a narrow band.

ii) Karl Pearson's Coefficient of Correlation

- Measure linear relationship ; r between -1 and $+1$

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \sum (Y-\bar{Y})^2}} \quad \text{—— (1)}$$

Let, $x = X-\bar{X}$; $y = Y-\bar{Y}$

$$r^* = \frac{\sum xy}{\sum x^2 \sum y^2}$$

$$r^{**} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{\sum x^2 - \frac{(\sum x)^2}{n}\right\}\left\{\sum y^2 - \frac{(\sum y)^2}{n}\right\}}}$$

Assumption

- The relationship between the two variables must be linear. Data points should form a straight line on scatter diagram.
- Both variables should be affected by numerous independent causes.
- There must be cause-and-effect linkage between the forces influencing the two variables.

CamScanner

Properties

- $-1 \leq r \leq +1$
- Independent of origin/scale change $\left[u = \dfrac{x-a}{i} \; ; \; v = \dfrac{y-b}{c} \right]$
- $r_{xy} = 0$, if variables independent
- $r = \sqrt{b_{yx} \times b_{xy}}$

$$* \quad u = \frac{x-a}{i} \quad ; \quad v = \frac{y-b}{c}$$

$$\Rightarrow \quad x = a + iu \quad ; \quad y = b + cv$$

$$\Rightarrow \bar{x} = a + i\bar{u} \quad ; \quad \bar{y} = b + c\bar{v}$$

$$\therefore r = \frac{\sum_i (u-\bar{u})(v-\bar{v})}{\sqrt{\sum_i (u-\bar{u})^2 \sum_i (v-\bar{v})^2}}$$

Q-2: Find correlation coefficient between the sales and expenses from the data given below:

Firm :	1	2	3	4	5	6	7	8	9	10
Sales (Rs. Lakhs) :	50	50	55	60	65	65	65	60	60	50
Expenses (Rs. Lakhs) :	11	13	14	16	16	15	15	14	13	13

Sol:

Firm	Sales (X)	$(X-\bar{X})$ x	x^2	Expenses (Y)	$(Y-\bar{Y})$ y	y^2	xy
1	50	-8	64	11	-3	9	+24
2	50	-8	64	13	-1	1	+8
3	55	-3	9	14	0	0	0
4	60	+2	4	16	+2	4	+4
5	65	+7	49	16	+2	4	+14
6	65	+7	49	15	+1	1	+7
7	65	+7	49	15	+1	1	+7
8	60	+2	4	14	0	0	0
9	60	+2	4	13	-1	1	-2
10	50	-8	64	13	-1	1	+8
N = 10	$\sum X = 580$	$\sum x = 0$	$\sum x^2 = 360$	$\sum Y = 140$	$\sum y = 0$	$\sum y^2 = 22$	$\sum xy = 70$

$$\bar{X} = \frac{\sum X}{N} = \frac{580}{10} = 58$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{70}{\sqrt{360 \times 22}} = \frac{70}{88.994} = 0.787$$

Hence, there is a high degree of positive correlation between the two variables.

Q-3: The following data relate to the age of 10 employees and the number of days which they reported sick in a month. Calculate Karl Pearson's coefficient of correlation and interpret its value.

Age :	20	30	32	35	40	46	52	55	58	62
sick days :	11	12	10	13	14	16	15	17	18	19

Sol:

Age X	$(X-\bar{X})$ x	x^2	Sick days Y	$(Y-\bar{Y})$ y	y^2	xy
20	-23	529	11	-3	9	+69
30	-13	169	12	-2	4	+26
32	-11	121	10	-4	16	+44
35	-8	64	13	-1	1	+8
40	-3	9	14	0	0	0
46	+3	9	16	+2	4	+6
52	+9	81	15	+1	1	+9
55	+12	144	17	+3	9	+36
58	+15	225	18	+4	16	+60
62	+19	361	19	+5	25	+95
$\sum X = 430$	$\sum x = 0$	$\sum x^2 = 1712$	$\sum Y = 145$	$\sum y = 5$	$\sum y^2 = 85$	$\sum xy = 353$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{430}{10} = 43$$

$$\bar{Y} = \frac{Y}{N} = \frac{145}{10} = 14.5 \approx 14$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{353}{\sqrt{1712 \times 85}} = 0.925$$

Thus, this is a high degree positive correlation between age and sick days.

Q-4: Find the coefficient of correlation by Karl Pearson's method between X and Y and interpret its value.

X :	57	42	40	33	42	45	42	44	40	56	44	43
Y :	10	60	30	41	29	27	27	19	18	19	31	29

Sol:

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
57	+13	169	10	-20	400	-260
42	-2	4	60	+30	900	-60
40	-4	16	30	0	0	0
33	-11	121	41	+11	121	-121
42	-2	4	29	-1	1	+2
45	+1	1	27	-3	9	-3
42	-2	4	27	-3	9	+6
44	0	0	19	-11	121	0
40	-4	16	18	-12	144	+48
56	+12	144	19	-11	121	-132
44	0	0	31	+1	1	0
43	-1	1	29	-1	1	+1
$\sum X = 528$	$\sum x = 0$	$\sum x^2 = 480$	$\sum Y = 340$	$\sum y = 10$	$\sum y^2 = 1828$	$\sum xy = -519$

$$\bar{X} = \frac{\sum x}{N} = \frac{528}{12} = 44$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{340}{12} = 28.33 \approx 28 \approx 30$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{-519}{\sqrt{180 \times 1828}} = -0.55$$

Thus, this is a moderate degree of negative correlation.

iii) Spearman's Rank Correlation Coefficient

- For ordinal data or non-quantifiable traits

- $$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

Q-7: Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The rankings are as follows:

Sol:

Employees	Rank by Manager-1 (R_1)	Rank by Manager-2 (R_2)	$(R_1 - R_2)^2$ (D^2)
A	10	9	1
B	2	4	4
C	1	2	1
D	4	3	1
E	3	1	4
F	6	5	1
G	5	6	1
H	8	8	0
I	7	7	0
J	9	10	1
N=10			$\sum D^2 = 14$

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 14}{10^3 - 10} = 1 - \frac{6 \times 14}{990}$$

$$= 1 - 0.085$$

$$= 0.915$$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers. A

Q-8: Two housewives, Geeta and Rita, asked to express their preference for different kinds of detergents, gave the following replies: To what extent the preferences of these two ladies go together?

Sol:

Detergent	Rank by Geeta (R_1)	Rank by Rita (R_2)	$(R_1 - R_2)^2$ (D^2)
A	4	4	0
B	2	1	1
C	1	2	1
D	3	3	0
E	7	8	1
F	8	7	1
G	6	5	1
H	5	6	1
I	9	9	0
J	10	10	0
N = 10			$\sum D^2 = 6$

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 6}{10^3 - 10} = 1 - \frac{36}{990} = 1 - 0.036 = 0.964$$

Thus the prefences of these two ladies agree very closely as far as their opinion on detergents is ~~connected~~ concerned.

Q-9: Calculate the rank correlation coefficient of on the following data of marks of 2 tests given to candidates for a clerical job.

Sol:

Preliminary test X	R_1	Final test Y	R_2	$(R_1 - R_2)^2$ (D^2)
92	10	86	9	1
89	9	83	7	4
87	8	91	10	4
86	7	77	5	4
83	6	68	4	4
77	5	85	8	9
71	4	52	2	4
63	3	82	6	9
53	2	37	1	1
50	1	57	3	4
$N = 10$				$\sum D^2 = 44$

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 44}{990} = 0.733$$

Thus, there is a high degree of positive correlation between preliminary and final test.

Regression Analysis

Regression

- After establishing that two variables are closely related, we may want to estimate or predict the values of one variable based on the value of another.
- Regression is the statistical tool used to estimate unknown values of one variable from known values of another.
- Example: If price (x) and demand (Y) are related, we can predict the most probable value of x for a given Y, or Y for a given x.

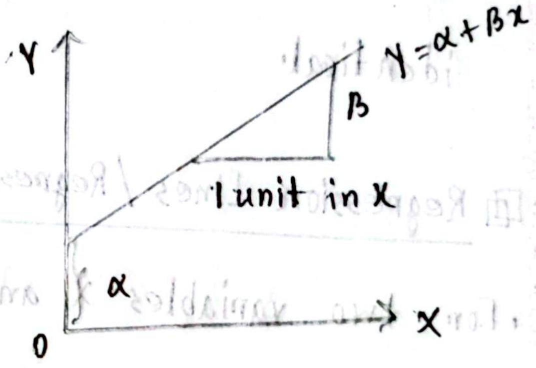
Difference Between Correlation & Regression

- Correlation measures the degree of relationship between x and Y. Regression studies the nature of the relationship.
- Regression clearly indicates cause-and-effect relations, unlike correlation

Linear Bivariate Regression Model

i) Dependent variable Y is random; independent variable x has fixed values selected and controlled by the experimenter.

ii) Average relationship between X and Y is described by a linear equation:



$$Y = \alpha + \beta x$$

intercept
slope

(a straight line geometrically)

iii) For each x value, there's a sub-population of Y values. Distribution may be normal or unknown.

iv) Mean of each sub-population (expected value of Y for given X):

$$E(Y|X) = \mu_{yx} = \alpha + \beta x$$

All means fall on a straight line under linearity assumption.

v) Individual Y value:

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

where e is the error term (stochastic disturbance), $E(e) = 0$.

If Y is normal, e is normal.

*Error: There always remain slight difference between the line of xy and ab graph. This is known as error.

vi) Variances of all sub-populations (variance of regression) are identical.

Regression Lines / Regression Model / Regression Equation

- For two variables X and Y, there are two regression lines:
  - → Regression line of Y on X: Gives most probable Y values for given X.
  - → Regression line of X on Y: Gives most probable X values for given Y.

- If perfect possitive or negative correlation ($r = \pm 1$), the two lines coincide (one line)
- Farther apart lines = lower correlation degree
- Closer lines = higher correlation degree
- If independent ($r = 0$), lines are at right angles (parallel to axes)

• Two equations:

a) Regression of Y on X:

$$y = \alpha + \beta x + \epsilon$$

Intercept

Error term

slope

Dependent variable

Independent variable

b) Regression of X on Y:

$$x = \alpha + \beta y + \epsilon$$

Intercept

Independent variable

Error term

slope

Dependent variable

Estimation of Parameters by Least Squares Method / Fitting

The Model

Regression equation of y on x is

$$y = \alpha + \beta x + \epsilon \qquad \text{———} \quad (*)$$

$$\Rightarrow \epsilon = y - \alpha - \beta x$$

$$\Rightarrow \sum \epsilon^2 = \sum (y - \alpha - \beta x)^2$$

$$\Rightarrow S = \sum (y - \alpha - \beta x)^2$$

Differentiating partially with respect to α and β & setting is equal to zero we get,

$$\frac{\partial S}{\partial \alpha} = 2 \sum (y - \alpha - \beta x)(-1) = 0$$

$$\Rightarrow \sum (y - \alpha - \beta x) = 0$$

$$\Rightarrow \sum y - N\alpha - \beta \sum x = 0 \qquad \text{———} \quad (i)$$

And

$$\frac{\partial S}{\partial \beta} = 2\sum_i (y-\alpha-\beta x)(-x) = 0$$

$$\Rightarrow \sum_i (y-\alpha-\beta x)x = 0$$

$$\Rightarrow \sum_i xy - \alpha \sum_i x - \beta \sum_i x^2 = 0 \quad \text{——— (ii)}$$

Dividing (i) and (ii) by N we get

$$\bar{y} - \alpha - \beta\bar{x} = 0 \quad \text{——— (iii)}$$

$$\frac{\sum_i xy}{N} - \alpha\bar{x} - \frac{\beta\sum_i x^2}{N} = 0 \quad \text{——— (iv)}$$

From (iii) $\times \bar{x}$ − (iv) we obtain,

$$\bar{x}\bar{y} - \alpha\bar{x} - \beta\bar{x} - \frac{\sum_i xy}{N} + \alpha\bar{x} + \frac{\beta\sum_i x^2}{N} = 0$$

$$\Rightarrow \beta\left(\frac{\sum_i x^2}{N} - \bar{x}^2\right) - \frac{\sum_i xy}{N} + \bar{x}\bar{y} = 0$$

$$\Rightarrow \beta\left(\frac{\sum_i x^2}{N} - \bar{x}^2\right) = \frac{\sum_i xy}{N} - \bar{x}\bar{y}$$

$$\Rightarrow \hat{\beta} = \frac{\sum_i (x-\bar{x})(y-\bar{y})}{\sum_i (x-\bar{x})^2}$$

from (iii) we get,

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$

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

Putting the value of $\hat{\alpha}$ and $\hat{\beta}$ in (*) we get the fitted regression model as

$$\hat{y} = \hat{\alpha} + \hat{\beta} \hat{x}$$

Regression Coefficients

a) Regression coefficient of Y on x (β_{yx}):

Measures change in Y per unit change in x

$$\beta_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2}$$

b) Regression coefficient of x on Y (β_{xy}):

Measures change in X per unit change in Y

$$\beta_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} = r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$

Properties of Regression Coefficients

i) Correlation coefficient r is the geometric mean:

$$r = \sqrt{\beta_{xy} \times \beta_{yx}}$$

ii) If one coefficient > 1, the other < 1 (since $|r| \leq 1$)

iii) Both coefficients have the same sign (positive or negative)

iv) r has the same sign as the coefficients

v) Average coefficient > r

$$\frac{(\beta_{xy} + \beta_{yx})}{2} > r$$

vi) Regression coefficients are independent of change of origin but not scale.

Proof: We know that,

$$\beta_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \quad \text{———} \quad (*)$$

origin

Let, $u = \dfrac{x-a}{h}$ and $v = \dfrac{y-b}{k}$

scale

$\Rightarrow x = a + hu \quad ; \quad y = b + kv$

$\Rightarrow \bar{x} = a + h\bar{u} \quad ; \quad \bar{y} = b + k\bar{v}$

Subtracting we get,

$x - \bar{x} = a + hu - a - h\bar{u} \qquad y - \bar{y} = b + kv - b - k\bar{v}$

$= h(u-\bar{u}) \qquad\qquad = k(v-\bar{v})$

Putting these values in $(*)$ we get,

$$\beta_{yx} = \frac{\sum hk(u-\bar{u})(v-\bar{v})}{\sum h^2 (u-\bar{u})^2}$$

$$= \frac{k}{h} \frac{\sum (u-\bar{u})(v-\bar{v})}{\sum (u-\bar{u})^2} = \frac{k}{h} \beta_{vu}$$

Similarly we have, $\beta_{xy} = \dfrac{h}{k} \beta_{uv}$

Hence the result.

Q-3: In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales. Calculate the regression equation of sales on test scores and estimate the probable weekly sales volume if a salesman makes a score of 100.

Sol:

Salesman	Test Score (X)	$(X-\bar{X})$ x	x^2	Sales (Y)	$(Y-\bar{Y})$ y	y^2	xy
1	40	-20	400	2.5	-1.5	2.25	+30
2	70	+10	100	6.0	+2.0	4.00	+20
3	50	-10	100	4.0	0	0	0
4	60	0	0	5.0	+1.0	1.00	0
5	80	+20	400	4.0	0	0	0
6	50	-10	100	2.5	-1.5	2.25	+15
7	90	+30	900	5.5	+1.5	2.25	+45
8	40	-20	400	3.0	-1.0	1.00	+20
9	60	0	0	4.5	+0.5	0.25	0
10	60	0	0	3.0	-1.0	1.00	0
$N=10$	$\sum X = 600$	$\sum x = 0$	$\sum x^2 = 2400$	$\sum Y = 40$	$\sum y = 0$	$\sum y^2 = 14$	$\sum xy = 130$

$$\bar{X} = \frac{\sum X}{N} = \frac{600}{10} = 60$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{10} = 4$$

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{130}{2400} = 0.054$$

The regression equation of sales and test score is given as:

$$Y - 4 = 0.054(X - 60)$$

$$Y = 0.76 + 0.054X$$

When X is 100. Y would be

$$Y = 0.76 + 0.054 \times 100 = 6.16$$

Q-25: In a correlation study the following values are obtained:

	X	Y
Mean :	65	67
SD :	2.5	3.5

Coefficient of Correlation, $r = 0.8$

Find the two regression equations.

Sol:

Regression equation of X on Y:

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\Rightarrow X - 65 = 0.8\frac{2.5}{3.5}(Y - 67)$$

$$\Rightarrow X - 65 = 0.571(Y - 67)$$

$$\therefore X = 0.571Y + 26.74$$

Regression equation of Y on X:

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$\Rightarrow Y - 67 = 0.8\frac{3.5}{2.5}(x - 65)$$

$$\Rightarrow Y - 67 = 1.12(x - 65)$$

$$\Rightarrow Y - 67 = 1.12x - 72.8$$

$$\therefore Y = 1.12x - 5.8$$