B.Com III sem.

QABD

Quantitative Analysis

For

Business Decisions - II

CORRELATION AND REGRESSION ANALYSIS - 1

When the relationship is a Quantitative nature, the appropriate statistical tool for discovering and measuring the relationship expressing it in brief formula — Correlation.

Eg:- Age of Husbands, Age of wife.
rainfall increasing then the production of rice
It determines.

Correlation Analysis:- The degree of relationship between the two variables under consideration.

Coefficient of correlation:- The numerical measure of correlation.

Covariation:- Increase in smoking causes and increase lung cancer.

would not prove that
smoking causes lung cancer.

That means the detection and analysis of correlation is called covariation.

Correlation examples:- The relation ship between
price and supply.
income and Expenditure.
students and marks / players.

Types of Correlation :- 1. +ve and -ve
2. Simple, partial and multiple
3. linear and non-linear

---

1. positive :- If the values of two variables deviate in the same direction.

Eg. X : 10 12 15 or X : 80 70 60
Y : 15 20 22 Y : 50 44 30

2. Negative :- If the values of the two variables deviate in the opposite direction.

X : 20 30 40 X : 100 90 60
Y : 40 30 22 Y : 10 20 30.

---

1. Simple :- When only two variables are studied. (which is in the same or opposite directions)
Eg:- height & weight
price & demand

2. Partial and multiple :- when we study three or more variables.
Eg:- Acc, Stat, Mat

---

1. linear :- The amount of change in one variable tends to bear constant ratio to the amount of change in other variable.

Eg:- 10 20 30 40 70
70 140 210 280 350.

2. Non-linear :- If the amount of change in one variable does not bear the const. ratio to the amount of change in other variable.

Eg:- If we double rainfall the production of rice or wheat would not necessarily be doubled.
of any correlation

Coefficient of Correlation :- The numerical measure of the amount of correlation existing between the two variables x and y.

Spearman's rank Correlation :- British Psychologist developed a formula to obtain the rank correlation coefficient in 1904

$$r_s = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

→ This method of rank correlation coefficient, the individual items of variables (x and y) are arranged in order of their ranks.

Common rank :- $\frac{\text{Normal ranks}}{\text{No. of ranks pooled}}$.

Probable Error :- Actual results - statistical results.

$$P.E = \text{probable factor} \times \text{Standard error}$$

$$= 0.6745 \times \left(\frac{1-r^2}{\sqrt{n}}\right)$$

Uses :-

1. It is used to determine the limits within which the population correlation coefficient may be expected to lie
2. It may be used to test if an observed value of Sample correlation coefficients is significant of any correlation in population.

Methods of studying correlation :-

1. Scatter Diagram method.
2. Graphic method.
3. ✓ Karl-pearson method
4. Concurrent deviation method.
5. ~~methods of least square~~.
   ✓ Spearman's Rank correlation

Karl-Pearson's coefficient of correlation.-

1867-1936, a great British biometrician & statistician

This formula based on A.M and S.D.

→ There is a possibility of linear relationship between the two variables.

The formula indicates whether the correlation is +ve or -ve.

$$r = \frac{\Sigma dx\, dy}{\sqrt{\Sigma dx^2}\ \sqrt{\Sigma dy^2}} \quad \text{(Actual mean)}$$

$$r = \frac{\Sigma dx\, dy - \frac{(\Sigma dx)(\Sigma dy)}{n}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{n}}\ \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{n}}} \quad \text{(Assumed mean)}$$

Uses:- To describing the degree of Correlation between two series.

Distinguish between correlation and Regression

Correlation	Regression.
1. It precedes regression.	1. It succeeds correlation
2. It tells the closeness between the two variables.	2. It studies the closeness between the two variables and estimates the values.
3. It measures the degree of covariation	3. It measures the nature of covariation.
4. It establishes just a relationship.	4. It studies the functional relationship with the two equations of lines.

Interpretation of correlation coefficient.

Karl-pearson coefficient of correlation lies between two limits +1 & -1

Value of correlation coefficient.	Interpretation.
(a) If $r = +1$	Perfect positive Correlation
(b) If $r = -1$	Perfect Negative Correlation
(c) r lies between +0.75 and +1	High Degree positive correlation
(d) r lies between -0.75 and -1	High Degree Negative correlation
(e) r lies between +0.25 and +0.75	Moderate Degree positive correlation
(f) r lies between -0.25 and -0.75	Moderate degree Negative Correlation
(g) r lies between 0 and +0.25	low degree positive correlation
(h) r lies between 0 and -0.25	low degree negative correlation
(i) when $r = 0$.	No correlation.

What are the characteristics of Karl-Pearson's coefficient of correlation.

1. It is based on arithmetic mean and standard deviation.
2. Determines the direction of relationship - positive or negative
3. Establishes the size of relationship
4. It is said to ideal measure of correlation.

write the merits of Karl-Pearson's Correlation coefficient.

(a) Takes into account all values
(b) more practical and popular.
(c) Numerically measures the 'r'
(d) measure the degree and direction of correlation
(e) Facilitates comparison
(f) Further algebraic treatment is possible.

write the demerits of Karl-Pearson's correlation efficient.

1. It assumes, linear relationship between variables even though there is no such relationship.
2. More time consuming.
3. It is affected by extreme items.
4. It is difficult to interpret.

4. What are the merits of Rank correlation

(a) Very simple and easy to understand.

(b) It is the only method applied in all the conditions

(c) Giving ranks is very simple which makes the fluctuations in the values of variables simple.

(d) When only ranks are available and not the values, this is the only method that can be used to find out co-variability or lack of it.

5. What are the demerits of Rank correlation?

1. It is not suitable for frequency distribution.
2. It is difficult to compute, when items increase beyond 20 or 30.
3. It lacks precision in results as compared to Karl pearson's method.

6. Differentiate between Correlation and Regression

Correlation Analysis refers to the technique used in measuring the closeness of the relationship between the variables.

Regression analysis is an absolute measure with the help of which unknown values are estimated from the known values of variables.

1 What are the Characteristics of Karl-Pearson's coefficient of correlation.

A 1. It is based on arithmetic mean and standard deviation.

2. Determines the direction of relationship - positive or negative

3. Establishes the size of relationship

4. It is said to ideal measure of correlation.

2. write the merits of Karl-Pearson's Correlation coefficient.

(a) Takes into account all values

(b) more practical and popular.

(c) Numerically measures the 'r'

(d) Measure the degree and direction of correlation

(e) Facilitates comparison

(f) Further algebraic treatment is possible.

3 write the demerits of Karl-Pearson's correlation coefficient

1. It assumes, linear relationship between variables even though there is no such relationship.

2. more time consuming.

3. It is affected by extreme items.

4. It is difficult to interpret.

Regression :- Act of returning or going back.

Regression :- Is the measure of the average relationship between two or more variables in terms of the original units of data.

Regression analysis :- It refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process.

Use :- The help of regression coefficients we can calculate the correlation coefficient.

Regression lines :- There are two regression lines.

1) Y on X
2) X on Y

Y on X :- The line gives the best estimate for the value of Y for any specified value of X

or

describe the variation in the value of Y for given changes in X

X on Y :- describe the variation in the value of X for given changes in Y

If the given data are plotted on a graph the points so obtained on the scatter diagram

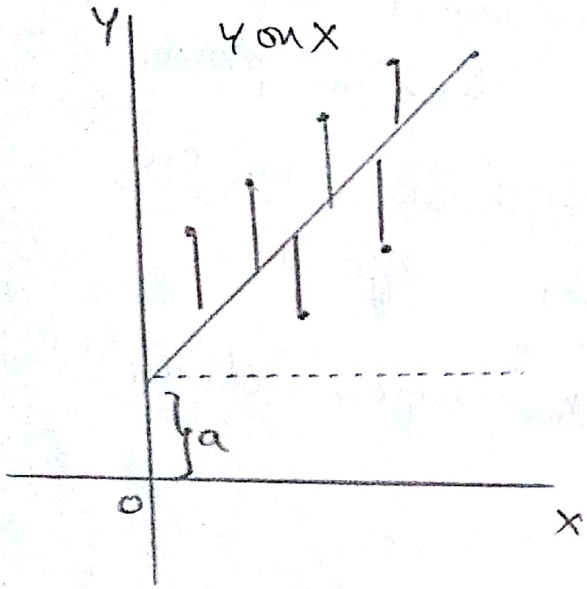
it is called Curve of regression

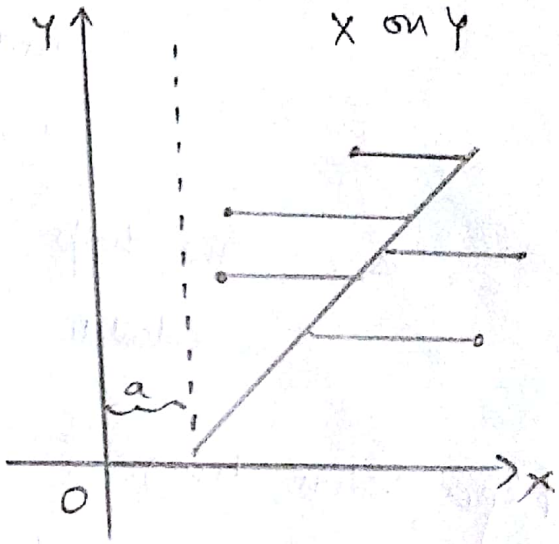
The curve is a line it is called linear regression

Y on X → $y = a + bx$ → a → y intercept

X on Y → $x = a + by$ → a → x intercept.

b → slope of the trend line.





$Y = a + bX$

$\Sigma Y = na + b\Sigma X$

$\Sigma XY = a\Sigma Y + b\Sigma Y^2$

$X = a + bY$

$\Sigma X = na + b\Sigma y$.

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Regression coefficients :- Deviations are taken from the actual means.

in case of secondary eq's Σx and $\Sigma y = 0$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \qquad b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

$y = bx$

$x = by$.

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad ; \quad (x - \bar{x}) = b_{xy}(y - \bar{y})$$

FORMULAS

1. Regression coefficient of Y on X (b_{yx})

$$b_{yx} = \frac{\text{Covariance}(x,y)}{\text{Variance}(x)} = r \cdot \frac{\sigma_y}{\sigma_x}$$

r = Coefficient of correlation

σ_x = Standard Deviation of x-series.

σ_y = Standard Deviation of y-series.

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

2. Regression coefficient of X on Y (b_{xy})

$$b_{xy} = \frac{\text{Covariance}(x,y)}{\text{Variance}(y)} = r \cdot \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2}$$

3. Regression equation Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

4. Regression equation X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

5. The regression coefficient $r = \sqrt{b_{xy} \times b_{yx}}$.

6. The regression coefficient $r = \sqrt{\frac{\Sigma xy}{\Sigma y^2} \times \frac{\Sigma xy}{\Sigma x^2}}$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

7. $$r = \frac{\Sigma xy}{n \times \sigma_x \times \sigma_y}$$

Probable Error :- $P = r \pm PE$.

$P \rightarrow$ correlation in Population.

Conditions for probable error :-

1. The data must approximately a normal frequency curve.
2. It must have been calculated from a sample
3. The sample must have been selected from an unbiased manner and the individual items must be independent.

1. $r < 6 \cdot P \cdot E_r$ the correlation is not all significant.
2. $r > 6 \cdot P E_r$ r is significant

Regression :- estimation or prediction of the unknown values of one variable from known values of another variable.

Regression Coefficient :- The slope of line of regression of one variable depending upon another variable.

Probable Error :- The difference between the results of samples (taken from the population) and the population.

2 Find the coefficient of correlation between the following two variables. Comment on result using the probable Error.

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

X	Y	$(x-18)=dx$	dx^2	$y-19=dy$	dy^2	$dx\,dy$
6	10	$6-18=-12$	144	$10-19=-9$	81	108
8	12	$8-18=-10$	100	$12-19=-7$	49	70
12	15	$12-18=-6$	36	$15-19=-4$	16	24
15	15	$15-18=-3$	9	$15-19=-4$	16	12
18	18	$18-18=0$	0	$18-19=-1$	1	0
20	25	$20-18=2$	4	$25-19=6$	36	12
24	22	$24-18=6$	36	$22-19=3$	9	18
28	26	$28-18=10$	100	$26-19=7$	49	70
31	28	$31-18=13$	169	$28-19=9$	81	117
$\Sigma X=162$	$\Sigma Y=171$	$\Sigma dx=0$	$\Sigma dx^2=598$	$\Sigma dy=0$	$\Sigma dy^2=338$	$\Sigma dx\,dy=431$

$$\bar{x}=\frac{\Sigma x}{n}=\frac{162}{9}=18$$

$$\bar{y}=\frac{\Sigma y}{n}=\frac{171}{9}=19$$

$$r=\frac{\Sigma dx\,dy}{\sqrt{\Sigma dx^2 \times \Sigma dy^2}}=\frac{431}{\sqrt{598\times 338}}=\frac{431}{\sqrt{202124}}=\frac{431}{449.582}$$

$$=0.95866827$$

2. Karl-Pearson method

3. Compute Karl-Pearson's coefficient of correlation between per capita National income and per capita consumer expenditure from the data given below.

Year	1999	2000	2001	2002	2003	2004	2005	2006
NI per capita	249	251	248	252	258	269	271	272
Per capita Con. Exp	237	238	236	240	245	255	254	252

2007	2008
280	275
258	251

Also calculate the probable error?

X	Y	dx X − 262	dy Y − 246	dx^2	dy^2	$dx\,dy$
249	237	−13	−9	169	81	117
251	238	−11	−8	121	64	88
248	236	−14	−10	196	100	140
252	240	−10	−6	100	36	60
258	245	−4	−1	16	1	4
269	255	+7	+9	49	81	63
271	254	+9	+8	81	64	72
272	252	+10	+6	100	36	60
280	258	+18	+12	324	144	216
275	251	+13	+5	169	25	65
$\Sigma X = 2625$	$\Sigma Y = 2466$	$\Sigma dx = 5$	$\Sigma dy = 6$	$\Sigma dx^2 = 1325$	$\Sigma dy^2 = 632$	$\Sigma dx\,dy = 885$

$$r = \frac{\Sigma dxdy - \frac{(\Sigma dx)(\Sigma dy)}{n}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{n}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{n}}}$$

$\Sigma dxdy = 885$

$\Sigma dx = 5$

$\Sigma dy = 6$

$\Sigma dx^2 = 1325$

$\Sigma dy^2 = 632$

$n = 10$

$$= \frac{885 - \frac{(5)(6)}{10}}{\sqrt{1325 - \frac{(5)^2}{10}}\sqrt{632 - \frac{(6)^2}{10}}}$$

$$= \frac{885 - 3}{\sqrt{1325 - 2.5}\sqrt{632 - 3.6}}$$

$$= \frac{882}{\sqrt{1322.5}\sqrt{628.4}} = \frac{882}{36.3662 \times 25.0679}$$

$$= \frac{882}{911.6244}$$

$= 0.9675$

Comment :- High degree positive correlation.

Probable Error $= 0.6745 \frac{1-r^2}{\sqrt{n}}$

$$= 0.6745 \left[\frac{1 - (0.9675)^2}{\sqrt{10}}\right]$$

$$= 0.6745 \left[\frac{1 - 0.9361}{3.16228}\right]$$

$$= 0.6745 \frac{0.0639}{3.16228} = 0.6745 \times 0.02021$$

$= 0.01363$

Calculation of 'R' when Ranks are given.

The ranks given by two Judges in a music competition of 8 singers are given below. Find the rank correlation coefficient.

Judge 1	2	1	4	3	5	8	7	6
Judge 2	1	3	2	8	7	4	5	6

Judge 1 R_1	Judge 2 R_2	$D = R_1 - R_2$	D^2
2	1	1	1
1	3	-2	4
4	2	2	4
3	8	-5	25
5	7	-2	4
8	4	4	16
7	5	2	4
6	6	0	0
			$\Sigma D^2 = 58$

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

$$= 1 - \frac{6(58)}{8^3 - 8} = 1 - \frac{348}{(512 - 8)}$$

$$= 1 - \frac{348}{504}$$

$$= 1 - 0.69047$$

$$= 0.309523.$$

Spearman's Rank Correlation.

Ten Competitors in a beauty Contest were ranked by three Judges as follows.

Judge A	3	5	4	10	8	8	1	6	8	2
Judge B	5.5	5.5	1	8.5	4	10	2	7	8.5	3
Judge C	9	9	7	5	2	2	2	5	5	9

			R_1-R_2		R_2-R_3		R_3-R_1	
R_1	R_2	R_3	d	d^2	d	d^2	d	d^2
3	5.5 ·	9.	-2.5	6.25	-3.5	12.25	6	36
5	5.5 ·	9.	-0.5	0.25	-3.5	12.25	4	16
4	1	7	3	9	-6	36	3	9
10	8.5 ×	5 ×	1.5	2.25	3.5	12.25	-5	25
8 *	4	2 ✓	4	16	2	4	-6	36
8 *	10	2 ✓	-2	4	8	64	-6	36
1	2	2 ✓	-1	1	0	0	+1	1
6	7	5 ×	-1	1	2	4	-1	1
8 *	8.5 ×	5 ×	-0.5	0.25	3.5	12.25	-3	9
2	3	9.	-1	1	-6	36	+7	49
				$\Sigma d^2=41$		$\Sigma d^2=193$		$\Sigma D^2=218$

For Judge A & B

8 is repeated 3 times
5.5 is repeated 2 times
8.5 is repeated 2 times.

$$R_{AB} = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[41 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{10(10^2-1)}$$

$$= 1 - \frac{6(44)}{990} = 1 - 0.02666 = 0.7334$$

(2) Judge B and Judge C

5.5 repeated 2 times
8.5 repeated 2 times
2 repeated 3 times
5 repeated 3 times
9 repeated 3 times

} 5 ranks are repeated.

$$R_{BC} = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[193 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3) + \frac{1}{12}(3^3-3) + \frac{1}{12}(3^3-3)\right]}{10(10^2-1)}$$

$$= 1 - \frac{6(200)}{900} = 1 - 1.2121 = -0.2121$$

(3) For Judge C and Judge A.

Rank 2 repeated 3 times
Rank 5 repeated 3 times.
Rank 9 repeated 3 times
Rank 8 repeated 3 times.

} 4 ranks are repeated.

$$R_{CA} = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[218 + \frac{1}{12}(3^3-3) + \frac{1}{12}(3^3-3) + \frac{1}{12}(3^3-3) + \frac{1}{12}(3^3-3)\right]}{10(10^2-1)}$$

$$= 1 - \frac{6(226)}{990} = 1 - 1.3696 = -0.36996.$$

Rank correlation coefficient is the maximum the pair of Judge AB has the nearest approach to taste in beauty.

From the following data obtain the regression equations X on Y and the regression equation Y on X

X	6	4	8	10	2
Y	9	8	7	5	11

Sol:-

$x = X - 6$; $y = Y - 8$

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
6	9	6-6=0	0	9-8=1	1	0
4	8	4-6=-2	4	8-8=0	0	0
8	7	8-6=2	4	7-8=-1	1	-2
10	5	10-6=4	16	5-8=-3	9	-12
2	11	2-6=-4	16	11-8=3	9	-12
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \quad ; \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8$$

Regression equation Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 8 = \frac{\Sigma xy}{\Sigma x^2}(X - 6)$$

$$Y - 8 = \frac{-26}{40}(X - 6)$$

$$Y - 8 = -0.65(X - 6)$$

$$Y - 8 = -0.65X + 3.9$$

$$Y = -0.65X + 3.9 + 8$$

$$\boxed{Y = -0.65X + 11.9}$$

Regression equation X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 6 = \frac{\Sigma xy}{\Sigma y^2}(Y - 8)$$

$$X - 6 = \frac{-26}{20}(Y - 8)$$

$$X - 6 = -1.3(Y - 8)$$

$$X - 6 = -1.3Y + 10.4$$

$$X = -1.3Y + 10.4 + 6$$

$$\boxed{X = -1.3Y + 16.4}$$

Time Series - 2

Time Series:- A Time series is a series of numerical data which have been recorded at different intervals of time. It is a record of changes in variables over a period of time.

Time Series Analysis:- The analysis of time series is of great significance not only to the economist and businessman but also to the scientist, geologist, research workers etc.,

1. It helps in understanding past behaviour.
2. It helps in planning future operations.
3. It help evaluating current accomplishments.
4. It facilitates comparison.

Components of time series:- Four basic types of variations

1. Secular trend → long term → (8 to 10 Y)
2. Seasonal variations → Short term → < 12 months (weekly)
3. Cyclical variations → medium term → 2 to 6 Y max 8 Y
4. Irregular (or) Random → Residual → Variations

The values of variables change suddenly or unexpectedly

Time series:- An arrangement of numerical data according time.

least square:- The sum of squares of deviation is least.

Computation of Trend Values :-

There are four methods commonly used in measuring the trend values

1. Graphic (Free hand curve fitting method)
2. method of semi average
3. method of moving average
4. least square method.

Trend Value :- The probable computed values having a specified or constant amount of change.

least square method :-

This is the method most accurate finding the trend values with the help of a mathematical technique which gives us a straight line trend.

* It is a line from which the actual values deviate on either sides. The sum of the deviations taken from the arithmetic mean will be zero. Consequently the sum of the squares of deviations will be least, as compared to the other alternatives

$$Y \text{ on } X \Rightarrow Y_c = a + bX$$

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

$Y_c \rightarrow$ Trend value
$a \rightarrow$ Y intercept when $X = 0$
$b \rightarrow$ slope of the trend line ; $X \rightarrow$ time.

Time series: Method of least squares

Below are given the figures of production of sugar factory.

Year:	2003	2004	2005	2006	2007	2008	2009
production (tonns):	80	90	92	83	94	99	92

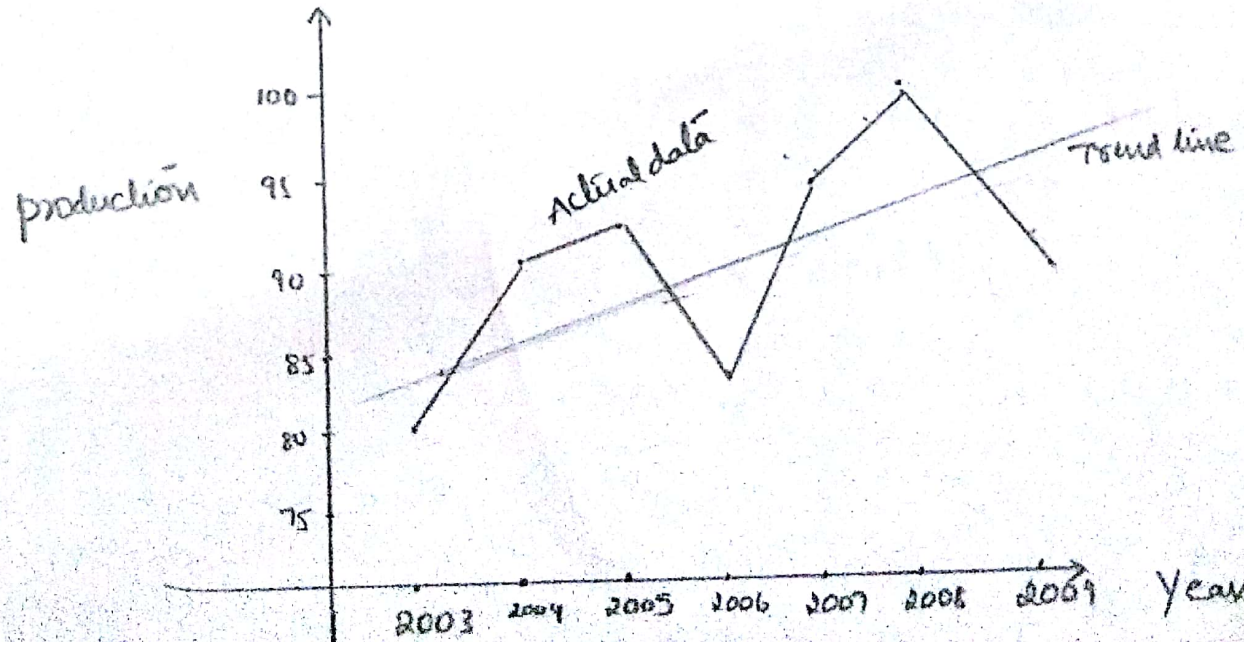
(i) Fit a straight line trend to these figures.

(ii) Plot these figures on a graph and show trend line

Year	production y	x	x^2	xy	Trend value $Y_c = a+bx$
2003	80	-3	9	-240	$90+2(-3) = 90-6 = 84$
2004	90	-2	4	-180	$90+2(-2) = 90-4 = 86$
2005	92	-1	1	-92	$90+2(-1) = 90-2 = 88$
2006	83	0	0	0	$90+2(0) = 90+0 = 90$
2007	94	1	1	94	$90+2(1) = 90+2 = 92$
2008	99	2	4	198	$90+2(2) = 90+4 = 94$
2009	92	3	9	276	$90+2(3) = 90+6 = 96$
$N=7$	$\Sigma y = 630$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xy = 56$	$\Sigma Y_c = 630$

$$a = \frac{\Sigma y}{N} = \frac{630}{7} = 90$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{56}{28} = 2$$



3. Interpolation and Extrapolation

Interpolation - Insertion of the most likely estimate under certain assumptions

Extrapolation - The estimation of the past or future figures that lie outside the two external points

Interpolation

1. estimation of figure within the given range of data.
2. The technique is also extended to know the past and future figures (outside the given range of data)

Extrapolation

1. Estimation of figures outside the given range of figures.
2. The technique can't be extended to know the figures within the given range of data.

Importance :-

1. The techniques are helpful in filling gaps that exist in statical data.
2. It is helpful in obtaining the median and mode in continuous series.
3. They are useful for forecasting the future events - sales, production, demand and other economic phenomena.

3. Interpolation and Extrapolation

Interpolation :- Insertion of the most likely estimate under certain assumptions

Extrapolation :- The estimation of the past or future figures that lie outside the two extreme points.

Interpolation

1. estimation of figure within the given range of data.
2. The technique is also extended to know the past and future figures (outside the given range of data)

Extrapolation

1. Estimation of figures outside the given range of figures.
2. The technique can't be extended to know the figures within the given range of data.

Importance :-

1. The techniques are helpful in filling gaps that exist in statical data.
2. It is helpful in obtaining the median and mode in continuous series.
3. They are useful for forecasting the future events - sales, production, demand and other economic phenomena.

Binomial Expansion method

This method is used for both the techniques interpolation and extrapolation.

1. The independent variable (x) advances by equal intervals like 2, 4, 6, 8 or 10, 15, 20, 25 or 6, 12, 18, 24, ...
2. The value of (x), for which the value of (y) is to be interpolated, is one of the class limits of x-series.

Newtons method of Advancing Differences -

The method is used only in interpolation

It is also applicable when the independent variable (x) advanced by equal intervals as in case of the Binomial Expansion method. However (x) value for which the (y) value to be interpolated need not be one of the class limits of X - Series.

The Newton method of Advancing Differences is based on finite differences.

Binomial Expansion method.

Known Values	Expansion of the formula	Symbol
2	$y_2 - 2y_1 + y_0 = 0$	$(y-1)^2 = 0$
3	$y_3 - 3y_2 + 3y_1 - y_0 = 0$	$(y-1)^3 = 0$
4	$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$	$(y-1)^4 = 0$
5	$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$	$(y-1)^5 = 0$
6	$y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0$	$(y-1)^6 = 0$
7	$y_7 - 7y_6 + 21y_5 - 35y_4 + 35y_3 - 21y_2 + 7y_1 - y_0 = 0$	$(y-1)^7 = 0$
8	$y_8 - 8y_7 + 28y_6 - 56y_5 + 70y_4 - 56y_3 + 28y_2 - 8y_1 + y_0 = 0$	$(y-1)^8 = 0$

Newton's method of Advancing Differences.

x	y	Δ^1	Δ^2	Δ^3	Δ^4
x_0	y_0	$y_1 - y_0$ (Δ_0^1)	$\Delta_1^1 - \Delta_0^1$ (Δ_0^2)	$\Delta_1^2 - \Delta_0^2$ (Δ_0^3)	$\Delta_0^3 - \Delta_1^3$ (Δ_0^4)
x_1	y_1	$y_2 - y_1$ (Δ_1^1)	$\Delta_2^1 - \Delta_1^1$ (Δ_1^2)	$\Delta_2^2 - \Delta_1^2$ (Δ_1^3)	
x_2	y_2	$y_3 - y_2$ (Δ_2^1)	$\Delta_3^1 - \Delta_2^1$ (Δ_2^2)		
x_3	y_3	$y_4 - y_3$ (Δ_3^1)			
x_4	y_4				

$$y = y_0 + x \cdot \Delta_0^1 + \frac{x(x-1)}{2!}\Delta_0^2 + \frac{x(x-1)(x-2)}{3!}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4!}\Delta_0^4 + \cdots$$

Binomial Expansion method.

The following data shown in the monthly average number of deaths under one year in Bangalore city Interpolate the monthly average number of deaths for year 1981.

Year :	1980	1981	1982	1983	1984
Monthly average number of Deaths :	940	-	907	843	798

Year	M.A. No. of Deaths.	
1980	940	y_0
1981	-	y_1
1982	907	y_2
1983	843	y_3
1984	798.	y_4

Known Values = 4.

Then Expansion of the formula is $(y-1)^4$

$$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$$

$$798 - 4(843) + 6(907) - 4y_1 + 940 = 0.$$

$$798 - 3372 + 5442 - 4y_1 + 940 = 0$$

$$3808 - 4y_1 = 0$$

$$3808 = 4y_1$$

$$\Rightarrow y_1 = \frac{3808}{4} = 952$$

Binomial Expansion method (2 Values missing)

Estimate the production for the year 1965 and 1975 from the following data

year :	1950	1955	1960	1965	1970	1975	1980
production :	100	120	150	–	210	–	320.
000 tonnes	y_0	y_1	y_2	y_3	y_4	y_5	y_6

Sol. There are two values missing

known values are 5.

The equation is $(y-5)^5 = 0$.

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0. \quad \text{— (1)}$$

Now change the subscripts.

$$y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 = 0 \quad \text{— (2)}$$

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$
$$y_6 - 5y_5 + 10y_4 - 10y_3 + 5y_2 - y_1 = 0$$

$$y_6 - 4y_5 + 5y_4 - 5y_2 + 4y_1 - y_0 = 0.$$

$$320 - 4(y_5) + 5(210) - 5(150) + 4(120) - 100 = 0$$

$$320 - 4y_5 + 1050 - 750 + 480 - 100 = 0$$

$$1000 - 4y_5 = 0$$

$$4y_5 = 1000 \Rightarrow y_5 = \frac{1000}{4} = 250.$$

$$y_5 = 250.$$

Substitute y_5 value (1) equation

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0.$$

$$250 - 5(210) + 10(y_3) - 10(150) + 5(120) - 100 = 0$$

$$250 - 1050 + 10y_3 - 1500 + 600 - 100 = 0$$

$$-1800 + 10y_3 = 0$$

$$-1800 = -10y_3$$

$$y_3 = \frac{1800}{10} = 180$$

In 1975 = 250000 tonns

In 1965 = 180000 tonns

Newton method of Advancing differences.

Below are given the wages by workers per month in a certain factory. Interpolate the number of workers earning in between 25/- and 35/-

Earning per day upto ₹	10	20	30	40	50	60
No. of workers	50	150	300	500	700	800

X	Y	Δ^1	Δ^2	Δ^3	Δ^4	Δ^5
10	50 (y_0)	100 (Δ^1_0)	50 (Δ^2_0)	0 (Δ^3_0)	-50 (Δ^4_0)	0 (Δ^5_0)
20	150 (y_1)	150 (Δ^1_1)	50 (Δ^2_1)	-50 (Δ^3_1)	-50 (Δ^4_1)	
30	300 (y_2)	200 (Δ^1_2)	0 (Δ^2_2)	-100 (Δ^3_2)		
40	500 (y_3)	200 (Δ^1_3)	-100 (Δ^2_3)			
50	700 (y_4)	100 (Δ^1_4)				
60	800 (y_5)					

Number of workers earning more than 35 ₹

$$x = \frac{35-10}{10} = \frac{25}{10} = 2.5.$$

$$Y_x = y_0 + x\Delta^1_0 + \frac{x(x-1)}{2}(\Delta^2_0) + \frac{x(x-1)(x-2)}{1\times2\times3}\Delta^3_0$$

$$+ \frac{x(x-1)(x-2)(x-3)}{1\times2\times3\times4}\Delta^4_0$$

$$+ \frac{x(x-1)(x-2)(x-3)(x-4)}{1\times2\times3\times4\times5}\Delta^5_0.$$

$$y_x = 50 + (2.5)(100) + \frac{(2.5)(2.5-1)}{1\times2}(50) + 0$$

$$+ \frac{(2.5)(2.5-1)(2.5-2)(2.5-3)}{1\times2\times3\times4}(-50) + 0.$$

$$= 50 + 250 + \frac{(2.5)(1.5)}{2}(50) + 0$$

$$+ \frac{(2.5)(1.5)(0.5)(-0.5)}{24}(-50) + 0$$

$$= 50 + 250 + (2.5)(1.5)(25) + 0 + \frac{(2.5)(1.5)(0.5)(-0.5)(-50)}{24} + 0$$

$$= 50 + 250 + 93.75 + 0 + 1.9531 + 0 = 395.7031$$

$$= 396 \text{ workers}$$

No. of workers earning more than 25/-

$$x = \frac{25-10}{10} = \frac{15}{10} = 1.5$$

$$y_x = 50 + (1.5)(100) + \frac{(1.5)(1.5-1)}{2}(50) + 0 +$$

$$\frac{(1.5)(1.5-1)(1.5-2)(1.5-3)}{1\times2\times3\times4}(-50) + 0.$$

$$= 50 + 150 + 18.75 + 0 - 1.1719 + 0 = 217.5781$$

$$= 218 \text{ workers}.$$

No. of workers earning more than ₹ 25 = 396

No. of workers earning more than ₹ 35 = 218

178.

Sampling and Sampling Distribution

Sample :- The set of observations that is taken from some source of observations for the purpose of obtaining information about the source is called sample

population :- The source of observations is called population

Sampling :- The study of a sample is refferred to as sampling.

Sampling units :- Items included in a population are called sampling units.

Census :- A survey in which the entire population is measured.

parameter :- A numerical descriptive measure of population

population :- An entire collection of units (individuals or objects or the list of measurements) about which we would like information.

Standard error :- A standard deviation of an estimate of some Quantity

Random Error :- An error arising due to chance.

Systematic Error :- An error arising due to bias views.

Sampling Error :- The error arising due to drawing inferences about the population on the basis of few observations is called sampling error.

methods to sampling :- 1. Probability Sampling (random)

2. non-probability sampling (non-random)

Sample size :- The number of elements included in the Sample is called sample size.

Theory of Probability - 5

✓A coin is tossed, A stone is thrown up in the air, A student has written an Examination, The sex of a baby to be born etc,

Experiment :- Any operation or act that results in two or more outcomes is called an Experiment.

Random Experiment :- When all the outcomes of an experiment can be enumerated but which particular outcome will result is not known, the experiment is called a random Experiment.

Event :- Constitutes one or more possible outcomes of an Experiment. It can be defined as a subset of the sample space.

Simple event :- It corresponds to a single possible outcome of an event.

Compound or Composite Event :- The joint occurence of two or more simple events is called a compound or composite event.

Independent events :- Independent events are those events, the occurence of which does not affect the occurence or non-occurence of other events.

Dependent events :- The occurence of which affects the occurence or non-occurence of other events.

Mutually Exclusive event :- Two events cannot happen simultaneously in a single trial.

$$A \cap B = \phi$$