ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category–I University by MHRD-UGC]
(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

DIRECTORATE OF DISTANCE EDUCATION

B.Sc. (Mathematics)

IV - Semester

113 44

STATISTICS

SYLLABI-BOOK MAPPING TABLE

STATISTICS

REFERENCE BOOKS:

1. Arumugam & Issac , Statistics, New Gamma Publishing House, 2007.
2. S.P.Gupta , Statistical Methods, Sultan Chand & Sons, 37th Edition, 2008.
3. Statistics by Dr.S.Arumugam and Mr. A. ThangapandiIssac, New Gamma Publishing House, Palayamkottai, June 2015.

CONTENTS

Course material prepared by
Dr. M.MUTHUSAMY M.Sc., M.Phil., Ph.D.
Assistant Professor
Department of Mathematics
Dr. Zakir Husain College
Ilayangudi - 630 702
Sivagangai District
Tamilnadu.

UNIT-I CENTRAL TENDENCIES

NOTES

1.1 INTRODUCTION

In this chapters we introduce several statistical constants which quantitatively describe some of the characteristics of a frequency distributions. These concepts are also helpful in comparing two similar frequency distribution.

The statistical constants that describe any given group of data are chiefly of four type viz.

(i) Measure of central tendency or measure of location.

(ii) Measure of dispersion

(iii) Measure of skewness

(iv) Measure of kurtosis

Here we introduce several commonly used measures of central tendencies

Definition. Measure of central tendency are "statistical constants which enable us to comprehend in a single effort the significance of the whole". Thus a measure of central tendency is a representative of the entire distribution. The following are the five measures of central tendencies which are in common use

1. Arithmetic mean(mean).

2.Median

3.Mode.

4.Geometric mean.

5.Harmonic mean.

1.2 ARITHMETIC MEAN

Definition. Arithmetic mean of n observations x_1, x_2, . . ., x_n is defined by

$$\bar{x} = \frac{x_1 + x_2 \ldots x_n}{n} = \frac{\Sigma_{x_i}}{n}$$

This definition is useful when n is so small that grouping f the values into a frequency distribution is not necessary.

Note: Suppose $x_1, x_2, \ldots, x_n$ be the distinct values of a variate with the corresponding frequencies $f_1, f_2, \ldots, f_n$

Then $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$ i=1,2,...,n

This maybe thought of as a weighted average where the weight of x_i is the corresponding frequency $f_{i.}$

Definition. Let $x_1, x_2, \ldots, x_n$ be n numbers. Suppose with each x_i there is associated a weight $w_{i.}$. then the weighted average or weighted means of $x_1, x_2, \ldots, x_n$ is defined by $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$ where i=1,2,...,n

The usual arithmetic mean $\bar{x}$ is the special case of the weighted arithmetic means where the corresponding frequencies

Example. Consider the 10 numbers 18, 15, 18, 16, 17,18,15, 19, 17, 17

Then $\bar{x} = \frac{18+15+18+16+17+18+15+19+17+17}{10}$

$= \frac{170}{10} = 17$

The frequency distribution for the above data is

x_i	15	16	17	18	19
f_i	2	1	3	3	1

$\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$ i=1,2,...,5

$= \frac{(2X\ 15)+(1\ X\ 16)+(3\ X\ 17)+(3\ X\ 18)+(1\ X\ 19)}{2+1+3+3+1}$

$= \frac{170}{10} = 17$

Suppose the variates $x_1, x_2, \ldots, x_{10}$ are assigned the weights 1,3,3,3,2,1,2,2,3,2 then the weighted average

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \text{ where i=1,2,...,10}$$

$$= \frac{18+45+54+48+34+18+30+38+51+34}{22}$$

$$= \frac{370}{22} = 16.82$$

Definition. The arithmetic mean (A.M) of a grouped frequency distribution is defined to be $\bar{x} = \frac{\sum f_i x_i}{N}$ where N=Σf_i and x_i is the mid-value of the true class interval.

Example. For the frequency distribution , the mid-value of the true class intervals x_i are given by 4.5, 14.5, 24.5, 34.5and 44.5 and the corresponding class frequencies are 11, 20, 16, 36 and 17 respectively.

$$\therefore \quad \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$$= \frac{(4.5\ X\ 11)+(14.5\ X\ 20)+(24.5\ X\ 16)+(34.5\ X\ 36)+(44.5\ X\ 17)}{11+20+16+36+17}$$

$$= \frac{49.5+290+392+1242+756.5}{100} = \frac{2730}{100}$$

$$= 27.30$$

Note. The calculation of arithmetic mean maybe considerably simplified arithmetically by shifting the origin of reference and at the same time altering the scale**.**

For example if we take A as the new origin and take h units of the variate x_i equal to one unit of the new variate u_i then $u_i = \frac{x_i - A}{h}$

(i.e) $x_i = hu_i + A$ Then the A.M. for the variate x_i is calculated as follows

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{\sum f_i (hu_i + A)}{N}$$

$$=h\left(\frac{\sum f_i u_i}{N}\right)+A\left(\frac{\sum f_i}{N}\right)$$

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

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

Example. For the frequency distribution , we have the following table for calculation of $\bar{x}$ by taking $u_i = \frac{x_i - 24.5}{10}$ where A=24.5 and h=10

Class	Mid x_i	Frequency f_i	u_i	f_iu_i
-0.5-09.5	04.5	11	-2	-22
0.95-19.5	14.5	20	-1	-20
19.5-29.5	24.5	16	0	0
29.5-39.5	34.5	36	1	36
39.5-49.5	44.5	17	2	34
Total	-	100	-	28

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

$$= 24.5+10 \text{ x}\left(\frac{28}{100}\right)=24.5+2.8$$

$$= 27.3$$

Theorem 1.1. The algebraic sum of the deviation of a set of n values from their arithmetic mean is zero.

Proof. Let $x_{1,}$ $x_{2,}$...,x_n be the values with frequencies $f_{1,}$ $f_{2,}$...,f_n respectively.

$$\therefore \bar{x}=\frac{\sum f_i x_i}{N} \text{ where } N=\Sigma_{f_i}$$

The deviation of x_i from the A.M. is given by $d_i= x_i - \bar{x}$(i=1,2,......,n)

$$\therefore \ \Sigma_{f_i d_i}= \Sigma_{f_i}(x_i-\bar{x}) = \sum f_i x_i - \bar{x}\sum f_i = N\bar{x} - N\bar{x}= 0$$

Hence the theorem.

NOTES

Theorem1.2. The sum of the squares of the deviations of a set of n values is minimum when the deviations are taken from their mean.

Proof: Let $x_{1,}x_2....,x_n$ be the set of n values with the corresponding frequencies $f_{1,}f_2,.....,f_n$

$$\therefore \bar{x} = \frac{\sum f_i x_i}{N} \text{ where } N=\Sigma_{f_i}$$

Now , the sum of the squares of the deviations of x_i from an arbitrary number A is given by $Z=\Sigma_{f_i}(x_i-A)^2$

The value of A for which Z is minimum is determined by the condition $\frac{dz}{dA}= 0$ and $\frac{d^2Z}{dA^2}>0$

Now $\frac{dz}{dA}=0 \quad => \; -2\Sigma_{f_i}(x_i-A)=0$

$$=> \sum f_i x_i - N.A = 0$$

$$=> A = \frac{\sum f_i x_i}{N}$$

$$=> A= \bar{x}$$

Also $\frac{d^2Z}{dA^2} =2\sum f_i =2N>0$

$\therefore Z$ is minimum when $A= \bar{x}$

Hence the theorem.

Theorem1.3 If $x_{1,}$ $x_{2,}...,x_k$ are the arithmetic means of $n_{1,}n_{2,}n_{3,}.....,n_k$ observations then the arithmetic mean of the combined set of observations is given by $\bar{x}= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ...+n_k\bar{x}_k}{n_1+n_2+\cdots+n_k}$

Proof. $n_1\bar{x}_1$ is the sum of all the n_1 observation in the first set.

$n_2\bar{x}_2$ is the sum of all the n_2 observation in the second set.

... ...

$n_k\bar{x}_k$ is the sum of all the n_k observations in the k^{th} set.

$\sum_{i=1}^{k} n_i\bar{x}_i$ is the sum of all the ($n_{1+}n_{2+}$.....$+n_k$) observations in the combined set.

$$\bar{x}=\frac{1}{N}\left(\sum_{i=1}^{k} n_i\bar{x}_i\right) \text{ where } N=\sum_{i=1}^{k} n_i. \text{ Hence the theorem.}$$

Solved Problems.

Problem 1. The heights of 10 students in c.m's chosen at random are given by 164, 159, 162, 168, 165, 170, 168, 171, 154, 169 Calculate A.M.

Solution. Here n=10

$$\bar{x}=\frac{1}{10}\left(\sum X_i\right)=\frac{1}{10}(1690) = 169 \text{ c.m.}$$

Problem 2. Calculate A.M. from the following frequency

Weight in Kgs	50	48	46	44	42	40
No. of persons	12	14	16	13	11	09

Solution. We have the following table.

Weight in Kgs x_i	**No.of persons** f_i	f_ix_i
50	12	600
48	14	672
46	16	736
44	13	572
42	11	462
40	09	360
Total	**75**	**3402**

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

NOTES

$= \frac{3402}{75}$

$=45.36$

Aliter. Choosing A=46 as the origin and h=2 as the scale so that

$u_i = \frac{x_i - 46}{2}$, we get the following table.

x_i	**f_i**	**u_i**	**$f_i\ u_i$**
50	12	2	24
48	14	1	14
46	16	0	0
44	13	-1	-13
42	11	-2	-22
40	09	-3	-27
Total	**75**	**-**	**-24**

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

$= 46 + \text{x} \left(\frac{-24}{75}\right)$

$=45.36$

Problem 3. Calculate the A.M for the following frequency distribution of the marks obtained by 50 students in a class.

Marks	**No. of students**	**Marks**	**No. of students**
05-10	5	25-30	5
10-15	6	30-35	4
15-20	15	35-40	3
20-25	10	40-45	2

Solution. Let us choose A = 22.5 as the origin and h = 5 as the scale,

$u_i = \frac{x_i - 22.5}{5}$ and we get the following table.

Class	Mid x_i	u_i	f_i	$f_i u_i$
05-10	7.5	-3	5	-15
10-15	12.5	-2	6	-12
15-20	17.5	-1	15	-15
20-25	22.5	0	10	0
25-30	27.5	1	5	5
30-35	32.5	2	4	8
35-40	37.5	3	3	9
40-45	42.5	4	2	8
Total	**-**	**-**	**50**	**-12**

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

$$= 22.5 + 5\left(\frac{-12}{50}\right) = 22.5 - 1.2$$

= 21.3

Problem 4. Find the mean mark of students from the following table.

Marks	No. of students
0 and above	30
10 and above	26
20 and above	21
30 and above	14
40 and above	10
50 and above	0

Solution. We express the above data in the form of a frequency table as follows :

Marks	*Mid* x_i	No. of students f_i	f_i x_i
00-10	5	4	20
10-20	15	5	75
20-30	25	7	175
30-40	35	4	140
40-50	45	10	450
50-	-	0	-
Total	**-**	**30**	**860**

$$\therefore \bar{x} = \frac{\sum f_i x_i}{N} = \frac{860}{30} = 28.67$$

NOTES

Problem 5. Calculate (i) mean price (ii) weighted mean price of the following food articles from the table given below.

Article of food	Quantity in Kgs	Price per Kg.
Rice	30	4.50
Wheat	10	2.75
Sugar	5.5	6.25
Oil	3.5	16.50
Flour	4.5	4.00
Ghee	1.5	40.00
Onion	9	3.25

Solution.

Article of food	**Price per Kg in Rs. x_i**	**Quantity in Kgs W_i**	**x_i w_i**
Rice	4.5	30	135.00
Wheat	2.75	10	027.50
Sugar	6.25	5.5	034.38
Oil	16.50	3.5	057.75
Flour	4.00	4.5	018.00
Ghee	40.00	1.5	060.00
Onion	3.25	9	029.25
Total	77.25	64.0	361.88

Mean Price = $\frac{\sum x_i}{N} = \frac{77.25}{7}$ = Rs.11.0

weighted mean price = $\frac{\sum w_i x_i}{N} = \frac{361.88}{64}$ = Rs.5.65

Problem 6. The four parts of a distribution are as follows.

	Frequency	Mean
Part 1	50	61
Part 2	100	70
Part 3	120	80
Part 4	30	83

Find the mean of the entire distribution.

Solution. $n_1 = 50; n_2 = 100; n_3 = 120; n_4 = 30$

$\bar{x}_1$=61; $\bar{x}_2 = 70$; $\bar{x}_3 = 80$; $\bar{x}_4 = 83$

Now, $\bar{x}= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n_1+n_2+n_3+n_4}$

$$= \frac{(50\ X\ 61)+(100\ X\ 70)+(120\ X\ 80)+(30\ X\ 83)}{50+100+120+30}$$

$$= \frac{22140}{300}$$

$$= 73.8$$

∴ Mean of the entire distribution is73.8

Problem 7. Mean weight of 80 students in two classes A and B is 50 kgs. There are 45 students in class A. The mean weight of the students in class B is 48. Find the mean weight of the students in class A.

Solutions. Here $n_1 = 45; n_2 = 35; \bar{x}_1$=50; $\bar{x}_2$=48

We need to find $\bar{x}_1$ from the formula $\bar{x}=\frac{n_1x_1+n_2x_2}{n_1+n_2}$

$$\therefore 50 = \frac{45\bar{x}_1+35 \text{ x } 48}{80}$$

45$\bar{x}_1$=(80 x 50)-(35 x 48)=4000-1680

= 2320

∴ $\bar{x}_1$ =51.56 kgs.

∴Mean weight of the students in class A is 51.56 kgs.

Problem 8. The average weight for a group of 40 students was calculated to be 58 kgs. It was later discovered that weight of one student was misread as 75 kgs instead of the correct weight 57 kgs. Find the correct average.

Solutions. Total weight of 40 students = 40 x 58 = 2320

Total weight after correction = 2320-75+57=2302

∴ After correction, the average = $\frac{2302}{40} = 57.55$

Problem 9. Show that (i) the A.M. of the first n natural numbers is $\frac{1}{2}(n+1)$. (ii) the weighted A.M. of first n natural numbers whose weights are equal to the corresponding numbers is equal to $\frac{1}{3}$(2n+1)

Solutions. (i) A.M. of the first n natural numbers = $\frac{\sum x_i}{n}$

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

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

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

(ii) The required Weighted A.M $= \frac{\sum w_i x_i}{\sum w_i}$

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

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

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

Problem 10. The frequencies of values 0,1,2,.....,n of a variable are given respectively by 1,$n_{c_1,}n_{c_2}$,......,n_{c_n}. Show that the mean is $\frac{1}{2}$n

Solution. $\sum_{i=1}^{n} f_i$ =1+n_{c_1}+n_{c_2}+......+n_{c_n}=$(1+n)^n$=2^n

$\sum_{i=1}^{n} f_i x_i$ = 0 + $(\mathbf{n_{c_1}})$ + 2$(\mathbf{n_{c_2}})$+3$(\mathbf{n_{c_3}})$+....+n$(\mathbf{n_{c_n}})$

= n+2$\left[\frac{n(n-1)}{1.2}\right]$ + 3 $\left[\frac{n(n-1)(n-2}{1.2.3}\right]$+........+n

= n$\left[1+(n-1)+\left[\frac{(n-1)(n-2}{1.2}\right]+\cdots+1\right]$

= n$\left[1+(n-1)_{c_1}+(n-1)_{c_1}+\cdots+(n-1)_{c_{n-1}}\right]$

= n$(1+1)^{n-1}$

= n 2^{n-1}

$\therefore \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{n2^{n-1}}{2^n}$ = $\frac{1}{2}$n

Exercises.

1. Calculate the A.M. from the following data.

Value	1	2	3	4	5	6	7	8	9
Frequency	7	11	16	17	26	31	11	1	1

2. Find the mean mark of the following frequency distribution

Marks	0-10	10-20	20-30	30-40	40-50
No . of students	3	5	9	3	2

3. Find the mean mark of students from the following table

Marks	**No. of students**	**Marks**	**No. of students**
0 and above	80	60 and above	28
10 and above	77	70 and above	16
20 and above	72	80 and above	10
30 and above	65	90 and above	8
40 and above	55	100 and above	0
50 and above	43		

4.Find the mean for the following data

(i)

Class	**Frequency**	**Class**	**Frequency**
0-9	32	50-59	167
10-19	65	60-69	98
20-29	100	70-79	46
30-39	184	80-89	20
40-49	228	90-99	0

(ii)

Temperature centigrades	No. of days	Temperature centigrades	No. of days
-40 to-30	10	0-10	65
-30 to-20	28	10-20	180
-20 to-10	30	20-30	10
-10to0	42	Total	365

NOTES

1.3 PARTITION VALUES (MEDIAN, QUARTILES, DECILES AND PERCENTILES)

Partial values are those values of the variate which divide the total frequency into a number of equal parts. Some important partition values are median quartiles, deciles and percentiles.

Median

Median of a frequency distribution is the value of the variate which divides the total frequency into equal parts. In other words median is the value of the variate for which the cumulative frequency is $\frac{1}{2}N$ where N is the total frequency.

In the case of ungrouped data if n values of the variate are arranged in ascending or descending order of magnitude the median is the middle value if n is odd and it is taken as the arithmetic mean of the middle values if n is even.

Example. Consider the values54,81,84,71,61,57,58,54,56,67,49 Arranging these values in ascending order of magnitude we get

49,54,54,56,57,61,67,68,71,81,84

Since there are 11 items 6th item , namely 61, is the median.

Note. In the case of the discrete frequency distribution we calculate the median as follows1. Calculate $\frac{1}{2}N = \frac{1}{2}\sum f_i$

2. Find the cumulative frequency just greater then$\frac{1}{2}N$

3. The corresponding value of the variate is the median**.**

Example. Consider the following discrete frequency distribution.

x	f	Less than c.f
1	5	5
2	9	14
3	18	32
4	12	44
5	9	53
6	7	60
Total	60	-

Here N=60 Hence$\frac{1}{2}N$=30

The values of x for which the c.f is just greater than 30 is given by x=3

$\therefore$x=3 is the median of the frequency distribution.

We now derive a formula for calculating the median in the case of a grouped frequency distribution.

Definition. For a grouped frequency distribution the **median class** is defined to be the class where the less than cumulative frequency is just greater than $\frac{1}{2}N$

Quartiles.

Definition. Consider a frequency distribution with the total frequency N. The value of the variate for which the cumulative frequency is N/4 is called the first quartile or lower quartile and it is denoted by $Q_{1.}$

Similarly, the value of variate for which the cumulative frequency is 3N/4 is called the third quartile or upper quartile and it is denoted by $Q_{3.}$

Clearly, median is the second quartile and it can also we denoted by $Q_{2.}$

In the case of ungrouped data with n items Q_1 is calculated as follows.

Let i=$\left[\frac{1}{4}(n+1)\right]$ = the integral part of $\frac{1}{4}(n+1)$

Let q=$\frac{1}{4}(n+1) - \left[\frac{1}{4}(n+1)\right]$. Hence q is the fractional part.

Then $Q_1 = x_i + q(x_{i+1} - x_i)$. Similarly $Q_3 = x_i + q(x_{i+1} - x_i)$

where i=$\left[\frac{3}{4}(n+1)\right]$and q=$\frac{3}{4}(n+1)$- $\left[\frac{3}{4}(n+1)\right]$

In this case of grouped frequency distribution the quartiles are calculated by using the formula

$$Q_1 = l + \frac{({N}/{4-m})h}{f_k} \text{ and } Q_3 = l + \frac{({3N}/{4-m})h}{f_k}$$

where l is the lower limit of the class in which the particular quartile lies, f_k is the frequency of this class, h is the width of the class and m is the cumulative frequency of the preceeding class.

Deciles.

Consider a frequency distribution with total frequency N. The value of the variates for which the cumulative frequencies are $\frac{iN}{10}(i = 1,2,....,9$ are called **deciles.** The ith decile is denoted by Di Clearly median is the fifth decile. Hence the median can also be denoted by D_5

In the case of the ungrouped data with n items, for k=1,2,3,......,9 $D_k = x_i + q(x_{i+1} - x_i)$ where i = $\left[\frac{k(n+1)}{10}\right]$ and q=$\frac{k(n+1)}{10} - \left[\frac{k(n+1)}{10}\right]$

As before for a grouped frequency distribution we can prove that $D_i = l + \frac{({iN}/{10-m})h}{f_k}$; (i=1,2,......,9) with corresponding notations.

Percentiles

Percentiles are the values of variates for which the cumulative frequencies are$\frac{iN}{100}$; (i=1,2,.....,9) and the i^{th} percentile is denoted by P_i Clearly median is 50^{th} percentile and hence median can alsobe denoted by P_{50}

In this case of ungrouped data with n items, for k=12,....,99

$P_k = x_i + q(x_{i+1} - x_i)$ where i = $\left[\frac{k(n+1)}{100}\right]$ and q=$\frac{k(n+1)}{100} - \left[\frac{k(n+1)}{100}\right]$

Percentile are got from the following formulae in the case of grouped frequency distribution $P_i = l + \frac{({iN}/{100-m})h}{f_k}$; i=1,2,......,99

NOTES

Solved Problems

Problems 1. Find the median and quartiles of the heights in c.m. of eleven students given by 66,65,64,70,61,60,56,63,60,67,62

Solution. Arranging the given data in ascending order of magnitude we get 56, 60, 60, 61, 62, 63, 64, 65, 66, 67, 70

Here n= 11. Since n is odd, median is the sixth item which is equal to 63

Q_1 =size of $\frac{1}{4}(n+1)^{\text{th}}$ item.

$\therefore$ Q_1 = third item = 60

$\therefore Q_1$ = $\frac{3}{4}(n+1)^{\text{th}}$ item = 9th item = 66

Problem 2. Find the median and quartile marks of 10 students in Statistics test whose marks are given as 40,90,61,68,72,43,50,84,75,33

Solution. Arranging in ascending order of magnitude we get 33,40,43,50,61,68,72,75,84,90 Here n = 10

Hence median is the average of the two middle items viz 61 and 68

$\therefore$Median = $\frac{1}{2}$(61+68) = 64.5 marks.

First quartile.

Here $\left[\frac{1}{4}(n+1)\right] = 2\ and\ q = \frac{1}{4}(n+1)$ - $\left[\frac{1}{4}(n+1)\right]$ = .75

$\therefore Q_1 = x_2 + .75\ (x_3 - x_2)$ = 40 +.75 (43-40)=42.5

Third quartile. $\left[\frac{3}{4}\ (n+1)\right]$ = 8 and q=$\frac{3}{4}\ (n+1) - \left[\frac{3}{4}\ (n+1)\right]$ =.25

$\therefore Q_3 = x_8 + .25\ (x_9 - x_8)$ = 75 + .25 (84-75)=77.25

Problem 3. From the following data calculate the percentage of tenants paying monthly rent (i) more than 105 (ii)between 130 and 190

Monthly rent	**No. of tenants**	**Monthly rent**	**No. of tenants**
60-80	18	140-160	88
80-100	21	160-180	75
100-120	45	180-200	18
120-140	85	**Total**	**350**

Solution. (i) Number of tenants paying more than Rs.105 is

$$=\left(\frac{120-105}{20}\right) \text{ x } 45+85+88+75+18$$

$$= 34+266 = 300\text{(approximately)}$$

$$\therefore \text{ Required percentage} = \frac{300}{350} \text{ x } 100$$

$$=85.7 \text{ (approximately)}$$

(ii) No. of tenants paying the rent between Rs.130 and Rs.190

$$= \left(\frac{140-130}{20}\right) \text{ x } 85+88+75+\left(\frac{190-180}{20}\right) \text{ x } 18$$

$$= 42.5 + 88 +75 + 9 =215 \text{ (approximately)}$$

$$\therefore \text{Required percentage} = \frac{215}{350} \text{ x } 100$$

$$= 61.43$$

Problem 4. An incomplete distribution is given below.

Class	Frequency	Class	frequency
0-10	10	40-50	?
10-20	20	50-60	25
20-30	?	60-70	15
30-40	40	**Total**	**170**

The median is 35. Find the missing frequencies.

Solution. Let the frequency corresponding to the class 20-30 be f_1 and that of class 40-50 be f_2

$$\therefore f_1 +f_2 =170 - (10+20+40+25+15)$$

$$\therefore f_1 +f_2 =60$$

Now, the median 35 lies in the median class 30-40

$\therefore$ L=30 ; M=10+20 + f_1 ; $f_{K=}$ 40 and h=10

We have median = l $+\left(\frac{85-(10+20+f_1)}{40}\right)$ x 10

$\therefore 35 = 30+\frac{1}{4}(55-f_1)$

$\therefore f_1 = 35$

Using (2) in(1) we get $f_2 = 25$

$\therefore f_1 = 35$ and $f_2 = 25$ are the missing frequencies.

NOTES

Exercises

1. Obtain the median for the following frequency distribution.

x:	1	2	3	4	5	6	7	8	9
f:	8	10	11	16	20	25	15	9	6

2.Find the median lower and upper quartiles, 8^{th} decile, and 56^{th} percentile for the following distribution of245 workers.

Monthly wages	**No. of workers**	**Monthly wages**	**No. of workers**
1-2.99	6	9-10.99	21
3-4.99	53	11-12.99	16
5-6.99	85	13-14.99	4
7-8.99	56	15-16.99	4

3.Calculate the median of the following series.

Wages in Rs.	**No of workers**
More than 100	5
More than 90	17
More than 80	37
More than 70	43
More than 60	49
More than 50	49
More than 40	51

4. Find the three quartiles, for the following distribution.

Marks	No of Students	Marks	No. of students
5-10	5	25-30	5
10-15	6	30-35	4
15-20	15	35-40	2
20-25	10	40-45	2

1.4 Mode

In a distribution the value of the variate which occurs most frequently and around which the other values of variates cluster densely is called the **mode or modal** value of the distribution.

In the case of a discrete frequency distribution mode is the value of the variate corresponding to the maximum frequency.

Example. Consider the discrete frequency distribution

x:	1	2	3	4	5	6	7	8	9	10
f:	8	13	47	105	28	9	5	3	2	1

Here the maximum frequency is 105

The value corresponding to this maximum frequency is 4. Hence mode is 4

In the case of grouped frequency distribution the mode is computed by the formula **Mode=** $\boldsymbol{l + \frac{(f-f_1)h}{2f-f_1-f_2}}$

Where l is the lower boundary of the modal class (class having maximum frequency); f is the maximum frequency ; f_1 and f_2 are the frequencies of the classes proceeding and following the modal class ; h is the width of the class.

An alternate formula for finding the mode is also given by

Mode = l +$\boldsymbol{\frac{h f_2}{f_1+ f_2}}$ with the above notations.

Note 1. In the case of irregularities in the distribution or when the maximum frequencies are repeated or the maximum frequency occurs in the very beginning or at the end, the modal class is determined by the method of grouping and then the mode is got by using any one of the formulae.

Note 2. A frequency distribution may have more than one mode in which it is called **multimodal** distribution. If there is only one mode it is called **unimodal** distribution.

Note 3. There is interesting empirical relationship between mean, median, mode which appears to hold for unimodal curves of moderate asymetry namely.

Mean - Mode = 3(Mean - Median) (i.e) **Mode = 3 Median - 2 Mean.**

Solved problems.

Problem 1. The following are the heights in c.m. of 10 students. Calculate the modal height 63,65,66,65,64,65,65,61,67,68

Solution. Since 65 occurs 4 times and no other item occurs 4 or more than four times 65 c.m. is the modal height.

Problem 2. Calculate the modal for the frequency distribution given in solved problem 3 in 2.2

Solution. Here maximum frequency 52 occurs in the class 30.5 - 35.5 (refer table in page 36) which is the modal class.

$$\therefore\ l = 30.5\ ;\ f_1 = 47\ ;\ f_2 = 41 \text{ and } h=5$$

$$\therefore \text{ Mode} = \mathbf{l} + \frac{hf_2}{f_1+f_2} = 30.5\ \frac{5\times 41}{47+41} = 30.5 + \frac{205}{88}$$

$$= 32.83$$

Problem 3. Given that the mode of the following frequency distribution of 70 students is 58.75 Find the missing frequencies f_1 and f_2

Class	frequency
52-55	15
55-58	f_1
58-61	25
61-64	f_2

Solution.

Class	f
52-55	15
55-58	f_1
58-61	25
61-64	f_2

Since N = 70 we get $f_1 + f_2 = 30$ ....(1)

Mode = l +$\frac{h(f-f_1)}{2f-f_1-f_2}$

$\therefore 58.75 = 58 + \frac{3\times(25-f_1)}{50-f_1-f_2}$

$\therefore 0.75 = \frac{3\times(25-f_1)}{50-f_1-f_2}$

$\therefore 3f_{1-}f_2$=50 (verify) ....(2)

From (1) and (2) we get $f_1 = 20$ and $f_2 = 10$

Exercises.

1.Find the mean, median and mode for the set of numbers.

(i) 6, 8,,2, 5, 9, 5, 6, 5, 2, 3

(ii) 61.7, 71.8, 65.3, 70, 69.8

2. In a moderately asymmetrical distribution if mean is 24.6 and mode is26.1 find the median.

3. In a skewed distribution mean and median are respectively 33 and 34.5. Find the mode.

4. Find the mean, median and mode of the following frequency distribution.

Class	**Frequency**	**Class**	**frequency**
20-24	3	40-44	12
25-29	5	45-59	6
30-34	10	50-54	3
35-39	20	55-59	1

5. Calculate the mode from the data given below.

Wages in Rs.	**Number of workers**	**Wages in Rs.**	**Number of workers**
Above 30	520	Above 70	104
Above 40	470	Above 80	45
Above 50	399	Above 90	7
Above 60	210	Above 100	0

6. Calculate the mode from the following frequency distribution.

X	1-9	9-17	17-25	25-33	33-41	41-49	49-57
f	20	31	27	15	10	7	8

UNIT-II GEOMETRIC MEAN AND HARMONIC MEAN

2.1 GEOMETRIC MEAN

NOTES

Definition. The **geometric mean** (G.M.) of a set of n observations $x_1, x_2, \dots \dots, x_n$ is the n^{th} root of their product. Thus, geometric mean is

$$G=(x_1, x_2, \dots \dots, x_n)^{\frac{1}{n}}$$

$$\therefore \log G = \frac{1}{n}(log x_1 + log x_2 + \dots\dots + log x_n)$$

$$= \frac{1}{n}\sum log x_i$$

$$\therefore G = \text{antilog}\left[\frac{\sum log x_i}{n}\right]$$

In case of a grouped frequency distribution geometric mean

$$G=\left(x_1^{f_1}\ x_2^{f_2} \dots \dots \dots x_n^{f_n}\right)^{1/N} \text{ where } N=\sum f_i$$

As before we can write G= $anti\ log\ \left[\frac{1}{N}(\sum f_i\ \ log x_i)\right]$

2.2 HARMONIC MEAN

Harmonic mean (H.M) of the set of n observations $x_1, x_2, \dots \dots, x_n$ is defined to be the reciprocal of the arithmetic mean of the reciprocal of the observations

Thus harmonic mean H= $\frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots.+\frac{1}{x_n}\right)}$

In the case of a grouped frequency distribution harmonic mean

$$H=\frac{1}{\frac{1}{N}[\sum(f_i/x_i)]} \text{ where } N=\sum f_i$$

Solved Problems.

Problem 1. Find the G.M. and H.M. of the four numbers 2, 4, 6, 27

Solution. G.M. $= G = (2 \times 4 \times 6 \times 27)^{1/4}$

$$= [2 \times (2 \times 2) \times (2 \times 3) \times (3 \times 3 \times 3)]^{1/4}$$

$$= (2^4 \times 3^4)^{1/4} = 6$$

$$\text{H.M} = H = \frac{1}{\frac{1}{4}\left(\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{27}\right)} = \frac{4 \times 108}{103} = 4.19$$

Problem 2. Find the G.M and H.M of the following distribution.

X :	1	2	3	4	5
f :	2	4	3	2	1

Solution. G.M. $= G = (1^2 \times 2^4 \times 3^3 \times 4^2 \times 5)^{1/12}$

$$= (16 \times 27 \times 80)^{1/12}$$

$$= \text{antilog}\left[\frac{1}{12}(\log 34560)\right]$$

$$= \text{antilog}\ (.3782) = 2.384$$

$$\text{H.M} = H = \frac{\mathbf{12}}{\mathbf{2\left(\frac{1}{1}\right) + 4\left(\frac{1}{2}\right) + 3\left(\frac{1}{3}\right) + 2\left(\frac{1}{4}\right) + 1\left(\frac{1}{5}\right)}}$$

$$= \frac{12}{2 + 2 + 1 + \frac{1}{2} + \frac{1}{5}}$$

$$= 2.11$$

Problem 3. Find the G.M for the following frequency distribution.

Marks	0-10	10-20	20-30	30-40
No. of students	5	8	3	4

NOTES

Solution.

Mid x_i	Frequency f_i	$log_{10}\ x_i$	$f_i log_{10}\ x_i$
5	5	0.6990	3.4950
15	8	1.1761	9.4088
25	3	1.3979	4.1937
35	4	1.5441	6.1764
Total	**20**	**-**	**23.2739**

$$G = \text{antilog}\left[\frac{1}{N}\left(\sum f_i \log x_i\right)\right]$$

$$=\text{antilog}\left[\frac{1}{N}(23.2739)\right]$$

$$=\text{antilog}\ (1.1637)$$

$$= 14.38$$

Problem 4. Find the H.M. for the following frequency distribution.

Class	0-10	10-20	20-30	30-40	40-50
Frequency	15	10	7	5	3

Solution.

Mid x_i	f_i	$1/x_i$	f_i/x_i
5	15	.2000	3.0000
15	10	.0667	0.6670
25	7	.0400	0.2800
35	5	.0286	0.1430
45	3	.0222	0.0666
Total	**40**	**-**	**4.1566**

$$\text{Harmonic mean} = H = \frac{1}{(1/N)\sum(f_i\ /x_i)} = \frac{40}{4.1566}$$

Problem 5. Calculate the average speed of a train running at the rate of 20 k.m per hour during the first 100 k.m., at 25 k.m.p.h. during the second 100 k.m and at 30 k.m.p.h. during the third 100 k.m.

Solution. Clearly weighted H.M. is the proper average

$$\text{Weighted H.M} = \frac{\sum w_i}{\sum(w_i/x_i)}$$

$$= \frac{100+100+100}{\frac{100}{20}+\frac{100}{25}+\frac{100}{30}} = 24.32 \text{ k.m.p.h(verify)}$$

Exercises

1. Calculate A.M., G.M. and H.M. of the following observations and show that A.M > G.M > H.M. 32,35,36,37,39,41,43

NOTES

2. Calculate the H.M of the following series of monthly expenditure in Rs.of a batch of students 125,130,75,10,45,0.5,0.4, 500,15

3. Find the G.M for the distribution.

Class	**Frequency**	**Class**	**Frequency**
0-9	32	50-59	167
10-19	65	60-69	98
20-29	100	70-79	46
30-39	184	80-89	20
40-49	288	**Total**	**1000**

4. Find the H.M of the following distribution.

x	2	3	4	5	6
f	5	7	11	9	8

5. Calculate the G.M and H.M of the following frequency distribution.

Class	2-4	4-6	6-8	8-12
frequency	20	40	30	10

UNIT-III MEASURES OF DISPERSION

3.1 INTRODUCTION

NOTES

You have learnt various measures of central tendency. Measures of central tendency help us to represent the entire mass of the data by a single value. Can the **central tendency** describe the data fully and adequately? In order to understand it, let us consider an example.

The daily income of the workers in two factories are :

Factory A:	35	45	50	65	70	90	100
Factory B:	60	65	65	65	65	65	70

Here we observe that in both the groups the mean of the data is the same, namely, 65

(i) In group A, the observations are much more scattered from the mean.

(ii) In group B, almost all the observations are concentrated around the mean. Certainly, the two groups differ even though they have the same mean.Thus, there arises a need to differentiate between the groups. We need some other measures which concern with the measure of scatteredness (or spread).

To do this, we study what is known as **measures of dispersion.**

3.2 MEASURES OF DISPERSION

Definition. Dispersion of a distribution is the amount of scatteredness of the individual values from a measure of central tendency. There are four measures of dispersion which are in common use. They are as follows

(i) Range (ii) Quartile (iii) Mean deviation (iv) Standard deviation.

Range

It is the simplest method of studying dispersion. Range is the difference between the smallest value and the largest value of a series. While computing range, we do not take into account frequencies of different groups.

Example. The maximum value is 49 and the minimum value is 1. Hence the range is 48

Quartile Deviations (Q.D.) (Semi inter quartile range)

The quartile deviation (Q.D) or semi inter quartile range is defined by Q.D=$\frac{1}{2}(Q_3 - Q_1)$ where

Q_1 and $Q_3$are the first and the thord quartiles of the distribution.

Example. Q_1 =17 and $Q_3 = 38.5$

Hence Q.D =$\frac{\mathbf{1}}{\mathbf{2}}(38.5 - 17) = 10.75$

Mean Deviation. The mean deviation of a frequency distribution from any average A is defined by M.D.= $\frac{\sum f_i|x_i-A|}{N}$ where N=$\sum f_i$

NOTES

Example.For the data, $\bar{x} = 27.3$(refer example under A.M. in section 1.2) Now.

Mid x_i	f_i	$\mathbf{|x_1 - 27.3\ |}$	$\mathbf{f_1|x_1 - 27.3\ |}$
4.5	11	22.8	250.8
14.5	20	12.8	256.0
24.5	16	02.8	044.8
34.5	36	07.2	259.2
11.5	17	17.2	292.4
Total	**100**	**-**	**1103.2**

$\therefore$M.D about mean =$\frac{1103.2}{100} = 11.032$

Standard Deviation.

A common measure of dispersion which is preferred in most circumstances in statistics is the standard deviation.

Definition. The **Standard deviation σ** of a frequency distribution is defined by σ=$\left[\frac{\sum f_i(x_{i-}\bar{x})^2}{N}\right]^{1/2}$ where N=$\sum f_i$ and $\bar{x}$ is the arithmetic mean of the frequency distribution.

The square of the standard deviation of a frequency distribution is called the variance of the frequency distribution. Hence **variance** =σ^2

Note. If ${\sigma^2}_x$ is the variance of a sample of size n the "best" estimate for the population variance ${\sigma^2}_x$ is not ${\sigma^2}_x$

But $\left(\frac{\mathrm{N}}{\mathrm{N-1}}\right){\sigma^2}_x$ For this reason many authors define standard deviation by the formula σ=$\left[\frac{\sum f_i(x_{i-}\bar{x})^2}{N}\right]^{1/2}$

For large values of N the two formulae for standard deviation are practically indistinguishable. Throughout this book we use the first formula for finding standard deviation of a frequency distribution. Both the formulae for standard deviation find place in modern calculators.

Definition. The **root mean square deviation** of a frequency distribution is defined to be s=$\left[\frac{\sum f_i(x_{i-}A)^2}{N}\right]^{1/2}$

where A is any arbitrary origin and s^2 is called the **mean square deviation.**

Definition. Coefficient of variation of a frequency distribution is defined to be C.V = $\frac{\sigma}{\bar{x}}$ x 100

For comparing the variability of two sets of observations of a frequency distribution we calculate the C.V for each of the set of frequency distribution. The set having smaller C.V is said to be more consistent than the other.

Example1 . Consider the numbers 1, 2, 3, 4, 5, 5, 7

Their arithmetic mean $\bar{x}$=4

Now, $\sum(x_{i-}4)^2$=28.(verify)

$\therefore\sigma$=$\left[\frac{\sum(x_{i-}4)^2}{7}\right]^{1/2}$=$\left(\frac{28}{7}\right)^{1/2}$=2

Example2. For the frequency distribution , $\bar{x}$ =27.3

Hence we have the following table.

x_i	f_i	$x_i - 27.3$	$(x_{i-27.3})^2$	$f_i(x_{i-27.3})^2$
04.5	11	-22.8	519.84	5718.24
14.5	20	-12.8	163.84	3276.80
24.5	16	-2.8	7.84	125.44
34.5	36	7.2	51.84	1866.24
44.5	17	17.2	295.84	5029.28
Total	**100**	**-**	**-**	**16016**

$\therefore\sigma^2 = \frac{1}{N}\sum f_i\,(x_{i-}\bar{x})^2$=$\frac{16016}{100} = 160.16$

$\therefore\sigma$=12.66

We now establish a relation between the root mean square deviation s and standard deviation σ

Theorem3.1 $\sigma^2 = s^2 - d^2$ where d=$\bar{x}$- A

Proof. $s^2 = \frac{\sum f_i(x_{i-}A)^2}{N}$

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

=$\frac{1}{N}\,[\sum f_i\,(x_i-\bar{x})^2 + 2\sum f_i\,(x_i-\bar{x})(\bar{x}-A) +$

NOTES

$\sum f_i (\bar{x} - A)^2]$

$=\frac{\sum f_i(x_{i-}\bar{x})^2}{N}+\frac{2d}{n}\sum f_i (x_i - \bar{x})+d^2$

$= \boldsymbol{\sigma^2} + d^2$ (since$\sum f_i (x_i - \bar{x}) = 0$)

$\therefore \sigma^2 = s^2 - d^2$

Corollary. The standard deviation is the least possible root mean square deviation.

Proof. We have $s^2 = \boldsymbol{\sigma^2} + d^2$

$\therefore s^2$ is least when d = 0. Hence the least value of s^2 and σ^2.

The following theorem gives another formula for calculation of standard deviation of a frequency distribution.

Theorem3.2 $\boldsymbol{\sigma} = \left[\frac{\sum f_i x_i{}^2}{N} - \left(\frac{\sum f_i x_i}{N}\right)^2\right]^{1/2}$

Proof. $\sigma^2=(1/N)\sum f_i(x_{i-}\bar{x})^2$

$=(1/N)[\sum f_i(x_i{}^2 - 2x_i\bar{x} + \bar{x}^2)]$

$=\frac{\sum f_i x_i{}^2}{N} - \boldsymbol{2}\bar{x}\left(\frac{\sum f_i x_i}{N}\right)+\bar{x}^2\left(\frac{\sum f_i}{N}\right)$

$\therefore \sigma = \left[\frac{\sum f_i x_i{}^2}{N} - \left(\frac{\sum f_i x_i}{N}\right)^2\right]^{1/2}$

Theorem 3.3 The standard deviation σ is independent of change of origin and is dependent on change of scale.

Proof. We have $\sigma_x^2 = (1/N)\sum(x_{i-}\bar{x})^2$

Suppose we change the variable x_i and u_i where $u_i = x_i - A$,

A being an arbitrary origin.

We know that $\bar{u} = \bar{x} - A$

Now, $u_i - \bar{u} = x_i - A$

Now,

$\sigma_x^2 = (1/N)\sum f_i(x_{i-}\bar{x})^2 = (1/N)\sum f_i(u_{i-}\bar{u})^2$

$=\sigma_2^u$

Hence σ is independent of change of origin.

Now, suppose we change the variable x_i and v_i where $v_i = x_i/h$.

Then$\bar{v} = \bar{x}/h$

$\therefore v_i - \bar{v} = (1/h)(x_{i-}\bar{x})$

Now, σ_x^2=$(1/N)\sum f_i(x_{i-}\bar{x})^2 = (h^2/N)\sum f_i(v_{i-}\bar{v})^2$

$= h^2\sigma_v^2$

$\therefore$ S.D is dependent on change of scale.

NOTES

Note. When we effect a change in origin as well as in scale σ^2 is multiplied by the square of the scale introduced.

Hence $\sigma_x^2 = h^2\left[\frac{\sum f_i u_i{}^2}{N} - \left(\frac{\sum f_i u_i}{N}\right)^2\right]$

Theorem 3.4 (Variance of combined set). Let the mean and standard deviation of two sets containing n_1 and n_2 Members be $\bar{x}_1, \bar{x}_2$ and σ_1, σ_2 respectively.Suppose the two sets are grouped together as one set of $(n_1 + n_2)$ members. Let $\bar{x}$ be the mean and σ be the standard deviation of this set. Then $\sigma^2 = \frac{1}{n_1+n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$

Where $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

Proof. $\sigma^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1+n_2} f_i(x_{i-}\bar{x})^2\right]$

$= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1} f_i(x_{i-}\bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2} f_i(x_{i-}\bar{x})^2\right]$

Now, $\sum_{i=1}^{n_1} f_i(x_{i-}\bar{x})^2 = \sum_{i=1}^{n_1} f_i(x_i - \bar{x}_1 + \bar{x}_1 - \bar{x})^2 = \sum_{i=1}^{n_1} f_i[(x_i - \bar{x}_1) + d12$

$=\sum_{i=1}^{n_1} f_i\,(x_i - \bar{x}_1)$+2d$\sum_{i=1}^{n_1} f_i\,(x_i - \bar{x}_1) + d_1^2\sum_{i=1}^{n_1} f_i$

$=n_1\sigma_1^2 + n_1 d_1^2$ (since $\sum f_i\,(x_i - \bar{x}_1) = 0$)

Similarly $\sum_{i=n_1+1}^{n_1+n_2} f_i(x_{i-}\bar{x})^2 = n_2\sigma_2^2 + n_2 d_2^2$

Hence $\sigma^2 = \frac{1}{n_1+n_2}[(n_1\sigma_1^2 + n_1 d_1^2) + (n_2\sigma_2^2 + n_2 d_2^2)]$ ……..(1)

$= \frac{1}{n_1+n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$

Solved Problems.

Problem 1. Find (i) mean (ii) range (iii) S.D (iv) mean deviation about mean and (v) coefficient of variation for the following marks of 10 students.

20, 22, 27, 30, 40, 48, 45, 32, 31, 35

NOTES

Solution. (i) Mean = $1/n \sum x_i = \frac{330}{10} = 33$

(ii) Range = Maximum value - Minimum Value

=48-20=28

(iii) $\sigma = \left[\frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2\right]^{1/2}$

Here we have $\sum x_i^2 = 11652(\text{verify})$

$\therefore \sigma = \left[\frac{11652}{10} - \left(\frac{330}{10}\right)^2\right]^{1/2} = (76.2)^{1/2} = 8.73$

(iv) Mean deviation about mean $= \frac{1}{10}[\sum |x_i - 33]$

$= \frac{1}{10}[13 + 11 + 6 + 3 + 7 + 15 + 12 + 1 + 2 + 2]$

= 7.2

(v) C.V = $\left(\frac{\sigma}{\bar{x}}\right) \times 100 = \left(\frac{8.73}{33}\right) \times 100 = 26.45$

Problem 2. Show that the variance of the first n natural numbers is $\frac{1}{12}(n^2 - 1)$

Solution. $\sigma^2 = \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2$

We have $\sum x_i = 1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$and

$$\sum x_i^2 = 1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1).$$

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

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

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

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

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

Problem 3. The following table gives the monthly wages of workers in a factory. Compute (i) standard deviation (ii) quartile deviation and (iii) coefficient of variation.

NOTES

Monthly wages	No. of workers	Monthly wages	No. of workers
125-175	2	375-425	4
175-225	22	425-475	6
225-275	19	475-525	1
275-325	14	525-575	1
325-375	3	**Total**	**72**

Solution. Let A=300; h=50 and $u_i = \frac{1}{50}(x_i - 300)$. The table is

Mid x_i	f_i	u_i	f_iu_i	$f_iu_i^2$	c.f
150	2	-3	-6	18	2
200	22	-2	-44	88	24
250	19	-1	-19	19	43
300	14	0	0	0	57
350	3	1	3	3	60
400	4	2	8	16	64
450	6	3	18	54	70
500	1	4	4	16	71
550	1	5	5	25	72
Total	**72**	**-**	**-31**	**239**	**-**

NOTES

(i) $\bar{x} = A + h\bar{u}$

$$= 300+50\left(\frac{-31}{72}\right) = 300 - \frac{1550}{72} \quad =300-21.53=278.47$$

(ii) $Q_1 = 175 + \frac{(18-2)\times 50}{22}$

$$= 175 + \frac{800}{22} = 211.36$$

$$Q_3 = 275 + \frac{(54-43)\times 50}{22}$$

$$= 275 + \frac{550}{14} = 314.29$$

$\therefore$ Q.D $= \frac{1}{2}(Q_3 - Q_1)$

$$= \frac{1}{2}(314.29 - 211.36)$$

$$= 51.45$$

(iii) $\sigma^2 = h^2\left[\frac{\sum f_i u_i^2}{N} - \left(\frac{\sum f_i u_i}{N}\right)^2\right]$

$$=50^2\left[\frac{239}{72} - \left(\frac{31}{72}\right)^2\right]$$

σ = 88.52(verify).

(iv) C.V = $\frac{88.52}{278.47}$ x 100

=31.79

Problem 4. Find the arithmetic mean $\bar{x}$, standard deviation σ and percentage of case within $\bar{x} \pm \sigma$, $\bar{x} \pm 2\sigma$ and $\bar{x} \pm 3\sigma$ in the following frequency distribution.

Marks	10	9	8	7	6	5	4	3	2	1
Frequency	1	5	11	15	12	7	3	3	0	1

NOTES

Solution.

x_i	f_i	$f_i\ x_i$	$f_i\ x_i^2$
10	1	10	100
9	5	45	405
8	11	88	704
7	15	105	735
6	12	72	432
5	7	35	175
4	3	12	48
3	2	6	18
2	0	0	0
1	1	1	1
Total	**57**	**374**	**2618**

$$x\perp -= \frac{\sum f_i x_i}{N} = \frac{374}{57} = 6.56$$

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

$$= \frac{2618}{57} - \left(\frac{374}{57}\right)^2 = \frac{2618 \times 57 - 374^2}{57^2} = \frac{9350}{572}$$

$$\therefore \sigma = \left(\frac{1}{57}\right)\sqrt{9350} = 1.7 \text{ (approximately)}$$

Now, $\bar{x} \pm \sigma = 6.56 \pm 1.7 = 8.26, 4.86$

There are 45 items [7+12+15+11] which lie within 4.86 and 8.26

∴ Percentage of cases lying within the range $\bar{x} \pm \sigma = \frac{45}{57} \times 100 = 79\%$

Now $\bar{x} \pm 2\sigma = 6.56 \pm 3.4 = 9.6, 3.16$

There are only 53 items [3+7+12+15+11+5] which is within 3.16 and 9.96

$\therefore$ Percentage of items lying within the range $\bar{x} \pm 2\sigma$ $=\frac{53}{57} \times 100$=93%

Similarly the percentage of items lying within the range $\bar{x} \pm 3\sigma$

is 98% (verify).

Problem 5. Mean and standard deviation of the marks of two classes of sizes 25 and 75 are given below.

	Class A	**Class B**
Mean	80	85
S.D	15	20

calculate the combined mean and standard deviation of the marks of the students of the two classes. Which class is performing a consistent progress?

Solution. Let $\bar{x}$ and σ be the mean and standard deviation of the combined classes.

Given $\bar{x}_1 = 80; \bar{x}_2 = 85;\ \sigma_1 = 15; \sigma_2 = 20; n_1 = 25; n_2 = 75$

$$\therefore \bar{x} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_1} = \frac{25 \times 80+75\times 85}{100} = \frac{8375}{100} = 83.75$$

Now, $d_1 = \bar{x}_1 - \bar{x}$ =80-83.75=- 3.75

$d_2 = \bar{x}_2 - \bar{x}$ =85-83.75=1.25

We have, $\sigma^2 = \frac{1}{n_1+n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$

$$\therefore \sigma^2 = \frac{1}{100}[25 \times 15^2 \times 75 \times 20^2 + 25(-3.75)^2 + 75(1.25)^2]$$

$$=\frac{1}{100}[5625 + 30000 + 351.5625 + 117.1875]$$

=360.9375

$\therefore$ σ=19 (approximately).

C.V of marks of class A=$\frac{\sigma_1}{\bar{x}_1} \times 100 = \frac{15}{80} \times 100$

= 18.75

C.V of marks of class B=$\frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{20}{85} \times 100$ = 23.53

Since the C.V of marks of class A is smaller than that of class B , class A is performing consistent progress.

NOTES

Problem 6. Prove that for any discrete distribution standard deviation is not less than the mean deviation from mean.

Solution. Let m=mean deviation from mean.

$\therefore$ m=$(1/N)\left[\sum f_i |x_i - \bar{x}|\right]$

We have to prove σ not less than m.

(i.e) to prove that $\sigma^2 \geq m^2$

Now, $\sigma^2 \geq m^2 \Leftrightarrow (1/N)\sum f_i(x_i - \bar{x})^2 \geq \left[(1/N)\sum f_i |x_i - \bar{x}|\right]^2$

$\Leftrightarrow (1/N)\sum f_i z_i^2 \geq \ [(1/N)\sum f_i z_i\ ]^2$ where $z_i = |x_i - \bar{x}|$

$\Leftrightarrow (1/N)[\sum f_i z_i^2 - \ (\sum f_i z_i)^2] \Leftrightarrow \sigma_2^2 \geq 0$ which is true.

Hence the result.

Problem 7. The scores of two cricketers A and B in 10 innings are given below. Find who is a better run getter and who is more consistent player.

A scores x_i	40	25	19	80	38	8	67	121	66	76
B scores y_i	28	70	31	0	14	111	66	31	25	4

Solution. For cricketer A: $\bar{x} = \frac{540}{10} = 54$

For cricketer B: $\bar{y} = \frac{380}{10} = 38$

NOTES

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
40	-14	196	28	-10	100
25	-29	841	70	32	1024
19	-35	1225	31	-7	49
80	26	676	0	-38	1444
38	-16	256	14	-24	576
8	-46	2116	111	73	5329
67	13	169	66	28	784
121	67	4489	31	-7	49
66	12	144	25	-13	169
76	22	484	4	-34	1156
Total	**-**	**10596**	**Total**	**-**	**10680**

$\sigma_x = [(1/N)\sum(x_{i-}\bar{x})^2]^{1/2} = \left[\frac{10596}{10}\right]^{1/2} = \sqrt{1059.6} =32.55$

Similarly $\sigma_y = [(1/N)\sum(y_{i-}\bar{y})^2]^{1/2} = \left[\frac{10680}{10}\right]^{1/2} = \sqrt{1068}$

$= 32.68$

C.V of A = $\left(\frac{\sigma_x}{\bar{x}}\right) \times 100 = \frac{32.35}{54} \times 100 = 60.28$

C.V of B = $\left(\frac{\sigma_y}{\bar{y}}\right) \times 100 = \frac{32.68}{38} \times 100 = 86$

Since $\bar{x} > \bar{y}$ cricketer A is better run getter. C.V of A< C.V of B, Cricketer A is also a consistent player.

Problem 8. The mean and standard deviation of 200 items are found to be 60 and 20. If at the time of calculation two items are wrongly taken as 3 and 67 instead of 13 and 17, Find the correct mean and standard deviation.

Solution. Here n=200; $\bar{x} = 60;\ \sigma = 20$

$$\bar{x} = 60 => \frac{\sum x_i}{200} = 60$$

$$\therefore \sum x_i = 12000$$

Corrected $\sum x_i = 12000 - (3 + 67) + (13 + 17) = 11960$

$\therefore$Corrected $\bar{x} = \frac{11960}{200} = 59.8$

$$\sigma^2 = \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2. \text{ Hence } 20^{2=} \frac{\sum x_i^2}{200} - (60)^2$$

$\therefore \sum x_i^2 = 200(20^2 + 60^2) = 800000$

After correction $\sum x_i^2 = 800000 - (3^2 + 67^2) + (13^2 + 17^2) = 795960$

$\therefore$ Corrected $\sigma^2 = \frac{795960}{200} - (59.8)^2$

$\therefore \sigma$ =20.09

NOTES

Problem9. Find (i) the mean deviation from the mean (ii) variance of the arithmetic progression a, a+d, a+2d, ..., a+2nd.

Solution. There are 2n +1 terms in the A.P

$$\therefore \bar{x} = \frac{1}{2n+1}[a + (a + d) + \cdots + (a + 2nd)]$$

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

= a+nd

(i) Mean deviation from mean = $\frac{1}{2n+1}\sum |x_i - \bar{x}|$

$$= \frac{1}{2n+1}[2d(1 + 2 + \cdots + n)]$$

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

(iii) Variance $\sigma^2 = \frac{1}{2n+1}\sum(x_i - \bar{x})^2$

$$= \frac{1}{2n+1}[2d^2(1^2 + 2^2 + \cdots + n^2)]$$

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

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

Exercises.

1. Calculate mean, S.D and C.V of the marks obtained by 20 students in an examination.

62	85	73	81	74	58	66	72	54	84
65	50	83	62	85	52	80	86	71	75

2.Calculate the standard deviation from the following data of income of 10 employees of a firm.

100	120	140	120	180	175	185	130	200	150

3. Prepare a frequency table from the following passage taking consonants and vowels in each word as two variable x and y. Find $\bar{x}, \bar{y}, \sigma_x$ andσ_y

4. Calculate the mean deviation from (i) mean (ii) median (iii) mode for the following data.

Size of item	Frequency	Size of item	Frequency
3-4	3	7-8	85
4-5	7	8-9	32
5-6	22	9-10	8
6-7	60	**Total**	**217**

5.Find the standard deviation of the following heights of 100 male students.

Height of inches	60-62	63-65	66-68	69-71	72-74
No. of students	5	18	42	27	8

UNIT-IV MOMENTS SKEWNESS AND KURTOSIS

NOTES

4.1 INTRODUCTION

In previous chapters we have introduced certain measures of central tendencies and measures of dispersion with the aim of finding a " few statistical constants" that represent the entire data. In this chapter we introduce some more statistical constants known as moments.

4.2.MOMENTS

Definition. The rth moment about any point A, denoted by μ_r of a frequency distribution (f_i/x_i) is defined by $\mu_r = \frac{\sum f_i(x_i - A)^r}{N}$

When A = 0 We get $\mu_r == \frac{\sum f_i(x_i - \bar{x})^r}{N}$ which is the rth moment about the origin.

The rth moment about the arithmetic mean $\bar{x}$ of a frequency distribution is given by $\mu_r = \frac{\sum f_i(x_i - \bar{X})^r}{N}$

μ_r is also called the rth central moment.

Note 1. The first moment about origin coincides with the A.M of the frequency distribution and μ_2 is nothing but the variance of the frequency distribution.

Note 2. $\mu_1 = \frac{\sum f_i (x_i - \bar{x})}{N}$ 0 ;

Note 3. $\mu_1 = \frac{\sum f_i(x_i - A)}{N} = \left[\frac{(\sum f_i x_i) - A \sum f_i}{N}\right] = \bar{x} - A$

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

We now establish a relation between μ_r' and μ_r

Theorem 4.1

$\mu_r = \mu'_r - r_{c_1} \mu'_{r-1} \mu'_1 + r_{c_2} \mu_{r-1}(\mu'_1)^2 .... +(-1)^{r-1} (r-1)(\mu'_1)^r$

Proof. $\mu_r = 1/N \sum f_i (x_i - \bar{x})^r$

$= 1/N \sum f_i (x_i - A + A - \bar{x})^r$

$= 1/N \sum f_i \ (x_i \ - A - d)^r$ where d=x̄-A

$= 1/N[\sum f_i \ (x_i \ - A)^r$-$r_{c_1} d \sum f_i \ + r_{c_2} d^2 \sum f_i \ (x_i \ - \ A)^{r-2} \ -$ ...+$_{rcr-1-dr-1}$fi xi $- Ar-1\mu1$ *+rcr(−d)r*fi]

$= \mu'_{r}$ - $r_{c_1} d\mu'_{r-1} \ + r_{c_2} \ d^2\mu'_{r-2}$.-......... +(-1) $^{r-1}$ $rd^{(r-1)}(\mu'_1)$ $+(-1)^r d^r$

= μ'_{r}- r_{c_1} μ'_{r-1} μ'_1 + r_{c_2} $\mu_{r-1}(\mu'_{1})^{2}$ .... +(-1) $^{r-1}$ (r-1)(μ'_1) ^{r}

Note. Putting r = 2, 3, 4 in the above theorem we have

(i) $\mu_2 = \mu'_2 + (\mu_1{}')^2$
(ii) $\mu_3 = \mu'_3 - 3\mu'_2\mu_1{}' + 2(\mu_1{}')^3$
(iii) $\mu_4 = \mu'_4 - 4\mu'_3\mu_1{}' + 6\mu'_2(\mu_1{}')^2 - 3(\mu_1{}')^4$

Theorem 4.2 $\mu'_r = \mu_r + r_{c_1} \ \mu_{r-1} \ \mu'_1 + r_{c_2} \ \mu_{r-2}(\mu'_{1})^{2}$ .... +(μ'_1) ^{r}

Proof. $\mu'_r \ = 1/N \sum f_i \ (x_i \ - A\ )^r$

$= 1/N \sum f_i \ (x_i \ - \bar{x} + \bar{x} - A\ )^r$

$=1/N \sum f_i \ (x_i \ - \bar{x} + d\ )^r$ where $d = \bar{x} - A = \mu_1$

$=1/N \sum f_i \ [(x_i \ - \ \bar{x})^r$+$r_{c_1}(x_i \ - \ \bar{x})^{r-1}d$+$r_{c_2}(x_i \ - \ \bar{x})^{r-2}d^2 + \cdots + d^n]$

$= \mu_r + r_{c_1} \ \mu_{r-1} \ \mu'_1 + r_{c_2} \ \mu_{r-2}(\mu'_{1})^{2}$ .... +(μ'_1) ^{r}

Note. Putting r =2, 3, 4 in the above theorem and using $\mu_1 \ = 0$ we have.

(i) $\mu'_2 = \mu_2 + (\mu_1{}')^2$
(ii) $\mu'_3 = \mu_3 + 3\mu_2\mu_1{}' + (\mu_1{}')^3$
(iii) $\mu'_4 = \mu_4 + 4\mu_3\mu_1{}' + 6\mu_2(\mu_1{}')^2 + (\mu_1{}')^4$

Note. When the variable x_i are changed into another variable u_i where $u_i \ = \frac{x_i - A}{h}$ the r^{th} moment μ_r of the variable x_i is given by
$\mu_r = \ h^r \left[\frac{\sum f_i (u_i - \bar{u})}{N}\right]$
Thus the r^{th} moment μ_r of the variable x_i is h^r times the r^{th} moment of the variable u_i

Definition. Karl Pearson's $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ coefficients are defined as follows.

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

$$\gamma_1 = \sqrt{\beta_1} \text{ and } \gamma_2 = \beta_2 - 3$$

The above four coefficients depends upon the first four central moments. They are pure numbers independent of units in which the variable x_i is expresses. Also their values are not affected by change of origin and scale. These constants are used in section 4.3 in the study of skewness and kurtosis.

4.3 SKEWNESS AND KURTOSIS

If the values of a variable x_i are distributed symmetrically about the mean which is taken as the origin then for every positive value of x -$\bar{x}$ there corresponds a negative equal value. Hence when these values are clubed they retain their signs and cancel on addition.

$$\therefore \mu_3 = \frac{1}{N} \sum f_i \ (x_i - \bar{x})^3 = 0 \text{ . Hence } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

Thus in the case of symmetrical distribution fails to be symmetrical (asymmetrical) then we say that it is skewed distribution., Thus skewness means lack of symmetry. From the above discussion we see that β_1 can be taken as a measure of skewness. We say that a frequency distribution has positive skewness if β_1>0 and negative skewness if β_1< 0

For a symmetric distribution the mean ,median and mode coincide. Hence for an asymmetrical distribution the distance between the median and mean may be used as measures of skewness.

$\therefore$ Mean -Mode and Mean - Median may be taken as measures of skewness.

To make these measures free from units of measurements so that comparison with other distribution may be possible we divide them by a suitable measure of dispersion and obtain the following coefficients of skewness.

(i) Karl Person's coefficient of skewness.

$\frac{\text{Mean}-\text{Mode}}{\sigma}$ - $\frac{3(\text{Mean}-\text{Mode})}{\sigma}$ are called Karl Pearson's coefficients of skewness.

(ii)Bowley's coefficient of skewness is given by $\frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$

Kurtosis

Definition. Kurtosis is the degree of peakedness of a distribution usually taken relative to a normal distribution. It is measured by the coefficient β_2

For a normal curve β_2 = 3 or (γ_2 = 0) **messokurtic.**

For a curve which is flater than the normal curve $\beta_2 < 3$ or ($\gamma_2 < 0$) and such a curve is known as platykurtic.

For a curve which is more peaked than the normal curve $\beta_2 > 3$ or ($\gamma_2 > 0$) and such a cure is known as leptokurtic.

Solved problems.

Problem 1.Calculate the first four central moments from the following data to find β_1 and β_2 and discuss the nature of the distribution.

X	0	1	2	3	4	5	6
y	5	15	17	25	19	14	5

Solution.

Here $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{300}{100}$ =3

Choosing $\mu_i = x_i - \bar{x} = x_i - 3$ we have the following table.

x_i	f_i	u_i	$f_i\, u_i$	$f_i u_i^2$	$f_i u_i^3$	$f_i u_i^4$
0	5	-3	-15	45	-135	405
1	15	-2	-30	60	-120	240
2	17	-1	-17	17	-17	17
3	25	0	0	0	0	0
4	19	1	19	19	19	19
5	14	2	28	56	112	224
6	5	3	15	45	135	405
Total	**100**	**-**	**0**	**242**	**-6**	**1310**

$$\mu_1 = 1/N \sum f_i\ (x_i\ -\ \bar{x})\ \ = 0$$

$$\mu_2 =\ 1/N \sum f_i\ (x_i\ -\ \bar{x})^{\,2}\ \ = \frac{242}{100} = 2.42$$

$$\mu_3 = 1/N \sum f_i\ (x_i\ -\ \bar{x})^3 = -\frac{6}{100} = -0.06$$

$$\mu_4 = 1/N \sum f_i\ (x_i\ -\ \bar{x})^4 = \frac{1310}{100} = 13.10$$

NOTES

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-0.06)^2}{2.42^3}\ =\ \frac{.0036}{14.1725} = 0.0003$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{13.10}{2.42^2} = \frac{13.10}{5.8564} = 2.237$$

Since $\beta_1 > 0$ the distribution is positively skewed.
Since $\beta_2 = 2.237 < 3$ the distribution is platykurtic.

Problem 2. Calculate the values of β_1 and β_2 for the distribution.

Solution. Taking $u_i = \frac{x_i-24.5}{10}$ we get the following table.

x_i	f_i	u_i	$f_i\ u_i$	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$
04.5	11	-2	-22	44	-88	176
14.5	20	-1	-20	20	-20	20
24.5	16	0	0	2	0	0
34.5	36	1	36	36	36	36
44.5	17	2	24	68	136	272
Total	**100**	**0**	**28**	**168**	**64**	**504**

Here we have chosen A=24.5 and h=10

$$\mu_1' \ = 1/N \sum f_i\ (x_i\ \ - A\ )\ = \frac{1}{N}\sum f_i\ u_i \times h = \frac{28}{100} \times 10 = 2.8$$

$$\mu_2' \ = \frac{1}{N}\sum f_i\ {u_i}^2 \times h^2 = \frac{168}{100} \times 10^2 = 168$$

$$\mu_3' \ = \frac{1}{N}\sum f_i\ {u_i}^3 \times h^3 = \frac{64}{100} \times 10^3 = 640$$

$$\mu_4' \ = \frac{1}{N}\sum f_i\ {u_i}^4 \times h^4 = \frac{504}{100} \times 10^4 = 50400$$

Now, $\mu_1 = 0$

$\mu_2 = \mu_2' + (\mu_1')^2 = 168 - (2.8)^2$=160.1

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 640 - 3$ x 168 x 2.8+$2(2.8)^3$

= -727.296

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

= 50400-4 ×640×2.8+6×168×$(2.8)^2$-$3(2.8)^4$

= 50950.323

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ =0.129(verify)

$\beta_2 = \frac{\mu_4}{\mu_2^2}$ = 1.986(verify)

Problem 3. The first four moments of a distribution about x=2 are 1, 2.5, 5.5 and 16. Calculate the four moments (i) about the mean. (ii) about the zero.

Solution. Given $\mu_1 = 1; \mu_2 = 2.5; \mu_3 = 5.5;\ \mu_4 = 16$ where A=2

(i) Moments about mean**.**

$\mu_{\mathbf{1}} = 0$

$\mu_2 = \mu_2' + (\mu_1')^2 = 2.5 - 1 = 1.5$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$=5.5 - 3×2.5+2=0

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$
=16-4×5.5+6×2.5-3=6.

(ii) Moments about Zero.

We have $\bar{x} = A + \mu_1$ (refer Note 3 in 4.1)
= 2+1=3

Now the first moment about zero μ_1' =$1/\mathrm{N} \sum \mathrm{f_i}\ (\mathrm{x_i}\ - 0\ )$
= $\bar{x}$ =3

Now, $\mu_2' = \mu_2 + (\mu_1')^2$= 1.5+$3^2 = 10$

$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$=0+3×1.5+×3+$3^3 = 40.5$

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

= 6+(4×0×3)+$(6 \times 15 \times 3^2) + 3^4 = 168$

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

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

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

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

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

$$\mu_2 = \mu_2' - (\mu_1{}')^2 = \frac{1}{6}(n+1)(2n+1) - \left[\frac{1}{2}(n+1)\right]^2$$

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

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

$\mu_2 \neq 0$ if $n \neq \pm 1$

$\therefore$ When n>1, $\beta_1 = 0$

Hence the distribution is symmetric.

Problem 5. For a frequency distribution (f_i/x_i) show that $\beta_2 \geq 1$

Solution. We have $\beta_2 = \frac{\mu_4}{\mu_2^2}$

To prove $\beta_2 \geq 1$ it is enough to prove that $\mu_4 \geq \mu_2^2$

$$\text{Now , } \mu_4 - \mu_2^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} - \left[\frac{\sum f_i (x_i - \bar{x})^2}{N}\right]^2$$

$$= \frac{\sum f_i z_i{}^2}{N} - \left(\frac{\sum f_i z_i}{N}\right)^2 \quad \text{where } z_i = (x_i - \bar{x})^2$$

$$= \sigma_z^2 \geq 0$$

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

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

Hence $\beta_2 \geq 1$

Exercises.

1. For the following data calculate the Karl Pearson's co efficient of skewness.

(i)

Wages in Rs.	10	11	12	13	14	15
Frequency	2	4	10	8	5	1

(ii)

Size	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

NOTES

2. Find Karl Pearson's coefficient of skewness for the following data.

(i)

Age	**Students**	**Age**	**Students**
10-12	4	18-20	20
12-14	10	20-22	14
14-16	16	22-24	6
16-18	30	Total	100

(ii)

Wage	**No. of Workers**	**Wage**	**No. of Workers**
Above Rs.5	120	Above Rs.55	58
Above Rs.15	105	Above Rs.65	42
Above Rs.25	96	Above Rs.75	12
Above Rs.35	85	Above Rs.85	0
Above Rs.45	72	Total	590

3. Karl Pearson's coefficient of skewness of a distribution is 0.4, its S.D is 8 and mean 30. Find the mode and median.

4. Calculate the first four moments of the following distribution about the mean. Find β_1 and β_2 and hence comment on the nature of the distribution.

X	0	1	2	3	4	5	6	7	8
Y	1	8	28	56	70	56	28	8	1

UNIT-V CURVE CUTTING

5.1 Introduction

NOTES

So far we have introduced several statistical constant like measures of central tendencies, measures of dispersion and measures of skewness and kurtosis in order to characterize a given set of sample data drawn from a population. Another important and useful method employed to understand the parent population is to discover a functional relationship between the variable comprising the sample data.

Let x_i where i = 1,2,3,......n be the values of the dependent variables y_i . If the points (x_i, y_{i}) ; i=1,2,.....n are plotted on a graph paper and we obtain a diagram called scatter diagram . Hence if there is a functional relationship between x_i *and* y_i . The process of finding such a functional relationship between the variables is called curve fitting. Curve fitting is useful in the study of correlation and regression which will be dealt with in the next chapter. For example the lines of regression can be got by fitting a linear curve to a given bivariate distribution. The properties of the curve fitted to a given data can be used to know the properties of the parent population.

UNIT-VI PRINCIPLE OF LEAST SQUARES

NOTES

6.1 INTRODUCTION

Among the many methods available for curve fitting the most popular method is the principle of least squares. Let (x_i, y_i) where i = 1,2,.....n be the observed set of the variables (x, y) . Let y = f(x) be a functional relationship sought between the variables x and y .

Then $d_i = y_i - f(x_i)$ which is the difference between the observed value of y and the determined by the functional relation is called the residuals. The principle of least squares states that the parameters involved in f(x) should be chosen in such a way that $\sum d_i^2$ is minimum.

6.2 Fitting a straight line

Consider the fitting of the straight line y = ax + b to the data (x_i, y_i) , i=1,2,....,n

The residual d_i is given by $d_i = y_i - (ax_i + b)$

$\therefore \sum d_i^2 = \sum (y_i - ax_i - b)^2 = R$(say). According to the principle of least squares we have to determine the parameters a and b so that R is minimum.

$$\frac{\partial R}{\partial a} = 0 \Rightarrow -2\sum (y_i - ax_i - b)\, x_i = 0$$

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

$$\therefore a\sum x_i^2 + b\sum x_i = \sum x_i\, y_i \quad \text{............. (1)}$$

$$\frac{\partial R}{\partial a} = 0 \Rightarrow -2\sum (y_i - ax_i - b) = 0$$

$$\therefore a\sum x_i + nb = \sum y_i \quad \text{............ (2)}$$

Equations (1) and (2) are called normal equations from which a and b can be found.

6.3 Fitting a second degree parabola.

Consider the fitting of the parabola $y = ax^2 + bx + c$ to the data (x_i, y_i) where i=1,2,......n.

The residual d_i is given by $d_i = y_i - (ax_i^2 + bx_i + c)$

$$\therefore \sum d_i^2 = \sum (y\text{i} - ax_i^2 - bx_i - c)^2 = \text{R(say)}$$

By the principle of least squares we have to determine the parameters a,b and c so that R is minimum.

NOTES

$$\frac{\partial R}{\partial a} = 0 \Rightarrow -2\sum(y_i - ax_i^2 - bx_i - c)\, x_i^2 = 0$$

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

$$\therefore a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum x_i^2\, y_i \quad ....(1)$$

$$\frac{\partial R}{\partial a} = 0 \Rightarrow -2\sum(y_i - ax_i^2 - bx_i - c)\, x_i = 0$$

$$\Rightarrow \sum x_i y\text{i} - a\sum x_i^3 - b\sum x_i^2 - c\sum x_i = 0$$

$$\Rightarrow a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i y\text{i} \quad ....(2)$$

$$\frac{\partial R}{\partial a} = 0 \Rightarrow -2\sum(y_i - ax_i^2 - bx_i - c)\, x_i = 0$$

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

$$\therefore a\sum x_i^2 + b\sum x_i + nc = \sum y_i \quad ....(3)$$

Equations (1), (2), and (3) are called normal equations from which a,b and c can be found.

Note. If the given data is not in linear form it can be brought to linear form by some suitable transformations of variable. Then using the principle of least squares the curve of best fit can be achieved.

Curves of the form (i) $y = bx^a$ (ii) $y = ab^x$ (iii) $y = ae^{bx}$ are of special interest which are dealt with here in solved problems.

Solved Problems

Problem1. Fit a straight line to the following data.

X	0	1	2	3	4
Y	2.1	3.5	5.4	7.3	8.2

Solution. Let the straight line to be fitted to the data be y =ax + b

Then the parameters a and b are got from the normal equations.

$$\sum x_i = a\sum x_i + nb$$

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

NOTES

x_i	y_i	$x_i y_i$	x_i^2
0	2.1	0	0
1	3.5	3.5	1
2	5.4	10.8	4
3	7.3	21.9	9
4	8.2	32.8	16
Total	**26.5**	**69.0**	**30**

Hence the normal equations are

$$10a + 5b = 26.5 \quad ....(1)$$

$$30a + 10b = 69 \quad ....(2)$$

Solving (1) and (2) we get a=1.6 and b=2.1

∴ The straight line fitted for the is y = 1.6x + 2.1

Problem 2. Fit a straight line to the following data and estimate the value of y corresponding to x= 6

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

Solution.

Take $u_i = \frac{1}{5}(x_i - 15)$ and $v_i = y_i - 22$

Let v = au + b be the straight line to be fitted.

We get the following normal equations to get the parameters a and b. Then the normal equations are.

$$\sum v_i = a \sum u_i + nb$$

$$\sum u_i v_i = a\sum u_i^2 + b\sum u_i$$

x_i	y_i	u_i	v_i	$u_i\ v_i$	u_i^2
0	12	-3	-10	30	9
5	15	-2	-7	14	4
10	17	-1	-5	5	1
15	22	0	0	0	0
20	24	1	2	2	1
25	30	2	8	16	4
Total	**-**	**-3**	**-12**	**67**	**19**

∴ The normal equations are

-3a + 6b=-12 ....(1)

19a – 3b=67 ....(2)

Solving for a and b we get a=3.49 and b = -0.26

∴ The straight line to be fitted becomes y – 22 =3.49$\left(\frac{x-15}{5}\right)$ - 0.26

∴ 5y -110= 3.49x – 52.35 -1.30

∴5y= 3.49x + 56.35

∴y = .698x + 11.27

Now for x = 6 the estimated value of y is y=.698 × 6 +11.27 = 15.458

Problem 3. Fit a second degree parabola by taking x_i as the independent variable.

X	0	1	2	3	4
Y	1	5	10	22	38

Solution

Let the second parabola to be fitted to the data be y

Y = ax^2 +bx +c Then we have the normal equations to find a,b,c.

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

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

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

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

Now , the normal equations become

$354a + 100b + 30c = 851$ .... (1)

$100a + 30b + 10c = 243$ ....(2)

$30a + 10b + 5c = 76$ ....(3)

Solving for a,b and c we get a = 2.21 ; b = 0.26 and c = 1.42(verify)

∴ The second degree parabola is $y = 2.21 x^2 + 0.26 x + 1.42$

Problem 4. Fit the curve $y = bx^a$ to the following data.

X	1	2	3	4	5	6
Y	1200	900	600	200	110	50

Solution. $y = bx^a$

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

Let log y = Y and log x = X

Then the curve is transformed into Y = AX + B where A = a and B = log b. Hench the normal equations now become

$\sum Y = A \sum X + nB$

$\sum XY = A \sum x^2 + B \sum X$

X	Y	X	Y	XY	X^2
1	1200	0	3.0792	0	0
2	900	0.3010	2.9542	0.889	0.091
3	600	0.4771	2.7782	1.325	0.228
4	200	0.6021	2.3010	1.385	0.363
5	110	0.6990	2.0414	1.427	0.489
6	50	0.7782	1.6990	1.322	0.606
Total	**-**	**2.8574**	**14.8530**	**6.348**	**1.777**

NOTES

∴ **T**he normal equations are

2.9 A + 6 B = 14.9 approximately

1.8 A + 2.9 B = 6.6 approximately

∴ A = - 2.3 and b = 3.6 (verify)

∴ A = a = - 2.3 and B = log b = 3.6

∴ a = -2.3 and b = antilog 3.6 = 3981

∴ The required equation to the curve is $y = 3981\ x^{-2.3}$

Problem 5. Explain the method of fitting the curve of good fit $y = ae^{bx}$ (a>0)

Solution. $y = ae^{bx}$ ....(1)

∴ log y = log a + bx log e .... (2)

Let Y = log y; B = log a ; A = b log e

∴ (2) between y = Ax + B

This is linear equations in x and y whose normal equations are,

$\sum x_i y_i = A \sum x_i^2 + B \sum x_i$

$\sum y_i = A \sum x_i + nB$

From the two normal equations we can get the values of A and B and consequently a and b be obtained form a = antilog (B) and b $\frac{A}{\log e}$. Thus the curve of best fit (1) can be obtained.

Problem 6. Explain the method of fitting the curve $y = Ka^{bx}$ (a,k>0)

Obtaining the normal equations by the method of least squares.

Solution. The curve can be transferred to the form of a straight line as follows.

Log y = log k + b (log a) x ; (a,k > 0)

Let log y = Y; log k = B ; b log a = A

Hence the above equations takes the form Y = Ax + B

By the principle of least squares the normal equations to find A and B of the above straight line are

$$\sum x_i y_i = A \sum x_i^2 + B \sum x_i$$

$$\sum y_i = A \sum x_i + nB$$

After finding the values of A and B from the normal equations we can obtain the value of k,a and b hence the curve $y = k\,a^{bx}$ to the following data.

Problem 7 . Fit a curve of the form $y = ab^x$ to the following data.

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

Solution. $y = ab^x$ ....(1)

$\therefore$ log y = log a + log b ....(2)

Let log y = Y ; log a = B and log b = A

$\therefore$ (2) becomes Y = AX + B ....(3) where X = x – 1954.

x	Y	X = x – 1954	Y = log y	XY	X^2
1951	201	-3	2.3032	-6.9096	9
1952	263	-2	2.4200	-4.8400	4
1953	314	-1	2.4969	-2.4969	1
1954	395	0	2.5966	0	0
1955	427	1	2.6304	2.6304	1
1956	504	2	2.7024	5.4048	4
1957	612	3	2.7868	8.3604	9
Total		**0**	**17.9363**	**2.1491**	**28**

The normal equations for (3) are

$$\sum XY = A \sum x^2 + B \sum X$$

$$\sum Y = A \sum X + nB$$

$$28A = 2.1491 \quad .... (4)$$

$$7B = 17.9363 \quad ....(5)$$

Solving the above equation we get A = 0.0768 B = 2.5623

$\therefore$ b = antilog A = antilog 0.0768 = 1.19 (approximately)

A = antilog B = 2.5623 = 365.01 (approximately)

$\therefore$ The curve of good fit is $y = (365.01)(1.19)^X$

$= (365.01)(1.19)^{X-1954}$

Exercises

1.Fit a straight line to the following data regarding x as the independent variable.

(i)

X	0	1	2	3	4
Y	1	1.8	3.3	4.5	6.3

NOTES

(ii)

Year x	1911	1921	1931	1941	1951
Production in tons y	10	12	8	10	14

Also estimate the production in 1936.

2.Fit a second degree parabola to the following data taking x as the independent variable.

(i)

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	2.3

(ii)

X	1	2	3	4	5	6	7	8	9
Y	2	6	7	8	10	11	11	10	9

3.Fit a curve $y = ax^b$ for the following data.

X	1	2	3	4	5	6
Y	14	27	40	55	68	300

4.Fit a curve $y = ax^b$ for the following data.

X	1	2	3	4
Y	2.99	4.25	5.22	6.10

5.Fit the exponential curve $y = ae^{bx}$ to the following data.

X	0	2	4
Y	50.2	10	31.62

UNIT-VII CORRELATION

7.1 INTRODUCTION

NOTES

In statistical we have studied the methods of classifying and analysing data relative to single variable. However data presenting two sets of related observations may arise in many fields of activities giving n pairs of corresponding observations (x_i, y_i);i=1,2,..., n

For example, (i) x_i may represent height and y_i weight of a colletion of students. (ii) x_i may represent price of a commodity and y_i the corresponding demand. Such a data (x_i, y_i); i=1,2,..., n is called a bivariate data.

7.2 CORRELATION

Definition. Consider a set if bivariate data (x_i, y_i);i=1,2,..., n. If there is a change in one variable corresponding to change in the other variable we say that the variables are **correlated.**

If the two variables deviate in the same direction the correlation is said to be direct or positive. If they always deviate in the opposite direction the correlation is said to be inverse or negative. If the change in one variable corresponds to a proportional change in the other variable then the correlation is said to be **Perfect.**

Height and weight of a batch of students; Income and expenditure of a family are examples of variables with positive correlation.

Price and demand; volume ν and pressure ρ of a perfect gas which obeys the law ρν=k where k is a constant, are examples of variables with negative correlation.

Definition. Karl Pereson's coefficient or correlation between the variables x and y is defined by $\gamma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$ where $\bar{x}, \bar{y}$ are the arithmetic means and σ_x, σ_y the standard deviations of the variables x and y respectively.

Definition. The Covariance between x and y is defined by cov(x,y)= $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$ Hence $\gamma_{xy} = \frac{\text{cov(x,y)}}{\sigma_x\sigma_y}$

Example. The heights and weights of five students are given below.

Height in c.m. x	160	161	162	163	164
Weight in kgs. y	50	53	54	56	57

Hence $\bar{x} = 162; \bar{y} = 54; \sigma_x = \sqrt{2}$ and $\sigma_y = \sqrt{6}$ (verify)

Now

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (-2)(-4) + (-1)(-1) + 0 + (1 \times 2) + (2 \times 3)$$

$$= 17$$

$$\therefore \gamma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

$$= \frac{17}{5\sqrt{2}\sqrt{6}} = \frac{17\times\sqrt{12}}{60} = \frac{17\times 3.46}{60} = 0.98$$

Theorem 7.1. $\gamma_{xy} = \frac{n\sum x_i y_i - \sum x_i y_i}{\left[n\sum x_i^2 - (\sum x_i)^2\right]^{1/2}\left[n\sum y_i^2 - (\sum y_i)^2\right]^{1/2}}$

Proof. $\gamma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$ ....(1)

Now, $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}$

$$= \bar{x}(n\bar{y}) - \bar{y}(n\bar{x}) + n\bar{x}\bar{y}$$

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

$$= \sum x_i y_i - \left(\frac{1}{n}\right)\sum x_i \sum y_i$$

$$= \left(\frac{1}{n}\right)\left[n\sum x_i y_i - \sum x_i \sum y_i\right] \quad \text{....................(2)}$$

Also, $\sigma_2^x = \frac{1}{n}\sum(x_i - \bar{x})^2$

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

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

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

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

$$\therefore \sigma_x = \frac{1}{n}[n\sum x_i^2 - (\sum x_i)^2]^{1/2} \quad ....(3)$$

$$\text{Similarly, } \sigma_y = \frac{1}{n}[n\sum y_i^2 - (\sum y_i)^2]^{1/2} \quad ....(4)$$

NOTES

Substituting (2),(3) and (4) in (1) we get the required result.

The calculation of γ_{xy} may frequently be simplified by making use of the following theorem.

Theorem 7.2 The correlation coefficient is independent of the change of origin and scale.

Proof. Let $u_i = \frac{x_i - A}{h}$ and $\nu_i = \frac{y_i - B}{k}$ where h, k > 0.

$\therefore x_i = A + hu_i$ and $y_i = B + k\nu_i$.

Hence $\bar{x} = A + h\bar{u}$ and $\bar{y} = B + k\bar{\nu}$

$\therefore x_i - \bar{x} = h(u_i - \bar{u})$ and $y_i - \bar{y} = k(\nu_i - \bar{\nu})$

Also $\sigma_x = h\sigma_u$ and $\sigma_y = k\sigma_\nu$

$$\therefore \gamma_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y} = \frac{hk\sum(u_i-\bar{u})(\nu_i-\bar{\nu})}{n(h\sigma_u)(k\sigma_\nu)}$$

$$= \frac{\sum(u_i-\bar{u})(\nu_i-\bar{\nu})}{n\,\sigma_u\,\sigma_\nu}$$

$$= \gamma_{u\nu}$$

Hence $\gamma_{xy} = \gamma_{u\nu}$

Theorem 7.3 -1 ≤ γ ≤ 1

Proof. $\gamma_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y}$

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

Let $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$

$$\gamma_{xy}^2 = \frac{(\sum a_i b_i)^2}{(\sum {a_i}^2)(\sum {b_i}^2)}$$

By Schwartz inequality we have $(\sum a_i b_i)^2 \le \left(\sum {a_i}^2\right)\left(\sum {b_i}^2\right)$

Hence $\gamma^2_{xy} \le 1$

$\therefore |\gamma^2_{xy}| \le 1$

$-1 \le \gamma \le 1$

Note1. If $\gamma = 1$ the correlation is perfect and positive

Note 2. If $\gamma = -1$ the correlation is perfect and negative

Note 3. If $\gamma = 0$ the variables are uncorrelated.

Note 4. If the variables x and y are uncorrelated then Cov(x,y)= 0.

The following theorem gives another formula for γ_{xy} interms γ_x and γ_y.

Theorem 7.4 $\gamma_{xy} = \frac{\sigma_x^2+\sigma_y^2-(\sigma_{x-y})^2}{2\,\sigma_x\sigma_y}$

Proof. $(\sigma_{x-y})^2 = \frac{\sum[(x_i-y_i)-(\bar{x}-\bar{y})]^2}{n}$

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

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

$$= \sigma_x^2 - 2\gamma_{xy}\sigma_x\sigma_y + \sigma_y^2$$

$$\therefore \gamma_{xy} = \frac{\sigma_x^2+\sigma_y^2-(\sigma_{x-y})^2}{2\,\sigma_x\sigma_y}$$

Solved Problems.

Problem 1. Ten students obtained the following percentage of marks in the college internal test (x) and in the final university examination(y). Find the correlation coefficient between the marks of the two tests.

x	51	63	63	49	50	60	65	63	46	50
y	49	72	75	50	48	60	70	48	60	56

Solution. Choosing the origin A= 63 for the variable x and B= 60 for y and taking $u_i = x_i - A$ and $v_i = y_i - B$ we have the following table.

x_i	u_i	y_i	v_i	${u_i}^2$	${v_i}^2$	$u_i v_i$
51	-12	49	-11	144	121	132
63	0	72	12	0	144	0
63	0	75	15	0	225	0
49	-14	50	-10	196	100	140
50	-13	48	-12	169	144	156
60	-3	60	0	9	0	0
65	2	70	10	4	100	20
63	0	48	-12	0	144	0
46	-17	60	0	289	0	0
50	-13	56	-4	169	16	52
Total	**-70**	**-**	**-12**	**980**	**994**	**200**

$\gamma_{xy} = \gamma_{uv}$ (by theorem 7.2)

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

$$= \frac{10\times 500 - (-70)\times(-12)}{[10\times 980 - (70^2)]^{1/2}[10\times 994 - (12^2)]^{1/2}}$$

$$= \frac{4160}{70\times 98.97} = 0.6 \quad \text{(verify)}$$

Problem 2. If x and y are two variables prove that the correlation coefficient between ax+b and cy+d is $\gamma_{ax+b,cy+d} = \frac{ac}{|ac|}\gamma_{xy}$ if $a, c \neq 0$.

Proof. Let u= ax+b and v = cy+d

$\bar{u}$= a$\bar{x}$+b and $\bar{v}$ = c$\bar{y}$+d

$$\sigma_u^2 = \frac{1}{n}\sum(u-\bar{u})^2 = \frac{a^2}{n}\sum(x_i - \bar{x})^2 = a^2\,\sigma_x^2$$

Similarly, $\sigma_v^2 = c^2\sigma_y^2$

NOTES

Now, $\gamma_{uv} = \frac{\sum(u-\bar{u})^2(v-\bar{v})^2}{n\sigma_u\sigma_v} = \frac{\sum a(x-\bar{x})c(y-\bar{y})}{n|ac|\,\sigma_x\sigma_y}$

$= \frac{ac}{|ac|}\,\gamma_{xy}$

Problem 3. A programmer while writing a program for correlation coefficient between two variables x and y from 30 pairs of observations obtained the following results $\sum x = 300;\ \sum x^2 = 3718$
$\sum y = 210\ \sum y^2 = 2000\ \sum xy = 2100$

At the time of checking it was found that he had copied down two pairs (x_i, y_i)as(18,20) and (12,10) instead of the correct values (10,15) and (20,15). obtain the correct value of the correlation coefficient.

Solution. Corrected $\sum x = 300 - 18 - 12 + 10 + 20 = 300$

Corrected $\sum y = 210 - 20 - 10 + 15 + 15 = 210$

Corrected $\sum x^2 = 3718 - 18^2 - 12^2 + 10^2 + 20^2 = 3750$

Corrected $\sum y^2 = 2000 - 20^2 - 10^2 + 15^2 + 15^2 = 1950$

Corrected$\sum xy = 2100 - (18 \times 20) - (12 \times 10) + (10 \times 15) + (20 \times 15) = 2070$

After correction the correlation coefficient is

$$\gamma_{xy} = \frac{n\sum xy - \sum xy}{[n\sum x^2 - (\sum x)^2]^{1/2}[n\sum y^2 - (\sum y)^2]^{1/2}}$$

$$\therefore \gamma_{xy} = \frac{30\times2070 - 300\times210}{(112500-90000)^{1/2}(58500-44100)^{1/2}}$$

$$= \frac{-900}{(22500)^{1/2}(14400)^{1/2}} = -\frac{900}{150\times120} = -\frac{1}{20}$$

= -0.05

Problem 4. If x and y are uncorrelated variables each having same standard deviation obtain the coefficient of correlation between x+y and y+z.

Solution. Given $\sigma_x = \sigma_y = \sigma_z = \sigma$ (say)

x and y are uncorrelated =>$\sum(x-\bar{x})(y-\bar{y}) = 0$

y and z are uncorrelated =>$\sum(y-\bar{y})(z-\bar{z}) = 0$

z and x are uncorrelated =>$\sum(z-\bar{z})(x-\bar{x}) = 0$

Let u=x+y and v=y+z.

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

Now, $\sigma_u^2 = \frac{1}{n}\sum(u-\bar{u})^2 = \frac{1}{n}\sum[(x-\bar{x})+(y-\bar{y})]^2$

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

$= \sigma_x^2 + \sigma_y^2$ (since $\sum(x-\bar{x})(y-\bar{y}) = 0$

$= 2\sigma^2$

Similarly $\sigma_v^2 = 2\sigma^2$

Now, $\sum(u-\bar{u})(v-\bar{v})$

$= \sum[\{(x-\bar{x})+(y-\bar{y})\}\{(y-\bar{y})+(z-\bar{z})\}]$

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

$= 0 + n\sigma y^2 + 0 + 0 = n\sigma^2$

$\therefore\ \gamma_{uv} = \frac{\sum(u-\bar{u})\,(v-\bar{v})}{n\sigma_u\sigma_v} = \frac{n\sigma^2}{n(2\sigma^2)} = \frac{1}{2}$

Problem 5. Show that the variables u=xcos $\alpha + y\ sin\alpha$ and ν=y cosα- x sinα are uncorrelated if $\alpha = \tan^{-1}\left(\frac{2\gamma_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right)$

Solution. $u_i = x_i\cos\alpha + y_i\sin\alpha$ and $v_i = y_i\cos\alpha - x_i\sin\alpha$

$\therefore \bar{u} = \bar{x}\cos\alpha + \bar{y}\sin\alpha$ and $\bar{v} = \bar{y}\cos\alpha - \bar{x}\sin\alpha$

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

The variables u_i and v_i are correlated if $\sum(u_i - \bar{u})(v_i - \bar{v}) = 0$

$\therefore \sum[(x_i - \bar{x})\cos\alpha + (y_i - \bar{y})\sin\alpha]\,[(y_i - \bar{y})\cos\alpha - (x_i - \bar{x})\sin\alpha] = 0$

$\therefore \sum(x_i - \bar{x})\,(y_i - \bar{y})\cos^2\alpha - \sum(x_i - \bar{x})\,(y_i - \bar{y})\sin^2\alpha$

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

$\therefore n\gamma_{xy}\sigma_x\sigma_y(\cos^2\alpha - \sin^2\alpha) = n\cos\alpha\sin\alpha(\sigma_x^2 - \sigma_y^2)$

$$\therefore \tan 2\alpha = \frac{2\gamma_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

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

NOTES

Problem 6. Show that if X' and Y' are the deviations of the random variables X and Y from their respective means then (i) $\gamma = 1 - \frac{1}{2N}\sum\left(\frac{X_i'}{\sigma_x} - \frac{Y_i'}{\sigma_Y}\right)^2$. (ii)$\gamma = -1 + \frac{1}{2N}\sum\left(\frac{X_i}{\sigma_x} - \frac{Y_i}{\sigma_Y}\right)^2$. Deduce that $-1 \le \gamma \le 1$

Solution. (i) Given that $X'_i = X_i - \overline{X}$ and $Y'_i = Y_i - \overline{Y}$.

$$1 - \frac{1}{2N}\sum\left(\frac{X'_i}{\sigma_x} - \frac{Y'_i}{\sigma_Y}\right)^2 = 1 - \frac{1}{2N}\sum\left(\frac{X_i - \overline{X}}{\sigma_x} - \frac{Y_i - \overline{Y}.}{\sigma_Y}\right)^2$$

$$= 1 - \frac{1}{2N}\left(\sum\frac{(X_i - \overline{X})^2}{{\sigma_x}^2} + \sum\frac{(Y_i - \overline{Y})^2}{{\sigma_Y}^2} + \frac{2\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_x\sigma_Y}\right)$$

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

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

$$= 1 - (1 - \gamma) = \gamma$$

(ii) can similarly be proved

Since $\sum\left(\frac{X_i}{\sigma_x} - \frac{Y_i}{\sigma_Y}\right)^2$ is always positive we have $\frac{1}{2N}\sum\left(\frac{X_i'}{\sigma_x} - \frac{Y_i'}{\sigma_Y}\right)^2$ is positive.

Hence $1 - \frac{1}{2N}\sum\left(\frac{X_i'}{\sigma_x} - \frac{Y_i'}{\sigma_Y}\right)^2 \le 1$

$\therefore$ By (i) $\gamma \le 1$ Similarly by(ii) $-1 \le \gamma$

Hence $-1 \le \gamma \le 1$

Problem 7. Let x, y be two variables with standard deviation σ_x and σ_y respectively. If u=x+ky and v= $x + \left(\frac{\sigma_x}{\sigma_Y}\right)y$ and $\gamma_{uv} = 0$(ie u and v are uncorrelated) find the value of k

Solution. u=x+ky => $\bar{u} = \bar{x} + k\bar{y}$

$\therefore$u-$\bar{u} = (x - \bar{x}) + k(y - \bar{y})$ and ν-$\bar{\nu} = (x - \bar{x}) + \left(\frac{\sigma_x}{\sigma_Y}\right)(y - \bar{y})$

Now, $\gamma_{u\nu} = 0 => Cov(u, v) = 0$

NOTES

$$=>\sum(u - \bar{u})(v - \bar{v}) = 0$$

$$=>\sum[(x - \bar{x}) + k(y - \bar{y})]\left[(x - \bar{x}) + \left(\frac{\sigma_x}{\sigma_Y}\right)(y - \bar{y})\right] = 0$$

$$=>\sum(x - \bar{x})^2 + k\left(\frac{\sigma_x}{\sigma_Y}\right)\sum(y - \bar{y})^2 + k\sum(x - \bar{x})(y - \bar{y})$$

$$+\left(\frac{\sigma_x}{\sigma_Y}\right)\sum(x - \bar{x})(y - \bar{y}) = 0$$

$$=> n\sigma_x^2 + nk\left(\frac{\sigma_x}{\sigma_Y}\right)\sigma_y^2 + n\gamma_{xy}\sigma_x\sigma_y\left(k + \frac{\sigma_x}{\sigma_y}\right) = 0$$

$$=> n\sigma_x + \left[\sigma_x + k\sigma_y + \gamma_{xy}(k\sigma_y + \sigma_x)\right] = 0$$

$$=> n\sigma_x\left[(\sigma_x + k\sigma_y)(1 + \gamma_{xy})\right] = 0$$

$$=> \sigma_x(\sigma_x + k\sigma_y)\left(1 + \gamma_{xy}\right) = 0$$

$$=>\sigma_x + k\sigma_y = 0 \text{ or } \gamma_{xy} + 1 = 0 \text{ or } \sigma_x = 0$$

If $\gamma_{xy} = -1$ and $\sigma_x \neq 0$ we get k= - (σ_x/σ_y)

Exercises.

1. Find the correlation coefficient for the following data.

(i)	X	10	12	18	24	23	27
	Y	13	18	12	25	30	10

(ii)	X	20	18	16	15	14	12	12	10	8	5
	Y	12	14	10	14	12	10	9	8	7	2

(iii)	Age of husband	23	27	28	29	30	31	33	35	36	39
	Age of wife	18	22	23	24	25	26	28	29	30	32

NOTES

7.3 RANK CORRELATION

Suppose that a group of n individuals are arranged in the order of merit or efficiency with respect to some characteristics. Then the rank is a variable which takes only the values1, 2, 3,....., n assuming that there is no tie. Hence $\bar{x} = \frac{1+2+\cdots+n}{n} = \frac{n+1}{2}$ and the variance is given by
$\sigma_x^2 = \frac{1}{12}(n^2-1)$

Now suppose that the same individuals are ranked in two ways on the basis of different characteristics or by two different persons for a single characteristics . Let x_i and y_i be the ranks of the i^{th} individual in the first and second ranking respectively. The coefficient of correlation between the ranks x_i and y_i is called the rank correlation coefficient and is denoted by ρ

Theorem 7.5. Rank correlation ρ is given by $\rho=1-\frac{6\sum(x-y)^2}{n(n^2-1)}$

Proof. Consider a collection of n individuals. Let x_i and y_i be the ranks of i^{th} individual
in the two different rankings.

$$\therefore \bar{x} = \frac{n+1}{2} = \bar{y} \text{ and } \sigma_x^2 = \frac{1}{12}(n^2-1) = \sigma_y^2$$

Now,$\sum(x-y)^2 = \sum[(x-\bar{x})-(y-\bar{y})]^2$ (sice $\bar{x} = \bar{y}$)

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

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

$$= 2n\sigma_x^2(1-\rho) \text{ (since } \sigma_x^2 = \sigma_y^2)$$

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

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

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

Note. This known as **Spearman's formula** for rank correlation coefficient.

Note 2. If two or more individuals get the same rank in the ranking process with respect to different characteristics then Spearman's formula for calculating the rank correlation will not apply since in this case $\bar{x} \neq \bar{y}$. In such case we assign a common rank to be repeated values. This common rank is the average of the ranks which these items would have assumed if their ranks were different from each other and the next item will get the rank next to the rank already assumed. As a result of this in the formula for ρ we add the factor $\frac{1}{12}m(m^2-1)$ to $\sum(x-y)^2$ where m is the number of times an item has repeated values. This correction factor is added for each repeated rank of the variables x and y. For example, after assigning rank 2 if four items get the same rank 3 then these fo ur items are given the common rank $\frac{1}{4}(3+4+5+6)=4.5$ and the next item is given rank 7. In this case the correction factor to be added is $\frac{1}{12}\times 4\times(4^2-1)=5$

NOTES

Problem 1. Find the rank correlation coefficient between the height in c.m and weight in kg of 6 soldiers in Indian Army.

Height	165	167	166	170	169	172
Weight	61	60	63.5	63	61.5	64

Solution.

Height	**Rank in Height** **x**	**Weight**	**Rank in Weight** **y**	**x-y**	$(\mathbf{x-y})^2$
165	6	61	5	1	1
167	4	60	6	-2	4
166	5	63.5	2	3	9
170	2	63	3	-1	1
169	3	61.5	4	-1	1
172	1	64	1	0	0
Total	**-**	**-**	**-**	**-**	**16**

$$\rho=1-\frac{6\sum(x-y)^2}{n(n^2-1)}=1-\frac{6\times 16}{6\times 35}=1-0.457$$

$= 0.543$

Problem 2. From the following data of marks obtained by 10 students in Physics and Chemistry calculate the rank correlation coefficient

Physics(P)	35	56	50	65	44	38	44	50	15	26
Chemistry(Q)	50	35	70	25	35	58	75	60	55	35

Solution. We rank the marks of Physics and Chemistry and we have the following table

P	**Rank in P** **x**	**Q**	**Rank in Q** **y**	**x-y**	$\mathbf{(x-y)^2}$
35	8	50	6	2	4
56	2	35	8	-6	36
50	3.5	70	2	1.5	2.25
65	1	25	10	-9	81
44	5.5	35	8	-2.5	6.25
38	7	58	4	3	9
44	5.5	75	1	4.5	20.25
50	3.5	60	3	0.5	0.25
15	10	55	5	5	25
26	9	35	8	1	1
Total	-	-	-	-	185

We observe that in the values of x the marks 50 and 44 occurs twice. In the values of y the mark 35 occurs thrice. Hence in the calculation of the rank correlation coefficient $\sum(x-y)^2$ is to be corrected by adding the following correction factors $\left[\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}\right]+\frac{3(3^2-1)}{12}$=3.

$\therefore$ After correction $\sum(x-y)^2 = 188$

Now, $\rho=1-\frac{6\sum(x-y)^2}{n(n^2-1)} = 1-\frac{6\times188}{10\times99} = 1-\frac{1128}{990}$

= 1-1.139=-0.139

NOTES

Problem 3. Three judges assign the ranks to 8 entries in a beauty contest.

Judge in X	1	2	4	3	7	6	5	8
Judge in Y	3	2	1	5	4	7	6	8
Judge in Z	1	2	3	4	5	7	8	6

which pair of judges has the nearest approach to common taste in beauty?

Solution.

x	**Y**	**z**	**x-y**	$(\mathbf{x-y})^2$	**y-z**	$(\mathbf{y-z})^2$	**z-x**	$(\mathbf{z-x})^2$
1	3	1	-2	4	2	4	0	0
2	2	2	0	0	0	0	0	0
4	1	3	3	9	-2	4	-1	1
3	5	4	-2	4	1	1	1	1
7	4	5	3	9	-1	1	-2	4
6	7	7	-1	1	0	0	1	1
5	6	8	-1	1	-2	4	3	9
8	8	6	0	0	2	4	-2	4
Total			-	28	-	18	-	20

$\rho_{xy} = 1\text{-}\frac{6\sum(x-y)^2}{n(n^2-1)}$

$=1\text{-}\frac{6\times28}{8(8^2-1)}=1\text{-}\frac{168}{504}=1\text{-}0.333=0.667$

$$\rho_{yz} = 1-\frac{6\times18}{8\times63} = 1-\frac{108}{504} = 1-0.214 = 0.786$$

$\rho_{zx} = 1-\frac{6\times20}{8\times63}=1\text{-}\frac{120}{504} = 1-0.238 = 0.762$

Since ρ_{yz} is greater than ρ_{xy} and ρ_{xz} the judges Mr.Y and Mr.Z have nearest approach to common taste in beauty.

Problem 4. The coefficient of rank correlation of marks obtained by 10 students in Mathematics and Physics was found to be 0.8. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 5 instead of 8. Find the correct coefficient of rank correlation.

Solution. ρ_{xy}=1-$\frac{6\sum(x-y)^2}{n(n^2-1)}$

Given $\rho_{xy} = 0.8$ and $n = 10$

$\therefore 0.8$ = 1-$\frac{6\sum(x-y)^2}{10(10^2-1)}$=1-$\frac{6\sum(x-y)^2}{990}$

$\therefore \frac{6\sum(x-y)^2}{990} = 1 - 0.8 = 0.2.$

$6\sum(x-y)^2 = 990 \times 0.2 = 198$

$\therefore \sum(x-y)^2 = 33$

Corrected $\sum(x-y)^2 = 33 - 5^2 + 8^2 = 72$

Now, after correction $\rho_{xy} = 1 - \frac{6\times72}{10(10^2-1)}$

=1-$\frac{432}{990} = 1 - 0.436$

= 0.564

The correct coefficient of rank correlation is 0.564

Problem 5. Let $x_1, x_2, \dots, x_n$ be the ranks of n indi viduals according to a character A and $y_1, y_2, \dots, y_n$ the ranks of the same individualsaccording to another character B. It is given that $x_i + y_i = 1 + n$ for i=1, 2, 3,....., n Show that the value of the rank correlation coefficient ρ between the characters A and B is -1

Solution. Given $x_i + y_i = 1 + n$ ....(1)

Let the difference of ranks be d_i

$\therefore x_i + y_i = d_i$ ....(2)

Adding (1) and (2) we get $2x_i = 1 + n + d_i$

$$\therefore d_i = 2x_i - (n+1)$$

Now, $\sum d_i^2 = \sum [2x_i - (n+1)]^2$

$$= \sum [4x_i^2 + (n+1)^2 - 4(n+1)x_i]$$

$$= 4\sum x_i^2 + n\,(n+1)^2 - 4(n+1)\sum x_i$$

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

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

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

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

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

Now $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\left[\frac{1}{3}n(n^2-1)\right]}{n(n^2-1)} = 1-2$

$= -1$

Exercises.

1. Calculate the rank correlation coefficient for the following data.

(i)

X	5	2	8	1	4	6	3	7
Y	4	5	7	3	2	8	1	6

(ii)

X	10	12	18	18	15	40
Y	12	18	25	25	50	25

2. Two judges in a beauty contest rank the ten competitors in the following order.

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

Do the judges appear to agree in their standard?

3.Ten students got the following percentage of mark in two subjects.

Economics	78	65	36	98	25	75	82	92	62	39
Statistics	84	53	51	91	60	68	62	86	58	47

Calculate the rank correlation coefficient.

NOTES

UNIT-VIII REGRESSION

8.1 INTRODUCTION

There are two main problems involved in the relationship between x and y. The first is to find a measure of the degree of association or correlation between the values of x and those of y. The second problem is to find the most suitable form of equation for determining the probable value of one variable corresponding to a given value of the order. This is the problem of **regression**.

If there is a functional relationship between the two variables x_i and y_i the points in the scatter diagram will cluster around some curve called the curve of regression. If the curve is straight line it is called a line of regression between the two variables.

8.2 Definition. It we fit a straight line by the principle of least square to the points of the scatter diagram in such a way that the sum of the squares of the distance parallel to the y-axis from the points to the line is minimized we obtain a line of best fit for the data and it is called the **regression line of y on x.**

Similarly we can define the **regression line of x on y.**

Theorem 8.1 The equation of the regression line of y on x is given by $y\text{-}\bar{y} = \gamma \frac{\sigma_y}{\sigma_x}(x - \overline{x})$

Proof. Let y=ax+b be the line of regression of y on x.

According to the principle of least squares constants a and b are to be determined in such a way that $S=\sum[y_i - (ax_i + b)]^2$ is minimum.

$$\frac{\partial s}{\partial a} = 0 => -2\sum(y_i - ax_i - b)x_i = 0$$

$$=> \sum x_i y_i = a\sum x_i^2 + b\sum x_i \qquad \text{.................(1)}$$

$$\frac{\partial s}{\partial a} = 0 => -2\sum(y_i - ax_i - b) = 0$$

$$=> \sum y_i = a\sum x_i + nb \qquad \text{...............(2)}$$

Equations (1) and (2) are called normal equations.

From (2) we obtain $\bar{y} = a\bar{x} + b$ ....(3)

∴ The line of regression passes through the point $(\bar{x}, \bar{y})$.
Now Shifting the origin to this point $(\bar{x}, \bar{y})$ by means of the transformation $X_i = x_i - \bar{x}$ and $Y_i = y_i - \bar{y}$ we obtain $\sum x_i = 0 = \sum y_i$
And the equation of the line of regression becomes Y=aX ....(4)

Corresponding to this line Y=aX, the constant a can be determined from the normal equation. $a\sum x_i^2 = \sum X_i Y_i$

$$\therefore \mathbf{a} = \frac{\sum X_i Y_i}{a\sum x_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\gamma \sigma_x \sigma_y}{\sigma_x^{\mathbf{2}}} = \gamma \frac{\sigma_y}{\sigma_x}$$

∴ The required regression line (4) becomes $Y = \left(\gamma \frac{\sigma_y}{\sigma_x}\right) X$

$$\therefore y - \bar{y} = \left(\gamma \frac{\sigma_y}{\sigma_x}\right)(x - \bar{x})$$

Theorem 8.2 The equation of regression line of x on y is given by
$$x - \bar{x} = \left(\gamma \frac{\sigma_x}{\sigma_y}\right)(y - \bar{y})$$
Proof. Proof is similar to that of theorem 8.1

Note. ($\bar{x}, \bar{y}$)is the point of intersection of the two regression lines.

Definition. The slope of the regression line of y on x is called the **regression coefficient of y on x** and it is denoted by $b_{yx.}$ Hence $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$.
The **regression coefficient of x on y** is given by $b_{xy} = \gamma \frac{\sigma_x}{\sigma_y}$

We now give some properties of the regression coefficients.

Theorem 8.3. Correlation coefficient is the geometric mean between the regression coefficients. (i.e) $\gamma = \pm\sqrt{b_{xy} b_{yx}}$
Proof. We have $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$ **and** $b_{xy} = \gamma \frac{\sigma_x}{\sigma_y}$

$$\therefore b_{yx} b_{xy} = \gamma^2$$

$$\therefore \gamma = \pm\sqrt{b_{xy} b_{yx}}$$

Note. The sign of the correlation coefficient is the same as that of regression coefficients.

NOTES

Theorem 8.4. If one of the regression coefficient is greater than unity the other is less than unity.

Proof. We have $b_{xy}b_{yx} = \gamma^2 \leq 1$ so that $b_{xy}b_{yx} \leq 1$

Hence $b_{xy} > 1 => b_{yx} < 1$

Hence the theorem.

Theorem 8.5. Arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

Proof. Let b_{xy} and b_{yx} be the regression coefficients.

We have to prove that $\frac{1}{2}\left(b_{xy} + b_{yx}\right) \geq \gamma$

Now , $\frac{1}{2}\left(b_{xy} + b_{yx}\right) \geq \gamma \Leftrightarrow b_{yx} + b_{xy} \geq 2\gamma$

$$\Leftrightarrow \boldsymbol{\gamma}\frac{\sigma_y}{\sigma_x} + \boldsymbol{\gamma}\frac{\sigma_x}{\sigma_y} \geq 2\gamma$$

$$\Leftrightarrow \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2$$

$$\Leftrightarrow \sigma_x{}^{\mathbf{2}} + \sigma_y{}^{\mathbf{2}}\text{-}\mathbf{2}\sigma_x\sigma_y \geq 0$$

$$\Leftrightarrow(\sigma_x - \sigma_y)^2 \geq 0$$

This is always true. Hence the theorem.

Theorem 8.6 Regression coefficients are independent of the change of origin but dependent on change of scale.

Proof. Let $u_i = \frac{x_i - A}{k}$ and $v_i = \frac{y_i - B}{k}$

Let $x_i = A + hu_i$ and $y_i = B + kv_i$

We know that $\sigma_x = h\sigma_u$; $\sigma_y = k\sigma_v$ and $\gamma_{xy} = \gamma_{uv}$

Now, $b_{yx} = \gamma_{xy}\frac{\sigma_y}{\sigma_x} = \gamma_{uv}\left(\frac{k\sigma_v}{h\sigma_u}\right) = \frac{k}{h}b_{uv}$ .... (1)

Similarly $b_{xy} = \left(\frac{h}{k}\right) b_{uv}$ ....(2)

From (1) and (2) we observe that b_{yx} and b_{xy} depend upon the scales h and k but not on the origin A and B.

Hence the theorem.

Theorem 8.7 The angle between the two regression lines is given by $\theta = \tan^{-1}\left[\left(\frac{\gamma^2-1}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)\right]$.

Proof. The equations of lines of regression of y on x and x on y respectively are $y-\bar{y}=\left(\gamma\frac{\sigma_y}{\sigma_x}\right)(x-\bar{x})$ ....(1)

$$x-\bar{x}=\left(\gamma\frac{\sigma_x}{\sigma_y}\right)(y-\bar{y}) \quad ....(2)$$

(2) can also be written as $y-\bar{y}=\left(\frac{1}{\gamma}\frac{\sigma_y}{\sigma_x}\right)(x-\bar{x})$ ....(3)

∴ Slopes of the two lines (1) and (2) are $\gamma\frac{\sigma_y}{\sigma_x}$ **and** $\frac{1}{\gamma}\frac{\sigma_y}{\sigma_x}$.

Let θ be the acute angle between the two lines or regression.

$$\therefore \tan\theta = \frac{\gamma\frac{\sigma_y}{\sigma_x}\sim\frac{1\sigma_y}{\gamma\sigma_x}}{\mathbf{1}+\left(\gamma\frac{\sigma_y}{\sigma_x}\right)\left(\frac{1\sigma_y}{\gamma\sigma_x}\right)}$$

$$=\frac{\gamma^2-1}{\gamma}\left(\frac{\sigma_x+\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)$$

$$=\frac{1-\gamma^2}{\gamma}\left(\frac{\sigma_x+\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right) \text{ (since } \gamma^2 \leq 1 \; and \; \theta \, is \; acute).$$

$$\therefore \theta = \tan^{-1}\left[\left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)\right]$$

Note1. The obtuse angle between the regression lines is given by $\tan^{-1}\left[\left(\frac{\gamma^2-1}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)\right]$

Note2. If γ=0 then tanθ=∞. Hence θ=π/2. Thus if the two variables are uncorrelated then the lines of regression are perpendicular to each other.

Note 3. If γ=±1 then tanθ - 0

Hence θ=0 or π

∴ The two lines of regression are parallel .

Further the two lines have the common point $(\bar{x}, \bar{y})$ and hence they must be coincident.

Therefore if there is a perfect correlation (positive or negative between the two variables then the two lines of regression coincide.

Solved Problems.

Problem 1. The following data relate to the marks of 10 students in the internal test and the university examination for the maximum of 50 in each.

Internal marks	25	28	30	32	35	36	38	39	42	45
University marks	20	26	29	30	25	18	26	35	35	46

(i) Obtain the two regression equations and determine

(ii) the most likely internal mark for the university mark of 25

(iii) the most likely university mark for the internal mark of 30

Solution. (i) Let the marks of internal test and university examination be denoted by x and y respectively.

we have $\bar{x} = \frac{1}{10}\sum x_i = 35$ and $\bar{y} = \sum y_i = 29$

For the calculation of regression we have the following table.

$\boldsymbol{x_i}$	$\boldsymbol{x_i - 35}$	$\boldsymbol{(x_i - 35)^2}$	$\mathbf{y_i}$	$\mathbf{y_i - 29}$	$\mathbf{(y_i - 29)^2}$	$\boldsymbol{(x_i - 35)\,(y_i - 29)}$
25	-10	100	20	-9	81	90
28	-7	49	26	-3	9	21
30	-5	25	29	0	0	0
32	-3	9	30	1	1	-3
35	0	0	25	-4	16	0
36	1	1	18	-11	121	-11
38	3	9	26	-3	9	-9
39	4	16	35	6	36	24
42	7	49	35	6	36	42
45	10	100	46	17	289	170
Total	**0**	**358**	**-**	**0**	**598**	**324**

$$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1}{10}\sum(x_i - 35)^2 = 35.8$$

$\sigma_y^2 = \frac{\sum(y_i-\bar{y})^2}{n} = \frac{1}{10}\sum(y_i - 29)^2 = 59.8$

$\therefore \sigma_x = 5.98$ and $\sigma_y = 7.73$

$\therefore \gamma = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x\sigma_y} = \frac{324}{10\times5.98\times7.73}$

$= \frac{324}{462.254} = 0.7$ (approximately)

NOTES

Now, the regression of y on x is (y-$\bar{y}$) $= \gamma\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

$\therefore \gamma\frac{\sigma_y}{\sigma_x} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\sigma_x^2} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$

$= \frac{324}{358} = 0.905$

Similarly $\gamma\frac{\sigma_x}{\sigma_y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(y_i-\bar{y})^2} = \frac{324}{598} = 0.542$

The regression line of y on x is y-29=0.905(x-35)

(i.e) y=0.905x-2.675 .... (1)

The regression line of x on y x-35=0.542(y-29)

(i.e) x=0.542y+19.282 ....(2)

(1) and (2) are the required regression equations.

(ii) The most likely internal mark for the university mark of 25 is got from the regression equation of x on y by putting y=25

$\therefore$ x=0.542 ×25+19.282=32.83

(iii) The most likely internal mark for the university mark of 30 is got from the regression equation of y on x by putting x=30

$\therefore$ y=0.905× 30 -2.675 =24.475

Problem 2. For the solved problem 1 of 6.1 estimate the university examination mark of a student who got 61 in the college internal test.

Solution. We have to find the equation of regression line of y on x and then estimate the value of y for the given value of x=61

The regression line of y on x is given by (y-$\bar{y}$) $= \gamma\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

$\bar{x} = A + h\bar{u} = 63 + \left(\frac{-70}{10}\right) = 56$

$\bar{y} = B + h\bar{v} = 60 + \left(\frac{-12}{10}\right) = 58.8$

$$\sigma_u = \left[\frac{\sum u_i^2}{n} - \left(\frac{\sum u_i}{n}\right)^2\right]^{1/2} = \left[\frac{980}{10} - \left(\frac{-70}{10}\right)^2\right]^{1/2} = 7$$

$$\sigma_\nu = \left[\frac{\sum \nu_i^2}{n} - \left(\frac{\sum \nu_i}{n}\right)^2\right]^{1/2} = \left[\frac{994}{10} - \left(\frac{-12}{10}\right)^2\right]^{1/2} = 9.898$$

since the scale factor for u_i and $\nu_i$is one we note $\sigma_x = \sigma_u$ and $\sigma_y = \sigma_\nu$

NOTES

we have γ=0.6 for the solved problem 1 of 6.1

∴Regression equation of y on x is y-58.8=0.6$\left(\frac{9.898}{7}\right)(x - 56)$

∴7y=5.9388x + 79.0272

when x=62, 7y=5.9388×60+79.0272=441.294

∴y=63 (approximately)

∴When the internal test mark is 61 the university examination mark is estimated to be 63

Problem 3. Out of the two lines of regression given by x+2y-5=0 and 2x+3y-8=0 which one is the regression line of x on y.?

Solution. Suppose x+2y-5=0 is the equation of the regression line of x on y and 2x+3y-8=0 is the equation of the regression line of y on x.

Then the two equation can be written as x=-2y+5 and y= -$\frac{2}{3}$x+$\frac{8}{3}$

Hence the two regression coefficient $b_{yx} = -\frac{2}{3}$ and $b_{xy} = -2$

Now $\gamma^2 = b_{yx}b_{xy} = \frac{4}{3} > 1.$ This is impossible.

Hence our assumption is wrong.

∴2x+3y-8=0 is the equation of the regression line of x on y.

Problem 4. The two variables x and y have the regression lines 3x+2y-26=0. and 6x+y-31=0

Find (i) the mean values of x and y

(ii) the correlation coefficient between x and y

(iii) the variance of y if the variance of x is 25

Solution. (i) Since the two lines of regression pass through $(\bar{x}, \bar{y})$ we have

$3\bar{x}+2\bar{y} = 26$ .... (1)

$6\bar{x}+\bar{y} = 31$ ....(2)

Solving (1) and (2) we get $\bar{x} = 4$ and $\bar{y} = 7$

(ii) As in the previous problem we can prove that $y= -\frac{3}{2}x + 13$ and $x=-\frac{1}{6}y + \frac{31}{6}$ represent the regression lines of y on x and x on y respectively .

Hence we get the regression coefficient as $b_{yx} = -\frac{3}{2}$ and $b_{xy} = -\frac{1}{6}$

$$\text{Now, } \gamma^2 = \left(-\frac{3}{2}\right) \times \left(-\frac{1}{6}\right) = \frac{1}{4}$$

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

Since both the regression coefficients are negative we take $\gamma= -\frac{1}{2}$.

(iii) Given $\sigma_x = 5$

We have $b_{yx} = \gamma \frac{\sigma_y}{\sigma_x}$

$\therefore -\frac{3}{2} = \left(-\frac{1}{2}\right)\left(\frac{\sigma_y}{5}\right)$

$\therefore \sigma_y = 15$

Problem 5. If x= 4y+5 and y = kx+4 are the regression lines of x on y and y on x respectively (i) show that $0 \leq k \leq 1/4$

(ii)If k=1/8 find the means of the two variables x and y and the correlation coefficient between them.

Solution. (i) The regression line of x on y is x= 4y+5

Hence $b_{xy} = 4$

The regression line of y on x is y =kx+4. Hence $b_{yx} = k$

Now, $b_{xy}\, b_{yx} = \gamma^2 \Rightarrow 4k = \gamma^2$

Now, $0\leq \gamma^2 \leq 1 \Rightarrow 0\leq 4k \leq 1$

$\Rightarrow 0\leq k\leq 1/4$

(ii)Given k=1/8. Hence $b_{yx} = 1/8$

$\therefore \gamma^2 = b_{yx}b_{xy} = \frac{1}{8} \times 4 = \frac{1}{2}$

Hence $\gamma=\frac{1}{\sqrt{2}}$. (Positive value of γ is taken since both regression coefficient are positive)

Let $\bar{x}$ and $\bar{y}$ be the means of the two variables x and y .

NOTES

Since the regression lines pass through $(\bar{x}, \bar{y})$ we have $\bar{x}= 4\bar{y}+5$ and

$\bar{y} = \frac{1}{8}\bar{x} + 4 \left(\text{Taking } k = \frac{1}{8}\right)$

Solving for $\bar{x}$ and $\bar{y}$ we get $\bar{x} = 42$ and $\bar{y} = 9.25$

Problem 6. The variables x and y are connected by the equation ax+by+c=0. Show that $r_{xy} = -1$ or 1 according as a and b are of the same sign or of opposite sign.

Solution. Writing ax + by + c=0 in the form $y= -\frac{b}{a}y - \frac{c}{a}$ we get the regression coefficient of x on y is $b_{xy} = -\frac{b}{a}$

Now $\gamma^2 = b_{yx}b_{xy} \Rightarrow \gamma^2 = -\left(\frac{a}{b}\right)\left(\frac{b}{a}\right)$

Suppose a and b are of same sign. Then $\gamma^2 = 1$

Hence γ=1 (since b_{yx} and b_{xy} are positive)

Problem 7. If θ is the acute angle between the two regression lines show that $\theta \leq 1- \gamma^2$.

Solution. We know that if θ is the acute angle between the two regression lines we have $\tan\theta =\left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}\right)$ ....(1)

We claim that $\sigma_x{}^2 + \sigma_y{}^2 \geq 2\sigma_x\sigma_y$

Suppose not, then $\sigma_x{}^2 + \sigma_y{}^2 < 2\sigma_x\sigma_y$

(ie) $\sigma_x{}^2 + \sigma_y{}^2 - 2\sigma_x\sigma_y < 0$

(ie) $(\sigma_x - \sigma_y)^2 < 0$ This is impossible.

Hence $\sigma_x{}^2 + \sigma_y{}^2 \geq 2\sigma_x\sigma_y$.

$\therefore \frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2} \leq \frac{1}{2}$From (1) we get $\tan\theta \leq \left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{1}{2}\right)$

$\therefore \tan\theta \leq \left(\frac{1-\gamma^2}{2\gamma}\right)$. Hence $\sin\theta \leq \left(\frac{1-\gamma^2}{1+\gamma^2}\right)$

$\therefore \sin\theta \leq 1 - \gamma^2$

Exercises.

1. Calculate the coefficient of correlation and obtain the lines of regression for the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

2. The following table shows the ages x and blood pressure y of 12 women. (i) Find the correlation coefficient between x and y.

(ii)Determine the regression equation of y on x.

(iii)Estimate the blood pressure of a women whose age is 45 years.

Age(x)	**Blood Pressure(y)**	**Age (y)**	**Blood Pressure(y)**
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

3.Calculate the coefficient of correlation for the following data.

x	3	6	5	4	4	6	7	5
y	3	2	3	5	3	6	6	4

UNIT-IX CORRELATION COEFFICIENT FOR A BIVARIATE FREQUENCY DISTRIBUTION

NOTES

9.1 INTRODUTION :

As in the case of a single variable large collection of data corresponding to two variables under consideration can also be conveniently summarised in a 2-way frequency table (bivariate frequency table)which is illustrated below.

If there are n classes for the first variable x and m classes for the second variable y then there are mn cells in the 2 - way table. If x_i and y_i denote the mid values of the i^th^ class for x and j^th^ class for y respectively then the frequency f_{ij} corressponding to (x_i y_i) is entered in the (i,j)^th^ cell. For i = 1,2,.....,n and j = 1,2,....,m we get all the mn cells, in the table.

From the bivariate frequency table we note the following :

(1) For any fixed i we have $\sum_{j=1}^{m} f_{ij}$ = g_i = the sum of all the cell frequencies of the i^th^ column.

(2) For any fixed j we have $\sum_{i=1}^{n} f_{ij}$ = f_j = the sum of the cell frequencies of the j^th^ row.

x y		Mid points of x						Total frequencies of y
		x_1	x_2	....	x_i	....	x_n	
Mid points of y	y_1	f_{11}	f_{21}	....	f_{i1}	....	f_{n1}	f_1
	y_2	f_{12}	f_{22}	....	f_{i2}	....	f_{n2}	f_2
	: : :	: : :	: : :	: : :	: : :	: : :	: : :	: : :
	y_i			....	f_{ij}	....	f_{nj}	f_j
	: : :	: : :	: : :	: : :	: : :	: : :	: : :	: : :
	y_m	f_{1m}	f_{2m}	....	f_{im}	....	f_{nm}	f_m
Total frequencies of x		g_1	g_2	....	g_i	....	g_n	N $\sum_{i=1}^{n} g_i =$ $\sum_{j=1}^{m} f_i$

(3) If the total frequency of all the mn cells is N then

$$N = \sum_{i=1}^{n} g_i = \sum_{j=1}^{m} f_i \text{ and } N = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij}$$

(4) $\bar{x} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij}\ x_i = \frac{1}{N} \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} f_{ij} \right) = \frac{1}{N} \sum_{i=1}^{n} x_i\ g_i$

$\frac{1}{N}\sum_{i=1}^{n} g_i x_i$ from (1)

(5) Similarly $\bar{y} = \frac{1}{N}\sum_{j=1}^{m} f_i y_i$

(6) $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}\, x_i^2 - \bar{x}^2 = \frac{1}{N}\sum_{i=1}^{n} g_i x_i^2 - \bar{x}^2$

(7) Similarly $\sigma_y^2 \sum_{j=1}^{n} f_i\, y_i^2 - \bar{y}^2$

(8) cov (x,y) = $\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}\, x_i y_{i} - \bar{x}\bar{y}$

The correlation coefficient between x and y is given by $\gamma_{xy} = \frac{\cos(x,y)}{\sigma_x\sigma_y}$

$$\therefore \gamma_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\, \mathbf{xi\, yi} - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)\left(\sum_{j=1}^{m} f_i y_i\right)}{\sqrt{\sum_{i=1}^{n} g_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{n} g_i x_i\right)^2}\sqrt{\sum_{j=1}^{m} f_i y_i^2 - \frac{1}{N}\left(\sum_{j=1}^{m} f_i y_i\right)^2}}$$

Note. Since correlation coefficient is independent of origin and scale if x and y are transformed to u and v by the formula u=$\frac{\mathbf{x-A}}{\mathbf{h}}$ and v=$\frac{\mathbf{y-B}}{\mathbf{k}}$ then we have $\gamma_{xy} = \gamma_{uv}$

Solved problems

Problem 1. Find the Correlation Coefficient between x and y from the following table**.**

x y	**5**	**10**	**15**	**20**
4	**2**	**4**	**5**	**4**
6	**5**	**3**	**6**	**2**
8	**3**	**8**	**2**	**3**

Solution :

x		x_1	x_2	x_3	x_4	Total
y		5	10	15	20	
y_1	4	2	4	5	4	$f_1 = 15$
y_2	6	5	3	6	2	$f_2 = 16$
y_3	8	3	8	2	3	$f_3 = 16$
Total		$g_1 = 10$	$g_2 = 15$	$g_3 = 13$	$g_4 = 9$	N=47

Correlation coefficient between x and y is given by

$$\gamma_{xy} = \frac{\sum\sum f_{ij}\, x_i y_j - \frac{1}{N}(\sum g_i x_i)\,(\sum f_i y_i)}{\sqrt{\sum g_i x_i^2 - \frac{1}{N}(\sum g_i x_i)^2}\;\sqrt{\sum f_i y_i^2 - \frac{1}{N}(\sum f_i y_i)^2}}$$

where i = 1,2,3,4 and j = 1,2,3

$\sum g_i\, x_i$=50 + 150 + 195 + 180 = 575

$\sum f_i\ \ y_i$= 60 + 96 + 128 = 284

$\sum g_i x_i^2$ = 250 + 1500 + 2925 + 3600 = 8275

$\sum f_j y_j^2$ = 240 + 576 + 1024 =1840

$\sum f_{ij}\ x_i y_i$ = (40 + 160 + 300 + 320) + (150 + 180 + 540 + 240)

+ (120 + 640 + 240 + 480) = 3410

$$\gamma_{xy} = \frac{3410 - \frac{1}{47}\ (575 \times 284)}{\sqrt{8275 - \frac{1}{47}\,(575)^2}\ \sqrt{1840 - \frac{1}{47}\,(284)^2}}$$

$$= \frac{3410 \times 47 - (575 \times 284\,)}{\sqrt{8275 \times \frac{1}{47} - 575^2}\ \sqrt{1840 \times \frac{1}{47} - 284^2}}$$

$$= \frac{160270 - 163300}{\sqrt{388925 - 330625}\quad \sqrt{86480 - 80656}}$$

$$= \frac{-3030}{\sqrt{58300}\ \ \sqrt{5824}} = \frac{-3030}{241.5 \times 76.3} = \frac{-3030}{18426.5}$$

= -0.16

NOTES

Problem 2 . Find the correlation coefficient between heights and weights of 100 students which are distributed as follows.

Height in c.m.	**Weight in kgs**					**Total**
	30-40	**40-50**	**50-60**	**60-70**	**70-80**	
150-155	1	3	7	5	2	**18**
155-160	2	4	10	7	4	**27**
160-165	1	5	12	10	7	**35**
165-170	-	3	8	6	3	**20**
Total	**4**	**15**	**37**	**28**	**16**	**100**

Solution : Let x_i denote the mid value of the classes of weights and y_i denote the mid value of the classes of height.

Let u_i $\frac{x_i - 55}{10}$ and $v_{i\,=}$ $\frac{y_j - 157.5}{5}$

Then the 2 way frequency table is given below.

X_i / yi	35	45	55	65	75	f_i	v_j	$f_i v_j$	$f_i v_j^2$	$f_{ij} u_i v_j$
152.5	(2) 1	(3) 3	(0) 7	(-5) 5	(-4) 2	18	-1	-18	18	(-4)
157.5	(0) 2	(0) 4	(0) 10	(0) 7	(0) 4	27	0	0	0	(0)
162.5	(-2) 1	(-5) 5	(0) 12	(10) 10	(14) 7	35	1	35	35	(17)
167.5	-	(-6) 3	(0) 8	(12) 6	(12) 3	20	2	40	80	(18)
g_i	4	15	37	28	16	100	-	57	133	(31)
u_i	-2	-1	0	1	2	-				
$g_i\ u_i$	-8	-15	0	28	32	37				
$g_i u_i^2$	16	15	0	18	64	123				
$f_{ij} u_i v_j$	(0)	(-8)	(0)	(17)	(22)	(31)				

$$\xleftrightarrow{\quad} \quad \boldsymbol{\Sigma\,\Sigma\, f_{ij}\ u_i\ v_j}$$

$$\gamma_{xy} = \gamma_{uv} = \frac{\sum f_{ij}\, u_i\, v_j - \frac{1}{N}\left(\sum g_i u_i\right)\left(\sum f_j\, v_j\right)}{\sqrt{\sum g_i u_i^2 - \frac{1}{N}\left(\sum g_i\, u_i\right)^2}\ \sqrt{\sum f_j v_j^2 - \frac{1}{N}\left(\sum f_j\, v_j\right)^2}}$$

$$= \frac{-31 - \frac{1}{100}\ (37 \times 57)}{\sqrt{123 - \frac{1}{100}\,(-37)^2\ \sqrt{133 - \frac{1}{100}(57)^2}}}$$

$$= \frac{3100 - 37 \times 57}{\sqrt{12300 - 37^2}\,\sqrt{13300 - 57^2}} = \frac{991}{\sqrt{10931}\,\sqrt{10051}}$$

$$= \frac{991}{104.5\ \times 100.25} = \frac{991}{10476} = .09$$

Exercises

1. Calculate the coefficient of correlation between the ages of husbands and wife from the following data.

Ages of wives in years	**Age of husbands in years**			**Total**
	20 - 25	25 - 30	30 - 35	
16 - 20	9	14	-	**23**
20 - 24	6	11	3	**20**
24 - 28	-	-	7	**7**
Total	**15**	**25**	**10**	**50**

2. Calculate the coefficient of correlation between the ages hundred mothers and daughters from the following data.

Ages of mothers in years	Ages of daughters in years					Total
	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	
15-25	6	3	-	-	-	9
25-35	3	16	10	-	-	29
35-45	-	10	15	7	-	32
45-55	-	-	7	10	4	21
55-65	-	-	-	4	5	9
Total	**9**	**29**	**32**	**21**	**9**	**100**

UNIT-X INTERPOLATION

NOTES

10.1 INTRODUCTION

Definition. Interpolation is the process of finding the most appropriate estimate for missing data. It is the "art of reading between the lines of a table". For making the most probable estimate it requires the following assumptions.

(i)The frequency distribution is normal and not marked by sudden ups and downs.

(ii) The changes in the series are uniform within a period.

Interpolation technique is used in various disciplines like economics, business, population studies, price determination etc, . It is used to fill in the gaps in the statistical data for the sake of continuity of information. For example, If we know the records for the past five years except the third year which is not available due to unforeseen conditions the interpolation technique helps to estimate the record for that year too under the assumption that the changes in the records over these five years have been uniform.

It is also possible that we may require information for future inwhich case the process of estimating the most appropriate value is known as extrapolation. There are two methods in interpolation.

(i) **Graphic method**
(ii) **Algebraic method**

(i)**Graphic method** is a simple method in which we just plot the available data on a graph sheet and read off the value for the missing period from the graph itself.

(ii) **Algebraic method.** There are several methods used for interpolation of which we deal with the following: (i) Finite differences. (ii) Gregory – Newton's formula

(iii)Lagrange's formula

10.2 FINITE DIFFERENCES.

The operator Δ. $\mathrm{U_x}$ is a function of the independent variable x and if a, a+h, a+2h,........ are a finite set of equidistant values then $\mathrm{U_a, U_{a+h}, U_{a+2h}, \dots \dots}$ are the corresponding values for $\mathrm{U_x}$. The values of the independent variable x are called arguments, the corresponding values of $\mathrm{U_x}$ are called entries and his known as the interval of differencing. Hence $\mathrm{U_{a+h}}$ is the entry for the argument a+h.

Definition. We define an operator Δ which is known as the first order difference on U_x as $\Delta U_x = U_{x+h} - U_x$ where x= a, a+h, a+2h,........ In particular (i) $\Delta U_a = U_{a+h} - U_a$ (ii) $\Delta U_x = 0$. If U_x is constant .

Higher order differences can similarly be defined.

Example. $\boldsymbol{\Delta}^{\mathbf{2}} U_x = \Delta(\Delta U_x) = \Delta(U_{x+h} - U_x)$

$= (U_{x+2h} - U_{x+h}) - (U_{x+h} - U_x)$ ....(1)

$= U_{x+2h} - 2U_{x+h} + U_x$

Also from (1) $\boldsymbol{\Delta}^{\mathbf{2}} U_x = U_{x+h} - \Delta U_x$

Note1. Unless otherwise stated interval of differencing is taken as 1.

Note 2. It is very easy to verify that the operator Δ satisfies the basic laws of algebra.

(i)Δis linear (ie) $\Delta(aU_x + bV_x) = a\Delta U_x + b\Delta V_x$

(ii)Δ satisfies the law of indices for multiplication.

(ie)$\Delta^m \Delta^n U_x = \Delta^{m+n} U_x$

We can construct the difference table for any number arguments and a sample difference table is exhibited below for five consecutive arguments.

Argument x	Entries U_x	1st diff ΔU_x	2nd diff $\Delta^2 U_x$	3rd diff $\Delta^3 U_x$	4th diff $\Delta^4 U_x$
A	U_a				
a+h		ΔU_a			
	U_{a+h}		$\Delta^2 U_a$		
a+2h		ΔU_{a+h}			
	U_{a+2h}			$\Delta^3 U_a$	
			$\Delta^2 U_{a+h}$		
a+3h		ΔU_{a+2h}			$\Delta^4 U_a$
	U_{a+3h}			$\Delta^3 U_{a+h}$	
			$\Delta^2 U_{a+2h}$		
a+4h		ΔU_{a+3h}			
	U_{a+4h}				

In this table U_a is known as the first entry and ΔU_a, $\Delta^2 U_a$, $\Delta^3 U_a$, $\Delta^4 U_a$ are called leading differences.

Example. The difference table for the following data is given below

x	0	1	2	3	4
U_x	8	11	9	15	6

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$
0	8				
		3			
1	11		-5		
		-2		13	
2	9		8		-36
		6		-23	
3	15		-15		
		-9			
4	6				

The differences ΔU_x, $\Delta^2 U_x$ etc are called forward differences. Since they involve functional values which are to the right of U_x. In cintrast to the forward differences we havw another kind of differences known as backward differences which involve functional values to the left of U_x. For this purpose we define another operator ∇(nebla) as

$\nabla U_x = U_x - U_{x-h}$ where x= a, a+h, a+2h,....., a+nh.

This is called the first order backward difference of U_x

We note that $\nabla U_{x+h} = \Delta U_x$. Thus there is a mutual relation between Δ and ∇.

The backward difference table for 5 consecutive arguments is given below.

NOTES

X	U_x	∇U_x	$\nabla^{2} U_x$	$\nabla^{3} U_x$	$\nabla^{4} U_x$
A	U_a				
		∇U_{a+h}			
a+h	U_{a+h}		$\nabla^2 U_{a+2h}$		
		∇U_{a+2h}			
a+2h	U_{a+2h}			$\nabla^3 U_{a+3h}$	
			$\nabla^2 U_{a+3h}$		
		∇U_{a+3h}			$\nabla^4 U_{a+4h}$
a+3h	U_{a+3h}			$\nabla^3 U_{a+4h}$	
			$\nabla^2 U_{a+4h}$		
		∇U_{a+4h}			
a+4h	U_{a+4h}				

Higher order backward differences can similarly be defined as

$$\nabla^n U_x = \nabla^{n-1} U_x - \nabla^{n-1} U_{x-h}$$

Further it can easily be verified that $\nabla^n U_{a+nh} = \nabla^n U_{a.}$

Hence the same forward difference table constructed for U_x can as well be employed to find out the backward differences of U_x and its higher order differences.

For example, from the above example, $\nabla^2 U_2 = -5 = \nabla^2 U_0$; $\nabla^4 U_4 = -36 = \nabla^4 U_0$etc.

The operator E.

Definition. The operator E on U_x is defined as $EU_x = U_{x+h}$

The higher order operator of E can similarly be defined.

Generally $E^n U_x = U_{x+nh}$

If h=1, then $E^n U_x = U_{x+nh}$

For example, $E^5 U_0 = U_{0+5} = U_5$

$E^3 U_4 = U_{4+3} = U_7$

$E^4 U_{-1} = U_{4-1} = U_3$

Theorem 10.1 (i)$E=1+\Delta$ (ii) $E=(1-\nabla)^{-1}$

Proof. Let U_x be a function of x

(i) $\Delta U_x = U_{x+h} - U_x$ (by definition)

$= EU_x - U_x$

$= (E-1) U_x$

$\therefore \Delta = E-1$. Hence $E=1+\Delta$

(iii) $\nabla U_x = U_x - U_{x-h}$

$\therefore U_{x-h} = U_x - \nabla U_x$

$= (1-\nabla)U_x$

$E^{-1}U_x = (1-\nabla)U_x$ (since $EU_{x-h} = U_x \Rightarrow U_{x-h} = E^{-1}U_x$)

$\therefore E^{-1} = 1-\nabla$. Hence $(1-\nabla)^{-1}$

Note 1. It is very easy to verify that the operator E also satisfies the basic laws of algebra such as linearity and law of indices for multiplication.

Lemma. The two operators Δ and E are commutative under composition of operations. (ie) $\Delta \circ E = E \circ A$

Proof. $(\Delta \circ E)(U_x) = \Delta(E(U_x)) = \Delta(U_{x+h}) = U_{x+2h} - U_{x+h}$

$= EU_{x+h} - EU_x = E(U_{x+h} - U_x)$

$= E(\Delta(U_x)) = (E \circ A)(U_x)$

Hence $\Delta \circ E = E \circ A$

Theorem 10.2 (Fundamental theorem for the finite differences)

If U_x is a polynomial of degree n then $\Delta^r U_x = \begin{cases} \text{constant} & \text{if } r = n \\ 0 & \text{if } r > n \end{cases}$

(ie) the n^{th} order difference of a polynomial of degree n is constant and differences of order higher than n are zero.

Proof. Let $U_x = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$

Where $a_0, a_1, a_2, \dots, a_n$ are constants and $a_n = 0$

$\Delta U_x = U_{x+h} - U_x$

$=[a_0(x+h)^n + a_1(x+h)^{n-1} + \cdots + a_{n-1}(x+h) + a_n]$

$-[a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n]$

$= \left[a_0\left(x^n + n_{c_1}x^{n-1}h + \cdots + h^n\right) + a_1\left(x^{n-1} + +n - 1_{c_1}x^{n-2}h +\right.\right.$
...+hn−1+...+an−a0xn+a1xn−1+*...+* an−1x +an

$= a_0nhx^{n-1} + b_2x^{n-2} + \dots + b_{n-1}x + b_n$ where
$b_2, b_3, \dots, b_n$ are constants independent of x and $a_0nh \neq 0$

$\therefore \Delta U_x$ is a polynomial of degree n-1

Continuing this process we get $\Delta^2 U_x$ is a polynomial of degree n-2

$\Delta^3 U_x$ is polynomial of degree n-3

....

....

$\Delta^n U_x = a_0 n(n-1)(n-2)\dots2.1h^n x^0 = a_0 n!\, h^n$ = constant....(1)

And $\Delta^r U_x = 0$ forr>n

Note. In Particular if the interval of differencing is unity and $U_x = a_0x^n$,

Then $\Delta^n(a_0x^n) = a_0n(n-1)\dots.2.1(\text{using } 1)$

$\therefore a_0\Delta^n(x^n) = a_0 n!$

$\therefore \Delta^n(x^n) = n!$

Solved Problems.

Problem 1. Find the second order differences for (i) $U_x = ab^{cx}$

(ii) $U_x = \frac{x}{x^2+7x+12}$ taking interval of differencing as 1.

Solution. (i) $\Delta U_x = U_{x+h} - U_x = ab^{c(x+h)} - ab^{cx} = ab^{cx}b^{ch} - ab^{cx}$

$$= ab^{cx}(b^{ch} - 1)$$

$$\Delta^2 U_x = (b^{ch} - 1)\Delta(\,ab^{cx}) = (b^{ch} - 1)^2 ab^{cx}$$

(ii) $U_x = \frac{x}{x^2+7x+12} = \frac{4}{x+4} - \frac{3}{x+3}$ (by partial fraction)

$$\therefore \Delta U_x = \left[\frac{4}{(x+1)+4} - \frac{3}{(x+1)+3}\right] - \left[\frac{4}{x+4} - \frac{3}{x+3}\right]$$

$$= \frac{4}{x+5} - \frac{3}{x+4} - \frac{4}{x+4} + \frac{3}{x+3}$$

$$= \frac{4}{x+5} - \frac{7}{x+4} + \frac{3}{x+3}.$$

Similarly $\Delta^2 U_x = \frac{4}{x+6} - \frac{11x}{x+5} + \frac{10}{x+4} - \frac{3}{x+3}$ (verify)

Problem 2. Find $\Delta^n sinx\ taking\ h = 1$

Solution. Δsinx=sin(x+1)-sinx=2 $\cos\left(x + \frac{1}{2}\right)\sin\left(\frac{1}{2}\right)$

$$= 2\sin\left(\frac{1}{2}\right)\sin\left(x + \frac{1}{2} + \frac{\pi}{2}\right)$$

$$\text{Now}\Delta^2 \sin x = \Delta\left[2\sin\left(\frac{1}{2}\right)\ \sin\left(x + \frac{1}{2} + \frac{\pi}{2}\right)\right]$$

$$= 2\sin\left(\frac{1}{2}\right)\left[\sin\left(x + \frac{1}{2} + \frac{\pi}{2} + 1\right) - \sin\left(x + \frac{1}{2} + \frac{\pi}{2}\right)\right]$$

$$= 2\sin\left(\frac{1}{2}\right)\left[2\cos\left(x + 1 + \frac{1}{2}\right)\sin\left(\frac{1}{2}\right)\right]$$

$$= \left[2\sin\left(\frac{1}{2}\right)\right]^2 \sin\left(x + 2\left(\frac{1}{2} + \frac{\pi}{2}\right)\right)$$

Proceeding like this we get $\Delta^n sinx = \left[2\sin\left(\frac{1}{2}\right)\right]^n \sin\left(x + n\left(\frac{1}{2} + \frac{\pi}{2}\right)\right)$

Problem3. Prove that $\Delta\,(\log U_x) = \log\left(1+\frac{\Delta U_x}{U_x}\right)$

Solution. $\Delta\,(\log U_x) = \log U_{x+h} - \log U_x = \log\left(\frac{\log U_{x+h}}{\log U_x}\right) = \log\left(\frac{EU_x}{U_x}\right)$

$$= \log\left[\frac{(1+\Delta)U_x}{U_x}\right] = \log\left[\frac{U_x+\Delta U_x}{U_x}\right]$$

$$= \log\left(1+\frac{\Delta U_x}{U_x}\right)$$

Problem 4. Evaluate. $\frac{\Delta^2 x^3}{Ex^2}$ taking h=1

Solution. $\Delta x^3 = (x+1)^3 - x^3 = 3x^2 + 3x + 1$

$$\Delta^2 x^3 = \Delta(\Delta x^3) = \Delta(3x^2 + 3x + 1)$$

$$= 3\Delta x^2 + 3\Delta x + \Delta(1)$$

$$= 3[(x+1)^2 - x^2] + 3[(x+1) - x] + 0$$

$$= 6(x+1)$$

Now $Ex^2 = (x+1)^2$

$$\therefore \frac{\Delta^2 x^3}{Ex^2} = \frac{6(x+1)}{(x+1)^2} = \frac{6}{x+1}$$

Problem 5. Evaluate $\Delta^3[(1-ax)(1-bx)(1-cx)]$

Solution. Let $U_x = (1-ax)(1-bx)(1-cx)$

The polynomial whose third order difference is to be calculated is a third degree polynomial and the coefficient of x^3 term is – abc

Since, $\Delta^r U_x = \begin{cases} 0 & \text{if } r > 3 \\ \text{constant} & \text{if } r = 3 \end{cases}$ we have

$$\Delta^3 U_x = \Delta^3(-abc\,x^3) = -abc\,\Delta^3(x^3) = -abc\,3!$$

$$= -6\,abc$$

Problem 6. If $U_0 = 1, U_1 = 5, U_2 = 8, U_3 = 3, U_4 = 7, U_5 = 0$ find $\Delta^5 U_0$

Solution. Consider $\Delta^5 U_0 = (E-1)U_0$

$$= (E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)U_0$$

$$= E^5U_0 - 5E^4U_0 + 10E^3U_0 - 10E^2U_0 + 5EU_0 - 1$$

$$= U_5 - 5U_4 + 10U_3 - 10U_2 + 5U_1 - U_0$$

$$= 0-(5\times7)+(10\times3)-(10\times8)+(5\times5)-1$$

$$= -35+30-80+25-1$$

$$= -61$$

NOTES

Alter. This can be obtained by forming the difference tale as shown below.

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$	$\Delta^5 U_x$
0	**1**					
		4				
1	5		**-1**			
				-7		
		3			**24**	
2	8		-8			**-61**
				17		
		-5			-37	
3	3		9			
				-20		
		4				
4	7		-11			
		-7				
5	0					

Hence $\Delta^5 U_0 = -61$

Problem 7. Estimate the missing term in the following table.

X	0	1	2	3	4
U_x	1	3	9	-	81

Explain why the resulting value differs from 3^3

Solution. Let the missing term in U_x be a

The difference table is given below

x	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$
0	**1**				
		2			
1	3		**4**		
				a-19	
		6			
2	9		a-15		**124-4a**
		a-9		105-3a	
3	a		90-2a		
		81-a			
4	81				

Since 4 values of U_x are given it is a polynomial of degree 3. Hence by fundamental theorem of finite differences $\Delta^4 U_x = 0$ for all x.

In particular $\Delta^4 U_0 = 0$. Hence 124-4a=0

$\therefore$ a=31

Exercises.

1.Find the first order differences for the following functions taking the interval of differencing as 1.

(i) e^x (ii) 2^x (iii) $\tan^{-1} x$ (iv) x(x-1)3^x (v) $\frac{x+1}{(x-1)(x-2)}$

(vi) $\frac{x-4}{(2-x)(3-x)}$

2.Prove that Δ tan cx = $\frac{\sin c}{\cos cx \cos[c(x+1)]}$; c being constant.

3. Prove that (i) $\Delta(U_x.V_x)$ = $U_x\Delta V_x + V_{x+h}\Delta U_x$

(ii) $\Delta\left(\frac{U_x}{V_x}\right) = \frac{V_x\Delta U_x - U_x\Delta V_x}{V_xV_{x+h}}$

NOTES

4.If U_x=$2x^3 - x^2 + 3x + 1$ find Δ^2U_x taking the interval of differencing as unity.

5. Show that Δ^3U_x=$6ah^3$ where If U_x= $ax^3 + bx^2 + cx + d$ and h is the interval of differencing.

6. Evaluate (i) Δ^2x^3 taking h as interval of differencing.

(ii) $\Delta^4[(1-2x)(1-3x)(1-5x)(1-6x)]$taking h=1.

(iii) $\Delta^n[(1-ax)(1-bx^2)(1-cx^3)(1-dx^4)]$ taking h=1.

7.Form the difference table for the following data
(i)

x	1	2	3	4	5
U_x	2	7	8	11	8

10.3 NEWTON,S FORMULA

Consider the function y =f(x). Let $f(x_0) = y_0, f(x_1) = y_1,$

$f(x_n) = y_n$. The replacement of f(x) by a simple function ϕ(x) which also assumes the values $y_0, y_1, \dots., y_n$ at the points $x_0, x_1, \dots, x_n$is the basic principle of interpolation. ϕ(x) is called a **formula for interpolation** and we say that ϕ(x) represents f(x). If ϕ(x) is a polynomial of degree n then ϕ(x) is called an **interpolating polynomial.** The existence of an interpolating polynomial is supposed by Weierstrass; approximation theorem which asserts that every continuous function on a closed interval can be approximated by a polynomial.

We now describe Newton –Gregory formula for constructing a n interpolating polynomial.

Theorem 10.3 (Newton's –Gregory Interpolating formula for equal Intervals)

Let $U_a, U_{a+h}, \dots., U_{a+nh}$ be the values of the function U_x at the points a, a+h, a+2h, ...,a+nh, which are of equal interval of difference.

Then $U_x = U_a + (x-a)\frac{\Delta U_a}{1!h} + (x-a)(x-a-h)\frac{\Delta^2 U_a}{2!h^2} + \cdots$

$$+(x\text{-}a)(x\text{-}a\text{-}h)\ldots.(x\text{-}a\text{-}\overline{n-1}h)\frac{\Delta^n U_a}{n!h^n}$$

NOTES

Proof. Let ϕ(x) be an interpolating polynomial of degree n which represents $U_x$in a≤x≤a+nh. Then ϕ(x) may be written in the form

ϕ(x) =$A_0 + A_1(x-a) + A_2(x-a)(x-a-h) +$

$$A_3(x-a)(x-a-h)(x-a-2h)+\ldots..$$

$$+A_n(x-a)(x-a-h)\ldots(x-a-\overline{n-1}h)\ \ldots\ldots\ldots..(1)$$

By definition of interpolating polynomial we have $U_a = \phi(a)$

Putting x=a, in(1) we get $A_0 = U_a$

Putting x=a+h in (1) we get $A_1 = \frac{1}{h}[U_{a+h} - A_0]$

$$=\frac{1}{h}[U_{a+h} - U_a]$$

$$\therefore A_1 = \frac{\Delta U_a}{h}$$

Putting x =a+2h in (1) we get

$$A_2 = \frac{1}{2h^2}[U_{a+2h} - A_0 - 2hA_1]$$

$$= \frac{1}{2h^2}[U_{a+2h} - U_a - 2\Delta U_a]$$

$$= \frac{1}{2h^2}[U_{a+2h} - 2U_{a+h} - U_a]$$

$$= \frac{\Delta^2 U_a}{2!h^2} \text{ (refer example in 2.1)}$$

Similarly substituting x=a+3h, ..., a+$\overline{n-1}$ h we get

$$A_3 = \frac{\Delta^3 U_a}{3!h^3}, \ldots\ldots\ldots., A_n = \frac{\Delta^n U_a}{n!h^n}$$

Substituting these values in (1) we get

ϕ(x)= $U_a + (x-a)\frac{\Delta U_a}{1!h} + (x-a)(x-a-h)\frac{\Delta^2 U_a}{2!h^2} + \cdots$

$$+(x\text{-}a)(x\text{-}a\text{-}h)\ldots.(x\text{-}a\text{-}\overline{n-1}h)\frac{\Delta^n U_a}{n!h^n}$$

Since ϕ(x) is the interpolating polynomial which represents U_x the function ϕ(x) can be written as U_x

$$\therefore U_x = U_a + (x-a)\frac{\Delta U_a}{1!h} + (x-a)(x-a-h)\frac{\Delta^2 U_a}{2!h^2} + \cdots$$

$$+(x-a)(x-a-h)....(x-a-\overline{n-1}h)\frac{\Delta^n U_a}{n!h^n}$$

NOTES

The above formula is known as **Newton –Gregory formula for forward interpolation.**

Corollary 1. If we take $\frac{x-a}{h}$ =r so that x =a+rh then the Newton –Gregory formula for forward interpolation reduces to

$$\phi(a+rh)= U_a + \frac{r}{1!}\Delta U_a + \frac{r(r-1)}{2!}\Delta^2 U_a + \cdots + \frac{r(r-1)......(r-\overline{n-1}h)}{n!}\Delta^n U_a$$

Note. Since $\phi(x)$ is the interpolating polynomial which represents U_x the function $\phi(a+rh)$ can be written as U_{a+rh}. Hence Newton Gregory formula for forward interpolation becomes,

$$U_{a+rh}=U_a + \frac{r}{1!}\Delta U_a + \frac{r(r-1)}{2!}\Delta^2 U_a + \cdots + \frac{r(r-1)......(r-\overline{n-1}h)}{n!}\Delta^n U_a$$

Corollary 2. The formula given in cor 1. Can also be used to **extrapolate** in the interval (a-h, a)

Newton's Gregory formula for forward interpolation cannot be used for interpolating a value of U_x near the end of the given data. To get a formula for this purpose we can write the interpolating polynomial $\phi(x)$ of degree n which represent $U_x$in a+nh≤x≤a as

$$\phi(x)=A_0 + A_1(x-\overline{a+nh}) + A_2(x-\overline{a+nh})(x-\overline{a+(n-1)h})+....+$$

$$A_n(x-\overline{a+nh})(x-\overline{a+(n-1)h}).....(x-\overline{a+h})$$

As in the proof of theorem 10.3 we can find

$$A_n = \frac{\nabla'' U_{a+nh}}{n!h^n};n=0,1,2, ..., n$$

Thus, $U_x = U_{a+nh} + \frac{\nabla U_{a+nh}}{1!h}\left(x-\overline{a+nh}\right) + \frac{\nabla^2 U_{a+nh}}{2!h^2}\left(x-\overline{a+nh}\right)(x-$ a+(n−1)h+...+∇nUa+nhn!hnx−a+nh...x−a+h

This is known as **Newton's formula for backward interpolation.**

Taking $\frac{x-\overline{a+nh}}{h}$-r we get x=a+nh+rh

Further using ,

h=(a+nh)=[a+(n-1)h], 2h=(a+nh)-[a+n-2)h] etc.

The above equation can be written as

$$U_{a+nh+rh} = U_{a+nh} + \frac{r\nabla U_{a+nh}}{1!} + \frac{r(r+1)\nabla^2 U_{a+nh}}{2!} + \frac{r(r+1)(r+2)\nabla^3 U_{a+nh}}{3!} + \cdots + \frac{r(r+1)\ldots(r+\overline{n-1}\nabla^n U_{a+nh}}{n!}$$

NOTES

Solved Problems.

Problem 1. If $U_{75} = 246; U_{80} = 202; U_{85} = 118$ and $U_{90} = 40$

Find $U_{79.}$

Solution.

Here a=75; h=5 We have to find $U_{a+rh} = U_{79}$

∴ a+rh=79 Hence 75+5r=79

∴ r=4/5=0.8

By Newton – Gregory formula for equal intervals,

$$U_{a+rh} = U_a + \frac{r}{1!}\Delta U_a + \frac{r(r-1)}{2!}\Delta^2 U_a + \cdots$$

We require ΔU_a, $\Delta^2 U_a$,...............

Hence we from the difference table as given below

$$\therefore U_{79} = 246 + \frac{0.8(-44)}{1} + \frac{08(0.8-1)}{1\ .\ 2}(-40) + \frac{0.8(0.8-1)(0.8-2)}{1\ .2.3}(46)$$

= 246-35.2+3.2+1.472

= 215.472

NOTES

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$
75	246			
		-44		
80	202		-40	
		-84		46
85	118		6	
		-78		
90	40			

Problem 2. By using Gregory-Newton's formula find U_x for the following data. Hence estimate (i) $U_{1.5}$ (ii) U_9

U_0	U_1	U_2	U_3	U_4
1	11	21	28	29

Solution. Let us from the difference table first.

X	U_x	ΔU_x	$\Delta^2 U_x$	$\Delta^3 U_x$	$\Delta^4 U_x$
0	**1**				
		10			
1	11		**0**		
				-3	
		10			**0**
2	21		-3		
				-3	
		7			
3	28		-6		
		1			
4	29				

Here the third order differences are constant and hence the required function is a polynomial of degree 3. In this case we use the formula.

$$U_x = U_a + (x-a)\frac{\Delta U_a}{1!\,h} + (x-a)(x-a-h)\frac{\Delta^2 U_a}{2!\,h^2} + \cdots$$

Here a=0 and h=1

$$\therefore U_x = 1 + (x-0) \times \frac{10}{1!} + x(x-1) \times \frac{0}{2!} + x(x-1)(x-2)x\frac{(-3)}{3!}$$

$$= 1+10x-\frac{x(x-1)(x-2)}{2}$$

$$= \frac{1}{2}(2 + 20x - x^3 + 3x^2 - 2x)$$

$$\therefore U_x = \frac{1}{2}(-x^3 + 3x^2 + 18x + 2)$$

NOTES

(i) $\therefore U_{1.5} = \frac{1}{2}(-(1.5)^3 + 3(1.5)^2 + 18(1.5) + 2)$

$= \frac{1}{2}(-3.375 + 6.75 + 27 + 2)$

$= 16.188$

(ii) $U_9 = \frac{1}{2}(-(9)^3 + 3(9)^2 + 18(9) + 2)$

$= \frac{1}{2}(-729 + 243 + 162 + 2) = -161$

Problem 3. Population was recorded as follows in village.

Year	1941	1951	1961	1971	1981	1991
Population	2500	2800	3200	3700	4350	5225

Estimate the population for the year 1945.

Solution.

Year x	**Populatio nU$_x$**	$\boldsymbol{\Delta U_x}$	$\boldsymbol{\Delta^2 U_x}$	$\boldsymbol{\Delta^3 U_x}$	$\boldsymbol{\Delta^4 U_x}$	$\boldsymbol{\Delta^5 U_x}$
1941	**2500**					
		300				
1951	2800					
			100			
		400				
1961	3200			**0**		
			100		**50**	
		500				
1971	3700			50		**-25**
			150		25	
		650				
1981	4350			75		
			225			
		875				
1991	5225					

We have to find $U_{1945.}$Here a=1941 and h=10

$$\therefore U_{a+rh} = U_{1945}$$

Hence 1941+10r=1945. Hence r=0.4

Applying Newton- Gregory formula we get

$U_{1945} = 2500 + 0.4 \times \frac{300}{1} + 0.4(0.4-1) \times \frac{100}{2!} + 0.4(0.4-1)(0.4-2) \times \frac{0}{3!} + 0.4(0.4-1)(0.4-2)(0.4-3) \times \frac{50}{4!} + 0.4(0.4-1)(0.4-2)(0.4-3)(0.4-4) \times \frac{25}{5!}$

NOTES

= 2500+120-12-2.08+0.75

= 2606.67

Exercises.

1.If $U_{75} = 2459; U_{80} = 2018;\ U_{85} = 1180$ and $U_{90} = 402$. Find U_{79}

2. Estimate the expectation of life at the age of 16 years from the following data.

Age in Years	10	15	20	25	30	35
Expectation of life in Years	35.4	32.3	29.2	26	32.2	20.4

3.If $\log_{10}$ 5 =0.6990; $\log_{10}$ 10 =1; $\log_{10}$ 15 =1.161; and

$\log_{10}$ 20 =1.3010 find $\log_{10}$ 12

4.Given that sin$45^o = 0.7071$; sin$50^o = 0.7660$;

Sin $55^o = 0.8192$; sin$60^o = 0.8660$. Find sin52^o

5.From the following table find U_{15}

x	0	6	12	18
U_x	23.1234	23.7234	24.6834	26.1330

10.4 LAGRANGE'S FORMULA

Newton-Gregory formula can be used for interpolation only when we know the values of U_x at points in equal intervals. The following formula due to the French Mathematician Lagrange can be used when we know the values of U_x at points which are not at equal intervals. This formula also enables us to determine the form of the function U_x.

Theorem 10.4 (Lagrange's theorem) Let $U_{a_1}, U_{a_2, \ldots\ldots} U_{a_n}$ be the values of U_x at $a_1, a_2, \ldots, a_n$ (not necessarily at equal intervals) then an interpolating Polynomial ϕ(x) for U_x is given by

$$\phi(x)=\frac{(x-a_2)(x-a_3)\ldots\ldots(x-a_n)}{(a_1-a_2)(a_1-a_3)\ldots(a_1-a_n)}\times U_{a_1}+\frac{(x-a_1)(x-a_3)\ldots\ldots(x-a_n)}{(a_2-a_1)(a_2-a_3)\ldots(a_2-a_n)}\times U_{a_2}+\ldots\ldots+$$

$$\frac{(x-a_1)(x-a_2)\ldots\ldots(x-a_{n-1})}{(a_n-a_1)(a_n-a_2)\ldots(a_n-a_{n-1})}\times U_{a_n}$$

Proof. Since n values for U_x are given we can assume ϕ(x) to be a polynomial of degree n-1 Let

$$\phi(x)=A_1(x-a_2)(x-a_3)\ldots\ldots(x-a_n)+A_2(x-a_1)(x-a_3)\ldots(x-a_n)$$

$$+\ldots.+A_n(x-a_1)(x-a_2)\ldots\ldots(x-a_{n-1})\ldots\ldots\ldots\ldots.(1)$$

When $x=a_1$, we get $U_{a_1}=A_1(a_1-a_2)(a_1-a_3)\ldots(a_1-a_n)$

$$\therefore A_1=\frac{U_{a_1}}{(a_1-a_2)(a_1-a_3)\ldots(a_1-a_n)}$$

Similarly when x = $a_2, a_3 \ldots, a_n$ we get

$$A_2=\frac{U_{a_2}}{(a_2-a_1)(a_2-a_3)\ldots(a_2-a_n)}$$

....

....

$$A_n=\frac{U_{a_n}}{(a_n-a_1)(a_n-a_2)\ldots(a_n-a_{n-1})}$$

Substituting these values in (1) we get Lagrange's formula.

Note. Since ϕ(x) is the interpolating polynomial which represents U_x the function ϕ(x) can be replaced by U_x in (1)

Hence Lagrange's formula becomes

$$U_x=\frac{(x-a_2)(x-a_3)\ldots\ldots(x-a_n)}{(a_1-a_2)(a_1-a_3)\ldots(a_1-a_n)}\times U_{a_1}+\frac{(x-a_1)(x-a_3)\ldots\ldots(x-a_n)}{(a_2-a_1)(a_2-a_3)\ldots(a_2-a_n)}\times U_{a_2}+\ldots\ldots+$$

$$\frac{(x-a_1)(x-a_2)\ldots\ldots(x-a_{n-1})}{(a_n-a_1)(a_n-a_2)\ldots(a_n-a_{n-1})}\times U_{a_n}$$

NOTES

Solved Problems.

Problem 1. Find U_5 given that $U_1 = 4; U_2 = 7;$ and $U_4 = 13; U_7 = 30$

Solution. The arguments 1, 2, 4, 7, are not at equal intervals and we use Lagrange's formula to find U_5 Take $a_1 = 1; a_2 = 2; a_3 = 4; a_4 = 7$ and x=5. Substituting in Lagrange's formula we get

$$U_5 = \left[\frac{(5-2)(5-4)(5-7)}{(1-2)(1-4)(1-7)}\right]\times4+\left[\frac{(5-1)(5-4)(5-7)}{(2-1)(2-4)(2-7)}\right]\times7\left[\frac{(5-1)(5-2)(5-7)}{(4-1)(4-2)(4-7)}\right]\times13$$

$$+\left[\frac{(5-1)(5-2)(5-4)}{(7-1)(7-2)(7-4)}\right]\times30$$

$$=\left[\frac{3\times1\times(-2)}{(-1)(-3)(-6)}\right]\times4+\left[\frac{4\times1\times(-2)}{1(-2)(-5}\right]\times7+\left[\frac{4\times3(-2)}{3\times2(-3)}\right]\times13+\left[\frac{4\times3\times1}{6\times5\times3}\right]\times30$$

$$=\frac{4}{3}-\frac{28}{5}+\frac{52}{3}+4 = 17.06$$

Problem 2. Find the form of the function U_x for the following data. Hence find U_3

x	0	1	2	5
U_x	2	3	12	147

Solution. Here $a_1 = 0; a_2 = 1; a_3 = 2; a_4 = 5$

$$\therefore U_{a_1} = 2; U_{a_2} = 1; U_{a_3} = 12; U_{a_4} = 147$$

Applying Lagrange's formula we get

$$U_x = \left[\frac{(x-1)(x-2)(x-5)}{(0-1)(0-2)(0-5)}\right]\times2\left[\frac{(x-0)(x-2)(x-5)}{(1-0)(1-2)(1-5)}\right]\times1+$$

$$\left[\frac{(x-0)(x-1)(x-5)}{(2-0)(2-1)(2-5)}\right]\times12+\left[\frac{(X-0)(X-1)(X-2)}{(5-0)(5-1)(5-2)}\right]\times147$$

$$=-\frac{(X^3-8X^2+17X-10)}{5}+\frac{3(X^3-7X^2+10X)}{4}-2(x^3-6x^2+5x)+$$

$$\frac{x^3-3x^2+2x}{60}\times147$$

$$=\frac{1}{60}[x^3(-12+45-120+147)+x^2(96-315+720-$$
$$441+x-204+450-600+294+120$$

$$=\frac{1}{60}[60x^3+60x^2-60x+120]$$

$$\therefore U_x = x^3+x^2-x+2$$

$$\therefore U_3 = 3^3+3^2-3+2$$

$$= 35$$

NOTES

Problem 3. Determine by Lagrange's formula the percentage number of criminals inder 35 years

Age	% number of criminals
Under 25 years	52.0
Under 30 years	67.3
Under 40 years	84.1
Under 50 years	94.4

Solution. We have to find U_{35}

Here $\mathrm{a}_1 = 25; \mathrm{a}_2 = 30; \mathrm{a}_3 = 40; \mathrm{a}_4 = 50$

$\therefore\ \mathrm{U}_{\mathrm{a}_1} = 52; \mathrm{U}_{\mathrm{a}_2} = 67.3; \mathrm{U}_{\mathrm{a}_3} = 84.1;\ \mathrm{U}_{\mathrm{a}_4} = 94.4$

Applying Lagrange's formula we get

$$\mathrm{U}_{35} = \left[\frac{(35-30)(35-40)(35-50)}{(25-30)(25-40)(25-50)}\right]\times 52.0\left[\frac{(35-25)(35-40)(35-50)}{(30-25)(30-40)(30-50)}\right]\times 67.3 +$$

$$\left[\frac{(35-25)(35-30)(35-50)}{(45-25)(40-30)(40-50)}\right]\times 84.1 + \left[\frac{(35-25)(35-30)(35-40)}{(50-25)(50-30)(50-40)}\right]\times 94.4$$

$$= \frac{(-1)\times 52.0}{5} + \frac{3\times 67.3}{4} + \frac{1\times 84.1}{2} - \frac{1\times 94.4}{20}$$

= -10.40+50.38+42.05-4.72

= 77.31

Hence the estimated % number of criminals under 35 years is 77.31

Problem 4. Prove that Lagrange's formula can be put in the form

$$\mathrm{U}_x = \sum_{i=1}^{n} \frac{\phi(x)\mathrm{U}_{\mathrm{a}_i}}{(x-\mathrm{a}_i)\phi(\mathrm{a}_i)} \text{ where } \phi'(x)=\frac{d\phi(x)}{dx} \text{ and}$$

$$\phi(x)= \prod_{i=1}^{n}(x - \mathrm{a}_i) = (x - \mathrm{a}_1)(x - \mathrm{a}_2) \ldots\ldots\ldots .(x - \mathrm{a}_n)$$

Solution. The Lagrange's formula is given by

$$\mathrm{U}_x=\frac{(x-\mathrm{a}_2)(x-\mathrm{a}_3).....(x-\mathrm{a}_n)}{(\mathrm{a}_1-\mathrm{a}_2)(\mathrm{a}_1-\mathrm{a}_3)...(\mathrm{a}_1-\mathrm{a}_n)}\times\mathrm{U}_{\mathrm{a}_1}+\frac{(x-\mathrm{a}_1)(x-\mathrm{a}_3).....(x-\mathrm{a}_n)}{(\mathrm{a}_2-\mathrm{a}_1)(\mathrm{a}_2-\mathrm{a}_3)...(\mathrm{a}_2-\mathrm{a}_n)}\times\mathrm{U}_{\mathrm{a}_2}+.....+$$

$$\frac{(x-\mathrm{a}_1)(x-\mathrm{a}_2).....(x-\mathrm{a}_{n-1})}{(\mathrm{a}_n-\mathrm{a}_1)(\mathrm{a}_n-\mathrm{a}_2)...(\mathrm{a}_n-\mathrm{a}_{n-1})}\times\mathrm{U}_{\mathrm{a}_n} \qquad ....(1)$$

NOTES

Now $\phi(x) = (x-a_1)(x-a_2)\ldots\ldots(x-a_n)$ ....(2)

$\therefore \log \phi(x) = \sum_{i=1}^{n} \log(x-a_i)$

Differentiating w.r.t. x we get

Differentiating w.r.t. x we get $\frac{\phi'(x)}{\phi(x)} = \sum_{i=1}^{n} \frac{1}{(x-a_i)}$

$$\therefore \phi'(x) = \sum_{i=1}^{n} \frac{\phi(x)}{x-a_i} = \sum_{i=1}^{n} \frac{(x-a_1)(x-a_2)\ldots..(x-a_n)}{(x-a_i)}$$

$$= (x-a_2)(x-a_3)\ldots\ldots(x-a_n) + (x-a_1)(x-a_3)\ldots\ldots(x-a_n)$$

$$+\ldots\ldots+(x-a_1)(x-a_2)\ldots\ldots(x-a_{n-1})$$

$\therefore \phi'(a_1) = (a_1-a_2)(a_1-a_3)\ldots(a_1-a_n)$

$\phi'(a_2) = (a_2-a_1)(a_2-a_3)\ldots(a_2-a_n)$

....

....

$\phi'(a_n) = (a_n-a_1)(a_n-a_2)\ldots(a_n-a_{n-1})$

Substituting these values in (1) and using (2) we get

$$U_x = \frac{\phi(x)U_{a_1}}{(x-a_1)\phi'(a_1)} + \frac{\phi(x)U_{a_2}}{(x-a_2)\phi'(a_2)} + \ldots\ldots + \frac{\phi(x)U_{a_n}}{(x-a_n)\phi'(a_n)}$$

$$= \sum_{i=1}^{n} \frac{\phi(x)U_{a_i}}{(x-a_i)\phi'(a_i)}$$

Exercises.

1.The following table gives the normal weight of a baby during first 6 months of life.

Age in month	0	2	3	5	6
Weight in lbs	5	7	8	10	12

Find the estimated weight of baby at the age of 4 months.

2. Apply Lagrange's formula to find U_2 from the data $U_0 = 2$; $U_1 = 5$; $U_3 = 29$; $U_7 = -19$

NOTES

3.The following table gives the premium payable at age in years completed. Interpolate the premium payable at age 35 completed.

Age completed	25	30	40	60
Premium in Rs.	50	55	70	95

4. Given $\log_{10} 654 = 2.8156;$

$\log_{10} 658 = 2.8182;$

$\log_{10} 659 = 2.8189;$

$\log_{10} 661 = 2.8202;$

Find $\log_{10} 656$

UNIT-11. THEORY OF ATTRIBUTES

11.1 INTRODUCTION

Statistics chiefly deals with collection of data, classification of data based on certain characteristics, calculation of statistical constants such as mean, median, mode, standards deviation etc., and analysis of data based on the statistical constants. The characteristics used for classification of data may be quantitative or qualitative or qualitative in nature. For example, when we consider the set of students in a class, their heights and weights are characteristics which are quantitative whereas their efficiency, intelligence, health and social status are characteristics which are qualitative. The qualitative characteristics of a population are called attributes and they cannot be measured by numeric quantities. Hence the statistical treatment required for attributes is different from that of quantitative characteristics. In this chapter we develop the statistical techniques used in the theory of attributes.

11.2 ATTRIBUTES

Suppose the population is divided into two classes according to the presence or absence of a single attribute. The positive class denotes the presence of the attributes and the negative class denotes the absence the attribute. Capital Roman letter such A,B,C.........are used to denote positive classes and the corresponding lower case Greek letters such as $\alpha,\beta,\gamma,\delta$.... are used to denote negative classes. For example if A represents the attribute richness , then α represents the attributes non-richness(poor).

The combinations of attributes are denoted by grouping together the letters concerned.

For example, if attributes A represents health and B represents wealth then AB represents the possession of both health and wealth; $A\beta$ represents health and non-wealth; αB represents non-health and non-wealth. A convenient way of representing two attributes in a 2× 2 tables is as follows.

Attributes	**B**	**β**
A	AB	$A\beta$
α	αB	$\alpha\beta$

A class represented by n attributes is called a class of n^{th} order.

For example, A,B,C, α,β,γ are all of first order, AB, A β, αB, $\alpha\ \beta$ are of second order, and ABC , A $\beta\gamma$, AβC, $\alpha\beta\gamma$ are of the third order.

The number of individuals possessing the attributes in a class of n^{th} order is called a class frequency of order n and class frequency are denoted by bracketing the attributes.

Thus(A) stands for the frequency of A the number of individuals possessing the attributes A and (A β)stands for the number of individuals possessing the attributes A and not B.

Note 1. Class frequencies of the type (a),(AB),(ABC)...... are known as positive class frequencies.

class frequencies of the type (α),(β),($\alpha\ \beta$),($\alpha\beta\gamma$).... are known as negative class frequencies.

Class frequencies of the type (αB),(A β),(A $\beta\gamma$),($\alpha\beta$C).... are known as contrary frequencies.

Note 2. If N is the total number of observations in a population (i.e., N is the total frequency) without any specification of attributes the N is considered to be a frequency of order zero.

The frequency classes for two attributes can be represented in the form of a table as shown below.

Attributes	**B**	**β**	**Total**
A	(AB)	(αB)	(A)
α	(Aβ)	($\alpha\beta$)	(α)
Total	**(B)**	**(β)**	**N**

N denotes the total number in the population.

In the population of size N , the relation between the class frequencies of various orders are given below.

$$N = (A) + (\propto) = (B) + (\beta) = (C) + (\gamma)\text{etc.,}$$

$$\left.\begin{aligned}(A) &= (AB) + (A\ \beta)\\ (B) &= (AB) + (\propto B)\end{aligned}\right\}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$\left.\begin{array}{l}(\propto) = (\propto B) + (\propto\beta)\\(\beta) = (A\,\beta) + (\propto\beta)\end{array}\right\}\text{.......................................(2)}$$

$$\left.\begin{array}{l}N=(A) + (\propto)\\N=(B) +(\beta)\end{array}\right\} > N= (AB)+(A\,\beta)+ (\propto B)+ (\propto\beta)$$

from (1) and (2).

For three attributes A,B and C we get similar results as shown below.

$(A) = (ABC) +(AB\gamma) + (A\,\beta C)+(A\,\beta\,\gamma)$

$(B) = (ABC) +(AB\gamma)+(\propto BC) + (\propto\beta\,\gamma)$

$(C) = (ABC) +(A\,\beta C) + (\propto BC) +(\propto\beta C)$

$(AB)= (ABC) + (AB\gamma)$

$(A\,\beta) = (A\beta C) + (A\beta\,\gamma)$

$(\propto B) = (\propto BC) +(\propto B\gamma)$

$(\propto\beta) = (\propto\beta C) +(\propto\beta\,\gamma)$etc.,

Note. Any class frequency can be expressed in terms of frequencies of higher order.

The following table gives the class frequencies of all orders and the total number of all class frequencies upto 3 attributes.

Order	Attributes	Class frequencies of all orders	Number in each order	Total number
0		**N**		**1**
	A			
0		N	1	3
1		(A),(∝)	2	
	A,B			
0		N	1	
1		(A),(B),(∝),(β)	4	9
2		(AB),(A β),(∝B);(∝β)	4	
	A,B,C			
0		N	1	
1		(A),(B),(C),(∝),(β),(γ)	6	
2				27
		(AB),(A β),(∝B),(∝β),	12	
3		(AC),(A γ),(∝C),(∝ γ)		
		(BC),(B γ),(βC),(β γ)		
		(ABC),(AB γ),(AβC),		
		(Aβγ),(∝B γ),(∝BC)	8	
		(∝βγ)		

The classes of height order are called the ultimate classes and their frequencies are called the ultimate class frequencies.

NOTES

Theorem 11 .1 Given n attributes,

(i) total number of class frequencies is 3^n

(ii)total number of positive class frequencies is 2^n

(iii) total number of negative class frequencies is 2^n-1

Proof.(i) The number of ways of choosing r attributes from the given set of n attributes is $\binom{n}{r}$.Since each attribute gives two symbols (one for positive and one for negative), the number of class frequencies of order r that can be obtained from r attributes is 2^r.

Hence the total number of class frequencies of order r is $\binom{n}{r}2^r$.

Thus the total number of frequencies (of all orders)

$$= \sum_{r=0}^{n} \binom{n}{r}2^r = 1+\binom{n}{1}2 +\binom{n}{1}2^2 +........+ \binom{n}{n}2^n$$

$$= (1+2)^n = 3^n$$

(ii) Any collection of r attributes gives rise to only one positive class frequency of order r (all possessing attributes only).

Hence total number of positive class frequencies (of all orders)

$$= \sum_{r=0}^{n} \binom{n}{r} = 1+\binom{n}{1} +\binom{n}{2} +........+ \binom{n}{n}$$

$$= (1+1)^n =2^n$$

(iii)There is no negative class frequency of order 0.Any collection of r attributes gives rise to one negative class frequency of order r (all non-possessing attributes).

Hence total number of negative class frequencies(of all orders)

$$= \sum_{r=1}^{n} \binom{n}{r}=2^n -1$$

Dichotomisation is the process of dividing a collection of objects into tow classes according to the possession or non-possession of an attribute.

Suppose a population consists of N objects. If A is an attribute we have N = (A) + (∝).

$$=> N = A.N + \propto.N$$

$$=(A + \propto).N$$

$$\Rightarrow 1 = A + \propto$$

Thus in symbolic expression A can be replaced by $1 - \propto$ and $\propto$ by $1 - A$. This concept is useful to express any class frequency in terms of higher order class frequencies and ultimately in terms of ultimate class frequencies examples 1, and 2 below). Also it is useful to express positive class frequencies (negative class frequencies) in terms of negative class frequencies (positive class frequencies)(refer examples 3,4, and 5 below).

Examples

1. $(AB) = (ABC) + (AB\gamma)$

Consider $(AB\gamma) = AB\gamma.N = AB(1 - C).N$

$$= AB\,.\,N - ABC\,.\,N$$

$$=(AB) - (ABC)$$

$$\therefore (AB) = (ABC) + (AB\gamma)$$

2. If there are two attributes A and B we have

$$N = (A) +(\propto) =(B) + (\beta)$$

Hence $N = (A) +(\propto) = (AB) +(A\beta) + (\propto B) + (\propto\beta)$

and $N = (B) + (\beta) = (AB) + (\propto B) + (A\beta) + (\propto\beta)$

If there are three attributes A,B,C we have

$$N = (A) + (\propto)$$

$\Rightarrow N = (AB) + (A\beta)+ (\propto B) + (\propto\beta)$. Thus

$$N = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) +(\propto BC) + (\propto B\gamma) + (\propto\beta C) + (\propto\beta\gamma).$$

3.Consider two attributes A and B.

Now $(\propto\beta) =\propto\beta\,.\,N =(1 -A)(1-B).N$

$$=(1-A-B+AB)$$

$$N = N-A.N-B.N+AB.N$$

$$= N-(A) - (B) + (AB)$$

Here negative class frequency has been expressed in terms of positive class frequencies.

4. $(AB) = AB.N$

$$(1 - \propto)(1-\beta).N = (1-\propto-\beta-\propto\beta).N$$

$$= N - \propto.N - \beta.N + \propto\beta.N = N - (\propto) - (\beta) + (\propto\beta)$$

Here positive class frequency has been expressed in terms of negative class frequencies.

5. $(\propto\beta\gamma) = \propto\beta\gamma.N = (1 - A)(1 - B)(1 - C).N$

$$= N - A.N - B.N - C.N + AB.N + AC.N + BC.N - ABC.N$$

$$= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$$

Note. $N = (A) + (B) + (C) - (AB) - (BC) - (AC) + (ABC) + (\propto\beta\gamma)$

Solved problems.

Problem1. Given $(A) = 30$; $(B) = 25$; $(\propto) = 30$; $(\propto\beta) = 20$

find (i) N (ii) (β) (iii) (AB) (iv) $(A\beta)$ (v) $(\propto\beta)$

Solution. (i) $N = (A) + (\propto) = 30 + 30 = 60$

(ii) $(\beta) = N - (B) = 60 - 25 = 35$

(iii) $(AB) = AB.N = (1 - \propto)(1-\beta).N$

$$= N - (\propto) - (\beta) + (\propto\beta)$$

$$= 60 - 30 - 35 + 20 = 15$$

(iv) $(A\beta) = A\beta.N = A(1-B).N = (A) - (AB)$

$$= 30 - 15 = 15$$

(v) $(\propto B) = \propto B . N = (1-A)B.N = (B) - (AB)$

$$= 25 - 15 = 10$$

Note. The result can also be got directly by completing the 2 × 2 contingency table for the attributes A and B

	(B)	(β)	
(A)	-	-	30
(∝)	-	20	30
	25	-	-

Problem 2. Given the following ultimate class frequency of two attributes A and B. Find the frequencies of positive and negative class frequencies and the total number of observations.

$$(AB) = 975\ ;\ (\propto\beta) = 100; (A\beta)=25;\ (\propto\beta) = 950$$

Solution. Positive class frequencies are (A) and (B)

$$(A) = (AB) + (A\,\beta) = 975 + 25 = 1000$$

$$(B) = (AB) + (\propto B) = 975 + 100 = 1075$$

Negative class frequencies are $(\propto)$ and (β)

$$(\propto) = (\propto B) + (\propto\beta) = 100 + 950 = 1050$$

$$(\beta) = (A\beta) + (\propto\beta) = 25 + 950 = 975$$

$$N = (A) + (\propto) = (B) + (\beta)$$

Taking $N = (A) + (\propto) = 1000 + 1050 = 2050$

Note. The results can also be got directly by completing the 2 × 2 contingency table for the attributes A and B.

Problem 3. Given the following positive class frequencies. Find the remaining class frequencies $N = 20;\ (A) = 9;\ (B)=12; (C) = 8; (AB)=6; (BC)=4; (CA) = 4; (ABC) = 3$

Solution. There are three attributes A,B,C

The total number of class frequencies is $3^3 = 27$

We are given only 8 class frequencies and we have to find the remaining 19 class frequencies. They are

Order 1. $(\propto) = N - (A) = 20-9 = 11$

$$(\beta) = N - (B) = 20-12 = 8$$

$$(\gamma) = N - (C) = 20-8 = 12$$

Order 2. $(A\beta) = A(1 - B).N = (A) - (AB) = 9 - 6 = 3$

$$(\propto B) = A(1 - B)B.N = (B) - (AB) = 12 - 6 = 6$$

$$(\propto B) = A(1 - B)B.N = (B) - (AB) = 12 - 6 = 6$$

$$(A\gamma) = A(1 - C).N = (A) - (AC) = 9 - 4 = 5$$

$$(\propto C) = (1 - A)C = (C) - (AC) = 8 - 4 = 4$$

$(B\gamma) = B(1 - C).N \quad = (B) - (BC) = 12 - 4 \ = 8$

$(\beta C) = (1 - B)C.N = (C) - (BC) = 8 - 4 = 4$

$(\propto\beta) = (1 - A)(1 - B).N = N - (A) - (B) + (AB)$

$= 20 - 9 - 12 + 6 = 5$

$(\beta\gamma) = (1 - B)(1 - C).N = N - (B) - (C) + (BC)$

$= 20 - 12 - 8 + 4 = 4$

$(\propto\gamma) = (1 - A)(A - C).N \quad = N - (A) - (C) + (AC)$

$= 20 - 9 - 8 + 4 = 7$

NOTES

Order 3. $(AB\gamma) = AB(1 - C).N \quad = (AB) - (ABC) = 6 - 3 = 3$

$(A\beta C) = A(1 - B)C.N = (AC) - (ABC) = 4 - 3 = 1$

$(AB\gamma) = A(1 - B)(1 - C).N = (A) - (AC) - (AB) + (ABC)$

$= 9 - 4 - 6 + 3 = 2$

$(\propto BC) = (1 - A)BC.N = (BC) - (ABC)$

$= 4 - 3 = 1$

$(\propto B\gamma) = (1 - A)(1 - C)\ B.N = (B) - (BC) - (AB) + (ABC)$

$= 12 - 4 - 6 + 3 = 5.$

$(\propto\beta C) = (1 - A)(1 - B)C.N = (C) - (AC) - (BC) + (ABC)$

$= 8 - 4 - 4 + 3 = 3.$

$(\propto\beta\gamma) = (1 - A)(1 - B)(1 - C).N = N - (A) - (B) - (C) + (AB) + (BC) + (CA) - (ABC)$

$= 20 - 9 - 12 - 8 + 6 + 4 + 4 - 3 = 2$

percentage of tube lights which pass the four tests

$= \frac{4615}{5000} \times 100$

$= 92.3\ \%$

Exercise

1. Given the frequencies (A) =1150; (∝)=1120; (AB) = 1075; (∝β)=985. Find the remaining class frequencies and the total number of observations.

2. Given the following ultimate class frequencies. Find the frequencies of the positive and negative classes and the total number of observations.

(AB)=733; (Aβ)=840; (∝B)=699; (∝β) = 783

3.Given the following data. Find the frequencies of (i) the remaining positive classes (ii) ultimate classes

N = 1800; (A)=850; (B) = 780; (C) = 326

(ABγ) = 200 ; (AβC)=94; (∝BC)=72 (ABC) = 50

4. Given The Following Ultimate Class frequencies. Find the frequencies of the positive classes.

(ABC) =298; (AβC) = 450; (∝BC) = 408;(∝βC) = 342

(ABγ) = 1476; (Aβγ) = 2292; (∝Bγ) = 3524 ;(∝βγ) = 43684

11.3 CONSISTENCY OF DATA

Consider a population with the attributes A and B . For the data observed in the same population (AB) cannot be greater than (A). Thus the figures(A) = 20 and (AB) = 25 are inconsistent. We observe that for the above figures,(Aβ) = (A) - (AB) = -5, which is negative. This motivated the following definition.

Definition: A set of class frequencies is said to be consistent if none of them is negative. Otherwise the given set of class frequencies is said to be inconsistent.

Since any class frequency can be expressed as the sum of the class frequencies , it follows that a set of independent class frequency a consistent if and only if no ultimate class frequency is negative.

We have the following set of criteria for testing the consistency in the set of single attribute, two attributes and three attributes.

Attributes	Condition of consistency	Equivalent positive class conditions	Number of conditions
A	$(A)\geq 0$ $(\propto)\geq 0$	$(A)\geq 0$ $(A)\leq N$(since $(\propto)$ = $(1-A).N\geq 0$	2
A,B	$(AB)\geq 0$ $(A\beta)\geq 0$ $(\propto B)\geq 0$ $(\propto\beta)\geq 0$	$(AB)\geq 0$ $(AB)\leq (A)$ $(AB)\leq (B)$ $(AB)\geq (A)+(B)-N$	2^2
A,B,C	$(ABC)\geq 0$ $(AB\gamma)\geq 0$ $(A\beta C)\geq 0$ $(\propto BC)\geq 0$ $(A\,\beta\gamma)\geq 0$ $(\propto B\gamma)\geq 0$ $(\propto\beta C)\geq 0$ $(\propto\beta\gamma)\geq 0$	(i) $(ABC)\geq 0$ (ii) $(ABC)\leq (AB)$ (iii)$(ABC)\leq (AC)$ (iv)$(\propto BC)\leq (BC)$ (v)$(ABC)\geq (AB)+(AC)+(A)$ (vi)$(ABC)\geq (AC)+(BC)-(C)$ (vii)$(ABC)\leq (AB)+(BC)+(AC)-(A)-(B)-(C)+(N)$	2^3

NOTES

Note 1. In the case of 3 attributes conditions.

(i) and (viii) => $(AB)+(BC)+(AC)\geq (A)+(B)+(C)-N$ ....(ix)similarly,

(ii) and (vii) => $(AC)+(BC)-(AB)\leq (C)$ ....(X)

(iii) and (vi) => $(AB)+(BC)-(AC)\leq (B)$...................(xi)

(iv) and (v) => $(AB)+(AC)-(BC)\leq (A)$ ....(X)

Conditions (ix) to (xii) can be used to check the consistency of the when the class frequency of first and second order alone are know.

NOTES

Note 2. If the given data are incomplete so that it is not possible to determined all the class frequencies then the conditions of consistency can be used determine the limits in which an unknown class frequency can lie.

Solved problem

Problem 1. Find whether the following data are consistent

$$N = 600; (A) = 300; (B)=400; (AB) = 50$$

Solution. We calculate the ultimate class frequencies ($\propto\beta$),($\propto$B) and Aβ)

$$(\propto\beta) = \propto\beta.N = (1 - A)(1\text{-}B).N = N - (A) - (B) + (AB)$$

$$=600 - 300 -400 + 50 = -50$$

Since $(\propto\beta) < 0$, the data are inconsistent

Problem 2. Show that there is some error in the following data : 50% people are wealthy ,35 % are wealthy but not healthy,20 % people healthy but not wealthy.

Solution. Taking 'wealth' as A and 'health' as B we get the following data $N = 100; (AB) = 50; (A\beta) = 35; (\propto B) = 20$

To check the consistency of data we find $(\propto\beta)$

$$(\propto\beta) = \propto\beta.N = (1 - A)(1\text{-}B).N$$

$$= N - (A) - (B) + (AB)$$

$$\text{But } (A) = (AB) + (A\beta) = 50 + 35 = 85$$

$$(B) = (AB) +(\propto B) = 50 + 20 = 70$$

$$\therefore (\propto\beta) = 100 - 85 - 70 + 50 = -5$$

$$\therefore (\propto\beta)< 0$$

Hence there is error in the data.

Problem 3. Of 2000 people consulted 1854 speak Tamil; 1507 speak Hindi;572 speak English;676 speak Tamil and Hindi; 286 speak Tamil and English;270 speak Hindi and English;114 speak Tamil, Hindi and English. Show that the information as it stands is incorrect.

Solution.

Let A C denote the attributes of speaking Tamil, Hindi, English respectively.

NOTES

$\therefore$ Give = 2000; (A) = 1854;(B) = 1507; (C) = 572;

(AB) = 676; (AC) = 286; (BC)= 270; (ABC) = 114

Consider $(\propto\beta\gamma) = \propto\beta\gamma.N$

$=(1 - A)(1 - B)(1 - C).N$

$= \quad N - (A) - (B) - (C) - (AB) + (BC) + (AC) - (ABC)$

$= 2000 - 1854 - 1507 - 572 + 676 + 270 + 286 -114$

$= - 815$

$\therefore (\propto\beta\gamma) < 0$. Hence the data are inconsistent

$\therefore$ The information is incorrect

Problem 4. Find the limits of (BC) for the following available data;

N = 125; (A) = 48; (B) = 62; (C)=45; (Aβ) 7 and (Aγ) = 18

Solution. First of all we find (AB) and (AC)

$(AB) = (A) - (A\beta) = 48 - 7 = 41$

$(AC) = (A) - (A\gamma) = 48 - 18 = 30$

Now, by condition of consistency (ix)

$(AB) + (BC) + (AC) \geq (A) +(B) +(C) - N$

$=> 41 + (BC) + 30 \geq 48 + 62 + 45 - 125$

$\therefore (BC) \geq - 41$ ....(i)

Also using (xii) , $(AB) + (AC) - (BC) \leq (A)$.

$=> (BC) \geq (AB) + (AC) - (A) = 41 + 30 - 48 = 23$

$\therefore$ (BC) ≥ 23 ....(ii)

Using (xi),$(AB) +(BC) - (AC) \leq (B)$

$=> (BC) \leq (B) +(AC) - (AB) = 62 + 30 - 41 = 51$

$\therefore$ (BC) $\leq$ 51 ....(iii)

Using (X),(AC) + (BC) - (AB) $\leq$ (C)

=> (BC) $\leq$(C) + (AB) - (AC) = 45 + 41 - 30 = 56

$\therefore$ (BC) $\leq$ 56 ....(iv)

From (i),(ii),(iii) and (iv) we find 23 $\leq$(BC) $\leq$ 56

Problem 5. Find the greatest and least values of (ABC) (A) = 50;

(B) = 60;

(C) = 80;(AB) = 35; (AC) = 45 and (BC) = 42

solution:

The problem involves 3 attributes and we are given positive class frequencies of first order and second order only.

Using positive class conditions (ii),(iii),(iv) of consistency for 3 attributes

$$\left.\begin{array}{l}(ABC) \leq (AB) \Rightarrow (ABC) \leq 35\\ (ABC) \leq (BC) \Rightarrow (ABC) \leq 42\\ (ABC) \leq (AC) \Rightarrow (ABC) \leq 45\end{array}\right\} \Rightarrow (ABC) \Rightarrow 35....(1)$$

Using (v),(vi) and (vii)

(ABC) $\geq$ (AB) + (AC) - (A)=> (ABC) $\geq$ 35 + 45 - 50 = 30

(ABC) $\geq$ (AB) + (BC) - (B)=> (ABC) $\geq$35 + 42 - 60 = 17

(ABC) $\geq$ (AC) + (BC) - (C)=> (ABC) $\geq$45 + 42 - 80 = 7

$$\text{Thus}\quad \left.\begin{array}{l}(ABC) \geq 30\\ (ABC) \geq 17\\ (ABC) \geq 7\end{array}\right\} \Rightarrow (ABC) \geq 30.....................(2)$$

From (1) and (2) we get 30 $\leq$ (ABC) $\leq$ 35

$\therefore$ The least value of (ABC) is 30 and the greatest value of (ABC) is 31

Exercise

1. Examine the consistency of data when

(i) (A) = 800, (B) = 700,(AB) = 660, N = 1000

(ii) (A) = 600, (B) = 500,(AB) = 50, N = 1000

(iii) N = 2100 (A) = 1000, (B) = 1300,(AB) = 1100

(iv) N = 100 (A) = 45, (B) = 55,(C) = 50(AB) = 15, (BC) = 25, (AC) = 20,

(ABC)= 12

(iv) N = 1800 (A) = 850, (B) = 780,(C) = 326(AB) =250, (BC) = 122,

(AC) =144, (ABC)= 50

2. If (A) = 50, (B) = 60, (C) = 50, (Aβ) = 5, (Aγ) = 20 and N = 100, find the least and greatest values of (BC)

3. Given that (A) = (B) = (C) = $\frac{N}{2}$ =50; (AB) = 30,(AC) = 25, Find the limits which (BC) will lie.

4. If N = 120, (A) = 60,(B) = 90, (C) = 30, (BC) = 15,(AC) = 15,

find the limits between which (AB)must lie.

11.4 INDEPENDENCE AND ASSOCIATION OF DATA

Two attributes A and B are said to be independent if there is same proportion of A's amongst B's as a amongst β's. Or equivalently the proportion of B's amongst A's is the same as amongst the ∝'s.

Thus A and B are independent if

$\frac{(AB)}{(B)} = \frac{(A\beta)}{\beta}$ ....(i) (or) $\frac{(AB)}{(A)} = \frac{(\propto B)}{\propto}$ ....(ii)

From (i) we get

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$$

(AB) = $\frac{(A)(B)}{N}$ ... ... (1) and (Aβ) = $\frac{(A)(\beta)}{N}$ ....(2)

Again from (i) we get 1 - $\frac{(AB)}{(B)}$ = 1 - $\frac{(A\beta)}{(\beta)}$

$$\therefore \frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

$$\therefore \frac{(\propto B)}{(B)} = \frac{(\propto\beta)}{(\beta)}$$

$$\therefore \frac{(\propto B)}{(B)} = \frac{(\propto\beta)}{(\beta)} = \frac{(\propto\beta)+(\propto B)}{(\beta)+(B)} = \frac{(\propto)}{N}$$

NOTES

$$\therefore (\propto \beta) = \frac{(\propto)(\beta)}{N} \text{ ....(3)}$$

and $(\propto B) = \frac{(\propto)(B)}{N}$ ....(4)

(1),(2),(3),(4) are all equivalent conditions for independence of the attributes A and B.

Note. In terms of second order class frequencies we get the condition of independence as (AB) ($\propto\beta$) = ($A\beta$)($\propto B$)

For if A and B are independent attributes we have

$$(AB)\ (\propto\beta) = \left[\frac{(A)(B)}{N}\right].\left[\frac{(\propto)(\beta)}{N}\right] = \left[\frac{(A)(\beta)}{N}\right]\left[\frac{(\beta)(B)}{N}\right] = (A\beta)(\ \propto B)$$

Thus A and B are independent if (AB) ($\propto\beta$) - ($A\beta$)($\propto B$) = 0 ....(5)

Association and coefficient of association

If (AB) = $\frac{(A)(B)}{N}$ we say that A and B are associated. There are two possibilities. If(AB) > $\frac{(A)(B)}{N}$ we say that A and B are positively associated and if(AB) < $\frac{(A)(B)}{N}$ we say that A and B negatively associated.

Notation. Let us denote δ = (AB) - $\frac{(A)(B)}{N}$

Thus δ = (AB) - $\frac{(A)(B)}{N}$ = $\frac{1}{N}$ [N(AB) - (A)(B)]

=$\frac{1}{N}$ [{(AB) + ($A\beta$) + ($\propto B$) + ($\propto\beta$)} (AB) - {(AB) + ($A\beta$)} + ($\propto B$)}]

=$\frac{1}{N}$ [(AB)($\propto\beta$) - ($A\beta$)($\propto B$)......................................(6)

Note 1. We Know That A and B are independent if δ = 0

$\therefore$ A and B are independent if Q = Y = 0

Note 2. If A and B are perfectly associated then (AB) = (A) hence ($A\beta$)= 0 or (AB) = (B) hence ($\propto B$) = 0. In either Q = 1 = Y note 3. If A and B are perfectly disassociated then either (AB) = 0 or ($\propto\beta$) = 0 and in this case Q = -1 = Y

Thus we get -1 ≤ Q ≤-1

Note. Yule's coefficient Q and the coefficient of colligation Y is related by the relation Q = $\frac{2Y}{1+ Y^2}$

Proof. Let x =$\frac{(A\beta)(\propto B)}{(AB)(\propto\beta)}$. Hence Y = $\frac{1-\sqrt{x}}{1+\sqrt{x}}$

$$\therefore Y^2 = \frac{(1-\sqrt{X})^2}{(1+\sqrt{X})^2}$$

$$\therefore 1+Y^2 = 1+\frac{1+X-2\sqrt{X}}{(1+\sqrt{X})^2}=\frac{2(1+X)}{(1+\sqrt{X})^2}$$

$$\therefore \frac{2Y}{1+Y^2} = \frac{2\left(\frac{1-\sqrt{X}}{1+\sqrt{X}}\right)}{\frac{2(1+X)}{(1+\sqrt{X})^2}.}=\frac{(1-\sqrt{X})(1+\sqrt{X})}{1+x}=\frac{1-x}{1+x.}$$

$$=\frac{1-\frac{(A\beta)(\propto B)}{(AB)(\propto\beta)}}{1+\frac{(A\beta)(\propto B)}{(AB)(\propto\beta)}}=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$=Q$$

From the above relationship between Y and Q we infer the following

Q=0=> Y=0; Q=-1 => Y=-1 and

Q=1 => Y=1 and conversely.

Solved Problems.

Problem 1. Check whether the attributes A and B are independent given that (i) (A)=30, (B)=60, (AB)=12, N=150

(ii) (AB)=256, (αB) = 768, (Aβ)=48, ($\alpha\beta$)=144

Solution. Since the given class frequencies are of first order the condition for independent is (AB)=$\frac{(A)(B)}{N}$

Consider $\frac{(A)(B)}{N} = \frac{30\times60}{150} = 12 = (AB)$

$\therefore$(AB)=$\frac{(A)(B)}{N}$. Hence A and B are independent.

(ii) (A)=(AB)+(Aβ)=256+48=304

NOTES

(B)= (AB)+(αβ)=256+768=1024

(α)=(αβ)+(αβ)=768+144=912

(β)=(A)+(α)=304+912 =1216

Now, $\frac{(A)(B)}{N} = \frac{304\times 1024}{1216} = 256 = (AB)$

(AB)= $\frac{(A)(B)}{N}$. Hence A and B are independent.

Aliter. Applying the condition (5) for independence,

(AB)(αβ)- (Aβ)(αB)=256×144-768×48=36864-36864=0

∴A and B are independent.

Note. By providing Q=10 also we can conclude A an B are independent.

Problem 2. In a class test in which 135 candidates were examined for profficiency in Physics and Chemistry, it was discovered that 75 students failed in Physics, 90 failed in Chemistry and 50 failed in both. Find the magnititude of association and state if there is any association between failing in Physics and Chemistry.

Solution. Denoting 'fail in Physics' as A and 'fail in Chemistry' as B we get

(A)=75, (B)=90, (AB)=50, N=135

The magnititude of association is measured by

Q=$\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

we now get the ultimate class frequencies.

(α)=N-(A)=135-75=60

(β)=N-(B)=135-90=45

(αβ)=(B)-(AB)=90-50=40

(Aβ)=(A)-(AB)=75-50=25

(αβ)=α-(αβ)=60-40=20

Q=$\frac{50\times 20-25\times 40}{50\times 20+25\times 40}$=0

∴A and B are independent. Hence failure in Physics and Chemistry are completely independent of each other.

NOTES

Problem 3. Show whether A and B are independent or positively asociated or negatively associated in the following cases.

(i) N=930, (A)=300, (B)=400, (AB)=230

(ii) (AB)=327, (Aβ)=545, (αβ)=741, (αβ)=235

(iii)(A)=470, (AB)=300, (α)=530, (αβ)=150

(iv) (AB)=66, (Aβ)=88, (αB)=102, (αβ)=136

Solution. (i) $\frac{(A)(B)}{N} = \frac{300\times400}{930} = 129.03$

Now, δ=(AB)- $\frac{(A)(B)}{N} = 230 - 129.03 = 100.97$

Hence $\delta>0$ Hence A and B are positively associated.

(ii) $Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}=\frac{327\times235-545\times741}{327\times235+545\times741}$

$= \frac{76845-403845}{76845+403845} = \frac{-32700}{480690} = -0.6803$

$\therefore Q<0$. Hence A and B are negatively associated.

(iii) N=(A)+(α)=470+530=1000

(B)=(AB)+(αB)=300+150=450

Now,$\frac{(A)(B)}{N} = \frac{470\times450}{1000} = 2115$

$\therefore\delta$=(AB)- $\frac{(A)(B)}{N}$=300-2115=-1815

$\therefore\delta<0$. Hence A and B are negatively associated.

(iv) $Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)} = \frac{66\times136-88\times102}{66\times136+88\times102} = 0$

$\therefore$A and B are independent

Problem 4. Calculate the coefficient of association between intelligence of father and son from the following data.

Intelligent fathers with intelligent sons 200

Intelligent fathers with dull sons 50

Dull fathers with intelligent sons 110

Dull fathers with dull sons 600

Comment on the result.

Solution. Denoting the 'intelligence of fathers' by A and 'intelligence of sons' by B we have.

(AB)=200; (Aβ)=50; (αB)=110; (αβ)=600

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}=\frac{200\times600-50\times110}{200\times600+50\times110}$$

$$=0.91235$$

Since Q is positive it means that intelligent fathers are likely to have intelligent sons.

Problem 5. Investigate from the following data between inoculation against small pox and preventation from attack.

	Attacked	**Not attacked**	**Total**
Inoculated	25	220	245
Not inoculated	90	160	250
Total	115	380	495

Solution. Denoting A as 'inoculated' and B as 'attacked' we have

(AB)=25; (Aβ)=220; (αB)=90; (αβ)=160

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}=\frac{25\times160-220\times90}{25\times160+220\times90}$$

$$=\frac{400-19800}{400+19800}=\frac{-15800}{23800}=-0.6638$$

This shows that the attributes A and B have negative association.

(ie., 'innoculation' and ' attack from small pox' are negatively associated.

Thus innoculation against small pox can be taken as the preventive measure.

Exercises. Show whether A and B are independent , positively associated or negatively associated.

1.Show whether A and B are indepentent, positiviely associated or negatively associated.

(i) N=392, (A)=154, (B)=168, (AB)=66

(ii) N=1000, (A)=470, (B)=620, (AB)=320

(iii) (A)=28, (B)=38, (AB)=12, N=60

(iv) (AB)=90; (Aβ)=65; (αB)=260; (αβ)=110

(v)(A)=245, (α)=285, (AB)=147, (αβ)=190

2. In an examination in Tamil and English 245 of the candidates passed in Tamil, 147 passed in both, 285 failed in Tamil and 190 failed in Tamil but passed in English. Howfar is the knowledge in the two subjects associated?

3. Calculate Yule's coefficient of association between marriage and failure of students

	Passed	**Failed**	**Total**
Married	90	65	155
Unmarried	260	110	370
Total	350	175	525

UNIT XII- INDEX NUMBERS

NOTES

12.1 INDEX NUMBERS

An Index number is a widely used statistical device for comparing the level of a certain phenomenon with the level of the same phenomenon at some standard period. For example , we may wish to compare the price of a food article at a particular period with the price of the same article at a previous period of time. This Compar4ision can be expressed as the percentage of ration of the prices in the two periods and this number serves as a single food-price index number. The comparison of prices of several food articles at two different periods is usually expressed as a suitable weighted average of the percentage changes in these prices. In the calculation of averages, various standard measures of central tendencies such as Arithmetic mean , Geometric Mean, Harmonic mean can be used.

In the computation of an index number, if the base year used for comparison is kept constant throughtout, then it is called **fixed base method.** If on the other hand, for every year the previous year is used as a base for comparison, then the method is called **chain base method**.

Index numbers can be broadly classified into two types.

(i)**Unweighted or simple index number.**

(ii)Weighted index number.

Two standard methods of computation are

(A) Aggregate method.
(B) Average of price relatives method

I-A Aggregate Method.

In this method total of current year prices for various commodities is divided by the total of the base year. In symbols, if p_0 denotes the price of the base year and p_1 the price of the current year

$p_{01} = \frac{\sum p1}{\sum p0} \times 100$, where $\sum p1$ is total of the current year and

$\sum p0$ is the total of the base year.

This is the simplest method inwhich aggregate of the prices for base year and current year alone are taken into consideration.

Example 1. From the following data construct the simple aggregative index number for 1992

Commodities	Price in 1991 Rs.	Price in 1992 Rs.
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
Flour	4.5	5.25

NOTES

Solution. Construction of price index taking 1991 as base year.

Commodities	Price in 1991 Rs.	Price in 1992 Rs.
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	85
Flour	4.5	5.25
Total	**133.0**	**147.00**

$\therefore$ Aggregate index number $p_{01}=\frac{\sum p1}{\sum p0}\times 100$

$$=\frac{147}{133}\times 100$$

$$= 110.5$$

I-B Average of Price Relatives Method (Simple index numbers).

Price relatives denoting the price of a commodity of a base year as p_0 and the price of te current year as p_1 the ratio of the prices $\frac{p_1}{p_0}$ is called the **Price relatives.**

Index number for the current year is $p_{01} = \frac{p_1}{p_0} \times 100$

In the average of price relatives method the average of price relatives for various item is calculated by using any one of the measures of central tendencies such as arithmetic mean, geometric mean, harmonic

NOTES

mean , etc.. Arithmetic and Geometric means are the very common average used in this method.

(i)The Arithmetic mean index number $p_{01} = \frac{\Sigma\left(\frac{p_1}{p_0}\right)\times 100}{n}$

(ii)The Geometric mean index number $p_{01} = \left[\prod\left(\frac{p_1}{p_0}\right)\right]^{1/n} \times 100$

where$\prod$ denotes the product.

Hence $\log p_{01}=\frac{\Sigma\left(\log\frac{p_1}{p_0}\times 100\right)}{n}$

Example. For the Example 1, we find the index number of the price relatives taking 1991 as the base year using (i) Arithmetic mean (ii) Geometric mean

Commodities	Price in 1991p_0	Price in 1992 p_1	$\frac{p_1}{p_0}\times 100$	$\log\left(\frac{p_1}{p_0}\times 100\right)$
Rice	7	8	114.3	2.0580
Wheat	3.5	3.75	107.1	2.0298
Oil	40	45	112.5	2.0512
Gas	78	85	109.0	2.0374
Flour	4.5	5.25	116.7	2.0671
Total			**559.6**	**10.2435**

(i) Using arithmetic mean the index number$p_{01}=\frac{559.6}{5} = 111.92$

(ii)Using geometric mean the index number

$$\log p_{01} = \frac{10.2435}{5} = 2.0487$$

$\therefore p_{01} = \text{antilog }(2.0487) = 111.87$

Solved Problems

Problem 1. From the following data of the whole sale price of rice for the 5 years construct the index numbers taking (i) 1987 as the base (ii) 1990 as the base:

Years	1987	1988	1989	1990	1991	1992
Price of rice per kg.	5.00	6.00	6.50	7.00	7.50	8.00

Solution. (i) Construction of index numbers taking 1987 as base

Years	**Price of rice** **per Kg.**	**Index numbers (base 1987)**
1987	5.00	$\frac{6}{5} \times 100 = 120$
1988	6.00	$\frac{6.5}{5} \times 100 = 130$
1989	6.50	$\frac{6.5}{5} \times 100 = 140$
1990	7.00	$\frac{7}{5} \times 100 = 150$
1991	7.50	$\frac{7.5}{5} \times 100 = 150$
1992	8.00	$\frac{8}{5} \times 100 = 160$

From the index number table we observe that from 1987 to 1988 these is a increase of 20% in the price of rice per Kg for 1987 to 1989 there is a increase of 30% in the price of rice per Kg. Etc.,

(ii)Construction of index numbers taking 1990 as base

Years	Price of rice per Kg.	Index number (Base 1990)
1987	5	$\frac{5}{7} \times 100 = 71.4$
1988	6	$\frac{6}{7} \times 100 = 85.7$
1989	6.50	$\frac{6.5}{7} \times 100 = 92.9$
1990	7	$\frac{7}{7} \times 100 = 100$
1991	7.5	$\frac{7.5}{7} \times 10 = 107.1$
1992	8	$\frac{8}{7} \times 100 = 114.3$

NOTES

Problem 2. Construct the whole sale price index number for 1991 and 1992 from the data given below using 1990 as the base year.

Commodity	Whole sale prices in Rupees per quintal		
	1990	1991	1992
Rice	700	750	825
Wheat	540	575	600
Ragi	300	325	310
Cholam	250	280	295
Flour	320	330	335
Ravai	325	350	360

Solution. Taking 1990 as base year

Commodity	1990 p_0	1991 p_1	1992 p_1	Relatives for 91	Relatives for 92
Rice	700	750	825	$\frac{750}{700} \times 100$ $= 107.1$	$\frac{825}{700} \times 100$ $= 117.9$
Wheat	540	575	600	$\frac{575}{540} \times 100$ $= 106.5$	$\frac{600}{540} \times 100$ $= 111.1$
Ragi	300	325	310	$\frac{325}{300} \times 100$ $= 108.3$	$\frac{310}{300} \times 100$ $= 103.3$
Cholam	250	280	295	$\frac{280}{250} \times 100$ $= 112$	$\frac{295}{250} \times 100$ $= 118$
Flour	320	330	335	$\frac{330}{320} \times 100$ $= 103.1$	$\frac{325}{320} \times 100$ $= 101.6$
Ravai	325	350	360	$\frac{350}{325} \times 100$ $= 107.7$	$\frac{360}{325} \times 100$ $= 110.8$
Total	**644.7**	**662.7**			
Index number (using A.M.)	107.5	110.5			

Index numbers for 1991 as base year 1990 is 107.5

Index number for 1992 as base year 1990 is 110.5

Exercises.

1. From the following data construct the aggregate index number for 1991 taking 1990 as the base:

Commodities	Price in 1990 Rs.	Price in 1991 Rs.
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

2. For the data given below calculate the index numbers taking (i)1984 as base year (ii)1991 as base year

Year	1984	1985	1986	1987	1989	1990	1991	1992
Price of Wheat per Kg.	4	5	6	7	8	10	9	10

II Weighted Index Numbers

All items in the calculation of unweighted index numbers(simple index numbers) are treated as of equal importance. But in actual practice we notice that some items command greater importance than others and as such need more weight in the calculation of index numbers.

Standard methods of computing weighted index number are:

II-A Weighted aggregative method.

II-B Weighted average of price relatives method.

II-A Weighted aggregative method. Though there are many formulae to calculate index number in this method we give below some standard formulae which are very often used.

(a) **Laspeyre's index number.** According to Laspeyre's method the prices of the commodities in the base year as well as the current year are known and they are weighted by the quantities used in the base year. Laspeyre's index number is defined to be

$$L_{Io1} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

NOTES

(b**) Paasche's index number.** According to Paasches' method current year quantities are taken as weights and hence Paasches' index number is defined

$$\mathbf{P_{Io1} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100}$$

(c) Marshall- Edgworth's index number. According to this method the weight is the sum of the quantities of the base period and current period. Hence **)** Marshall- Edgworth's formulae is defined by

$$\mathbf{M_{Io1} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100}$$

$$\mathbf{= \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100}$$

(d)**Bowley's Index number.** The arthemetic mean of Laspeyre's AND Paasche's index number defined to be Bowley's index number. Hence Bowley's Index number is given by

$$\mathbf{B_{Io1} = \frac{L_{Io1} + P_{Io1}}{2}}$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1}\right] \times \mathbf{100}$$

(e) Fisher's index number. Prof. Lrving Fisher, though suggested many index numbers, gives an 'ideal index number' as

$$\mathbf{I_{Io1}} = \sqrt{\left[\frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1}\right]} \times \mathbf{100}$$

we notice $\mathbf{I_{Io1}} = \sqrt{L_{Io1} \times P_{Io1}}$**.** That is Fisher's index number is the geometric mean of Laspeyre's index number and Paasche's index number.

(f)**Kelley's index number.** According to Kelley, Weight may be taken as the quantities of the period which is not necessarily the base year or current year. The average quantity of two or more years may be taken as the weight. Hence Kelley's index number can be defined as

$K_{Io1} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$ where q is the average quantity of two or more years. We notice that Marshall-Edgeworth index number is the same as Kelley's index number if the average quantity of two years is considered.

Example. Calculate (i) Laspeyre's (ii)) Paasche's (ii) Fisher's index numbers for the following data given below. Hence or otherwise find Edgeworth and Bowley's index numbers.

Commodities	Base Year 1990		Current Year 1992	
	Price	Quantity	Price	Quantity
A	2	10	3	12
B	5	16	6.5	11
C	3.5	18	4	16
D	7	21	9	25
E	3	11	3.5	20

Solution.

Commodities	1990		1992		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	p_0	q_0	p_1	q_1				
A	2	10	3	12	20	24	30	36
B	5	16	6.5	11	80	55	104	71.5
C	3.5	18	4	16	63	56	72	64
D	7	21	9	25	147	175	189	225
E	3	11	3.5	20	33	60	38.5	70
Total					**343**	**370**	**433.5**	**466.5**

(i)) Laspeyre's index number $= \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

$$= \frac{433.5}{343} \times 100 = 126.4$$

(ii) Paasche's index number $= \frac{\sum p_1q_1}{\sum p_0q_1} \times 100.$

$$= \frac{466.5}{370} \times 100$$

$$= 126.1$$

(iii) Fisher's index number$= \sqrt{\left[\frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1}\right]} \times 100$

$$= \sqrt{\frac{433.5 \times 466.5}{343 \times 370}} \times 100$$

$$=126.2$$

NOTES

(iv) Bowley's Index number $= \frac{L_{Io1}+P_{Io1}}{2}$

$$=\frac{126.4+126.1}{2} = 126.25$$

(v) Edgworth's index number = $\frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$

$$=\frac{433.5+466.5}{343+370} \times 100$$

$$= \frac{900}{713} \times 100$$

$$= 126.2$$

II-B Weighted Average of Price Relative Method

In this metod the index number is computed by taking the weighted Arithmetic mean of price relatives. Thus if P is the price relative and V is the value weights $p_0 q_0$ then the index number $p_{01} = \frac{\sum PV}{\sum V}$

Example.

Index number using weighted arithmetic mean of price relatives.

Commodity	Price in 1990 p_0	Price in 1992 p_1	Quantity in 1990 q_0	V $p_0 q_0$	P $\frac{p_1}{p_0} \times 100$	PV
Coconut oil	Rs. 50	Rs. 54	15 lit	750	108	81000
Ground nut oil	Rs. 45	Rs. 48	25 lit	1125	106.7	120037.5
Gingles oil	Rs.43	Rs.45	30 lit	1290	104.7	135063
Rice	Rs. 7	Rs. 9	350 kg	2450	128.6	315070
Total				**5615**	**-**	**651170.5**

$\therefore$ Weighted index number = $\frac{\sum PV}{\sum V} = \frac{651170.5}{5615} = 116$

Ideal index number. An index number is said to be ideal index number if it is subjected to the following three tests and found okeyed.

(i)**The time reversal test.**

(ii)The factor reversal tests

(iii)The commodity reversal tests.

NOTES

(i) **The time reversal test.** Let $I_{(01)}$ denote the indx number of the current year y_1 relative to the base year y_0 , without considering percentage, and $I_{(10)}$ denotes the index number of the base year y_0 relative to the current year y_1 without considering the percentage. If $I_{(01)} \times I_{(10)}$=, then we say that the indx number satisfies the time reversal test.

(ii) **The factor reversal tests.** In this test the prices and quantities are interchanged, without considering the percentage, satisfying the following relation $I_{(pq)} \times I_{(qp)} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$,where $I_{(pq)}$ is the price index of the current year relative to the base year and $I_{(q,p)}$ is the quantity index of the current year relative to the base year.

(iii) **The commodity reversal tests.** The index number should be independent of the order in which different commodities are considered. This test is satisfied by almost all index numbers.

Remark 1. Fisher's index number is an ideal index number.

We verify whether the Fisher's index number satisfies the threes tests for ideal number.

Fisher's index number is $I_{(01)} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

Time reversal test. Interchanging base year and current year

$$=I_{(10)} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}.$$

$$\text{Now, } I_{(01)} \times I_{(10)} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= 1$$

Factor reversal test. Denoting the fisher's index number I_{01} for the prices p and quantity q as

$$I_{(pq)} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

Interchanging the prices and quantities in $I_{(pq)}$ we get

$$I_{(qp)} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

NOTES

Now, $I_{(pq)} \times I_{(qp)} = \sqrt{\left(\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}\right)\left(\frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}\right)}.$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Obviously Fisher’s index number satisfies the commodity reversal test. Hence Fisher’s index number is an ideal index number.

Note. Of all the index numbers defined earlier Fisher’s index number is the only index number which is an ideal index number.

Remark 2. Laspeyer’s index number does not satisfy the time reversal test.] We have $p_{01}=\frac{\sum p_1 q_0}{\sum p_0 q_0}$and $p_{10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}$

$$\text{Now, } p_{01} \times p_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} = 1$$

Hence Laspeyre’s index number does not satisfy the time reversal test

Remark 3. Paasche’s index number does not satisfy the time reversal test. .(Verify)

Remark 4. Laspeyre’s and Paasche’s index number also do not satisfy factor reversal test. .(Verify)

Solved Problems.

Problem 1. Construct , with the help of data given below , Fisher’s index number and shown that it satisfies both the factor reversal test and time reversal test.

Commodity	**A**	**B**	**C**	**D**
Base year price in Rupees	5	6	4	3
Base Year quantity in Quintals	50	40	120	30
Current year price in Rupees	7	8	5	4
Current year quantity in Quintals	60	50	110	35

Solution.

Commodity	**Base year**		**Current year**		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	p_0	q_0	p_1	q_1				
A	5	50	7	60	250	300	350	420
B	6	40	8	50	240	300	320	400
C	4	120	5	110	480	440	600	550
D	3	30	4	35	90	105	120	140
Total					**1060**	**1145**	**1390**	**1510**

Fisher's index number is $I_{(01)} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$

$$= \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}} \times 100$$

Time reversal test.

Now, $I_{(01)} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}}$

$$I_{(10)} = \sqrt{\frac{\sum p_0q_1}{\sum p_1q_1} \times \frac{\sum p_0q_0}{\sum p_1q_0}} = \sqrt{\frac{1145}{1510} \times \frac{1060}{1390}}$$

Now, $I_{(01)} \times I_{(10)} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145} \times \frac{1145}{1510} \times \frac{1060}{1390}} = 1.$

Factor reversal test.

$I_{(pq)} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}}$

Interchanging the factors,

$I_{(qp)} = \sqrt{\frac{\sum p_0q_1}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_1q_0}} = \sqrt{\frac{1145}{1060} \times \frac{1510}{1390}}$

Now, $I_{(pq)} \times I_{(qp)} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145} \times \frac{1145}{1060} \times \frac{1510}{1390}}$

$$= \frac{1510}{1060} = \frac{\sum p_1q_1}{\sum p_0q_0}$$

Hence the factor reversal test is also satisfied.

NOTES

Exercises.

1.Construct index number of prices from the following data by applying (i) Laspeyre's Method (ii) Paasche's method (iii)Bowley's method (iv) Fisher's index method (v)Marshall-Edge worth method.

Commodities	**Base year**		**Current year**	
	Price	**Quantity**	**Price**	**Quantity**
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

2. Calculate Fisher's index number for 1992 for the following data.

Year	Rice		Wheat		Flour	
	Price	Quantities	Price	Quantities	Price	Quantities
1988	9.3	100	6.4	11	5.1	5
1992	4.5	90	3.7	10	2.7	3

3. For the data given below find the different weighted index numbers

Commodities	Base year		Current Year	
	Price	Quantities	Price	Quantities
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	26

NOTES

12.2CONSUMER PRICE INDEX NUMBERS (Cost of living index numbers)

It is the index number designed to measure changes in the living cost of various classes of people. No consumer price index will suti all classes of people because different classes of people differ widely from each other in the style of functioning, consumption habits, mode of expenditure etc., These index numbers are useful for wage negotiations and wage contracts. Dearness allowance is calculated based on the cost of living index numbers.

Formula to find the cost of living index numbers.

(i)Aggregate ecxpenditure method.

Cost of living index number $I_{01}=\frac{\sum p_1q_0}{\sum p_0q_0}\times 100$.(Laspeyre's method)

(ii) **Family budget method.**

Cost of living index $I_{01}=\frac{\sum PV}{\sum V}$

Where $P=\frac{p_1}{p_0}\times 100$ and V-value weight p_0q_0.

Also $I_{01}=\frac{\sum PW}{\sum W}$ where P= $P=\frac{p_1}{p_0}\times 100$ and W is the weight.

Solved Problems.

Problem 1. Find the cost of living index number for 1992 on the base of 1991 on the basis from the following data using(i)family budget method (ii)aggregate expenditure method.

Commodity	Price in Rs.		Quantity in Quintals in 1991
	1991	1992	
Rice	7	7.5	6
Wheat	6	6.75	3.5
Flour	5	5	0.5
Oil	30	32	3
Sugar	8	8.5	1

Solution. By family budget method.

Commodity	p_0	p_1	q_0	p_0q_0 V	$\frac{p_1}{p_0} \times 100$ P	PV
Rice	7	7.5	6	42	107.1	4498.2
Wheat	6	6.75	3.5	21	112.5	2362.5
Flour	5	5	0.5	2.5	100	250
Oil	30	32	3	90	106.7	9603
Sugar	8	8.2	1	8	106.3	850.4
Total				**163.5**	**-**	**17564.1**

NOTES

Cost of finding index = $\frac{\sum PV}{\sum V}$

$$= \frac{17564.1}{163.5}$$

$$= 107.4$$

(ii)By aggregate expenditure method

Commodity	p_0	p_1	q_0	p_0q_0	p_1q_0
Rice	7	7.5	6	42	45
Wheat	6	6.75	3.5	21	23.6
Flour	5	5	0.5	2.5	2.5
Oil	30	32	3	90	96
Sugar	8	8.2	1	8	8.5
Total				**163.5**	**175.6**

Cost of living index = $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

$$= \frac{175.6}{163.5} \times 100$$

$$= 107.4$$

Problem 2. An enquiry into the budgets of the middle class families in a city in India gave the following information.

NOTES

	Food	Rent	Clothing	Fuel	Misc
Weights	35%	15%	20%	10%	20%
Prices 1991	1500	300	450	70	500
Prices 1992	1650	325	500	90	550

What changes in cost of living index of 1992 as compared with that of 1991 are seen?

Solution. Base year is choosen as 1991(=100)

Item s	**Prices 1991**	**Prices 1992(I)**	**Index number 1992 p**	**W**	**PW**
Food	1500	1650	$\frac{1650}{1500} \times 10 = 110$	35	3850
Rent	300	325	$\frac{325}{300} \times 100$ =108.3	15	1624.5
Clot hing	450	500	$\frac{500}{450} \times 10 = 111.1$	20	2222.0
Fuel	70	90	$\frac{90}{70} \times 100 = 128.6$	10	1286
Misc.	500	550	$\frac{550}{500} \times 100 = 110$	20	2200
Total	**100**	**11182.5**			

Cost of living index numbers = $\frac{\sum PW}{\sum W}$

$$= \frac{11182.5}{100}$$

= 111.8

Hence the prices in 1992 compared with the prices in 1991 has risen to 11.8%

NOTES

Problem 3. Find the cost of living index for the following data in a middle class family.

Items	Price		Weight
	1991	1992	
Food	700	850	40
Clothing	300	280	15
Rent	200	225	7
Fuel	70	82	5
Medicine	100	135	9
Education	500	550	12
Entertainment	100	90	10
Misc.	475	425	23

Solution.

Items	Price		Index number 1992 P	W	PW
	1991	1992			
Food	700	850	$\frac{850}{700} \times 100 = 121.4$	40	4856
Clothing	300	280	$\frac{280}{300} \times 100 = 93.3$	15	1399.5
Rent	200	225	$\frac{225}{200} \times 100 = 112.1$	7	787.5
Fuel	70	82	$\frac{82}{70} \times 100 = 117.1$	5	585.5
Medicine	100	135	$\frac{135}{100} \times 100 = 135$	9	1215
Education	500	550	$\frac{550}{500} \times 100 = 110$	12	1320
Entertainment	100	90	$\frac{90}{100} \times 100 = 90$	10	900
Misc.	475	425	$\frac{425}{475} \times 100 = 89.5$	23	2058.5
Total				**121**	**13122**

Cost of living index number=$\frac{\sum PW}{\sum W} = \frac{13122}{121} = 108.4$

Exercises.

1. Calculate the index number of prices for 1992 on the bases of 1990 from the data given below.

Commodities	Weights	Price per unit 1990	Price per unit 1992
A	40	80	85
B	25	60	55
C	5	345	50
D	20	35	40
E	10	25	20

NOTES

2.The following are the group index numbers and the group weights of an average working class family's budget . Construct the cost of living index numbers by assigning the given weights.

Group	**Index number**	**Weight**
Food	352	48
Fuel & Electricity	220	10
Clothing	30	8
Rent	160	12
Miscellaneous	190	15

UNIT-X111. ANALYSIS OF TIME SERIES

NOTES

13.1 INTODUCTION

Economists and business are interested in studying the behaviour of sales, profits, national income , foreign exchange, industrial production, dividends of companies, share prices in share markets etc over a period of time. Such a study helps them in planning for the future. Analysis of time series deals with the methods of analysing such factors which vary with respect to time and deriving information regarding the likely future behaviour of those factors.

13.2 TIME SERIES

Definition. Time series is a series of values of a variable over a period of time arranged chronologically.

For example, the following table giving the price level of a commodity in different years forms a time series.

Year	1990	1991	1992	1993	1994	1995
Price Level	100	111	117	120	130	135

Mathematically a time series is a functional relationship y = f(t) where y is the value of the variable (phenomenon or factor) under consideration at a time t.

In general a time series is influenced by a large number of forces of different kinds . For examples, the time series of retail prices of rice is the result of combined influences of rain fall, availability of fertilisers, good yield, transport facilities, consumer' s demand and so on.

13.3 Uses of Time Series

- The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience
- It is helpful for business planning as it helps in comparing the actual current performance with the expected one
- From time series, we get to study the past behaviour of the phenomenon or the variable under consideration

- We can compare the changes in the values of different variables at different times or places, etc.

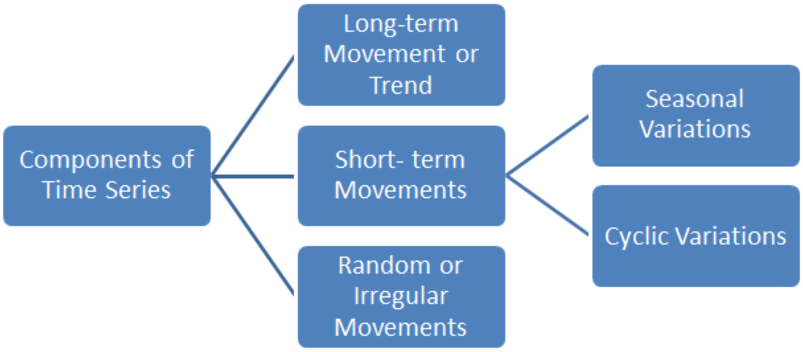
13.4 COMPONENTS OF A TIME SERIES

NOTES

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuation


Long-term Movement or Trend
Components of Time Series
Short- term Movements
Random or Irregular Movements
Seasonal Variations
Cyclic Variations

Trend

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Linear and Non-Linear Trend

If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear).

Periodic Fluctuations

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

Seasonal Variations

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.

The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Cyclic Variations

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the ‘Business Cycle’.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Random or Irregular Movements

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

UNIT-XIV MEASUREMENTS OF TRENDS

NOTES

14-1 MEASUREMENT OF TRENDS:

Secular trend is a long term movement in a time series. This component represents basic tendency of the series. The following methods are generally used to determine trend in any given time series. The following methods are generally used to determine trend in any given time series.

- Graphic method or eye inspection method
- Semi average method
- Method of moving average
- Method of least squares

Graphic method or eye inspection method

Graphic method is the simplest of all methods and easy to understand. The method is as follows. First plot the given time series data on a graph. Then a smooth free hand curve is drawn through the plotted points in such a way that it represents general tendency of the series. As the curve is drawn through eye inspection, this is also called as eye-inspection method. The graphic method removes the short term variations to show the basic tendency of the data. The trend line drawn through the graphic method can be extended further to predict or estimate values for the future time periods. As the method is subjective the prediction may not be reliable.

Advantages

It is very simplest method for study trend values and easy to draw trend.

Sometimes the trend line drawn by the statistician experienced in computing trend may be considered better than a trend line fitted by the use of a mathematical formula.

Although the free hand curves method is not recommended for beginners, it has considerable merits in the hands of experienced statisticians and widely used in applied situations.

Disadvantages:

This method is highly subjective and curve varies from person to person who draws it.

The work must be handled by skilled and experienced people.

Since the method is subjective, the prediction may not be reliable.

While drawing a trend line through this method a careful job has to be done.

Method of Semi Averages:

In this method the whole data is divided in two equal parts with respect to time. For example if we are given data from 1999 to 2016 i.e. over a period of 18 years the two equal parts will be first nine years i.e. from 1999 to 2007 and 2008 to 2016. In case of odd number of years like 9, 13, 17 etc. two equal parts can be made simply by omitting the middle year. For example if the data are given for 19 years from 1998 to 2016 the two equal parts would be from 1998 to 2006 and from 2008 to 2016, the middle year 2007 will be omitted. After the data have been divided into two parts, an average (arithmetic mean) of each part is obtained. We thus get two points. Each point is plotted against the mid year of the each part. Then these two points are joined by a straight line which gives us the trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

Advantages:

This method is simple to understand as compare to moving average methodand method of least squares.

This is an objective method of measuring trend as everyone who applies this method is bound to get the same result.

Disadvantages:

The method assumes straight line relationship between the plotted points regardless of the fact whether that relationship exists or not.

The main drawback of this method is if we add some more data to the original data then whole calculation is to be done again for the new data to get the trend values and the trend line also changes.

As the Arithmetic Mean of each half is calculated, an extreme value in any half will greatly affect the points and hence trend calculated through these points may not be precise enough for forecasting the future.

Method of Moving Average:

It is a method for computing trend values in a time series which eliminates the short term and random fluctuations from the time series by means of moving average. Moving average of a period m is a series of successive arithmetic means of m terms at a time starting with 1st, 2nd , 3rd and so on. The first average is the mean of first m terms; the second average is the mean of 2nd term to (m+1)th term and 3rd average is the mean of 3rd term to (m+2)th term and so on.

NOTES

If m is odd then the moving average is placed against the mid value of the time interval it covers. But if m is even then the moving average lies between the two middle periods which does not correspond to any time period. So further steps has to be taken to place the moving average to a particular period of time. For that we take 2-yearly moving average of the moving averages which correspond to a particular time period. The resultant moving averages are the trend values.

Advantages:

This method is simple to understand and easy to execute.

It has the flexibility in application in the sense that if we add data for a few more time periods to the original data, the previous calculations are not affected and we get a few more trend values.

It gives a correct picture of the long term trend if the trend is linear.

If the period of moving average coincides with the period of oscillation (cycle), the periodic fluctuations are eliminated.

The moving average has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than the statistician"s choice of mathematical function.

Disadvantages:

For a moving average of 2m+1, one does not get trend values for first m and last m periods.

As the trend path does not correspond to any mathematical; function, it cannot be used for forecasting or predicting values for future periods.

If the trend is not linear, the trend values calculated through moving averages may not show the true tendency of data.

The choice of the period is sometimes left to the human judgment and hence may carry the effect of human bias.

Method of Least Squares:

This method is most widely used in practice. It is mathematical method and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied.

$\sum(Y - Y_c) = 0$ i.e. the sum of the deviations of the actual values of Y and the computed values of Y is zero.

$\sum(Y - Y_c)^2$ is least, i.e. the sum of the squares of the deviations of the actual values and the computed values is least.

The line obtained by this method is called as the "line of best fit". This method of least squares may be used either to fit a straight line trend or a parabolic trend.

Measurement of seasonal variations:

There is a simple method for measuring the seasonal variation which involves simple averages.

Simple average method.

Step1. All the data are arranged by years and months(or quarters)

Step 2. Compute the simple averages (arithmetic mean) $\overline{x_i}$ for i^th^ month.

Step 3. Obtain the overall average $\bar{x}$ of these averages $\bar{x}_i$ and

$$\bar{x} = \frac{\bar{x}_1 + \cdots + \bar{x}_{12}}{12}$$

Step 4. Seasonal indices for different months are calculated by expressing monthly average as the percentage of the overall average $\bar{x}$

Thus seasonal index for ith month = $\frac{\bar{x}_i}{\bar{x}} \times 100.$

Solved problems.

problem 1. use the method of least and fit a straight line trend to the following data given from 82 to 92 . Hence estimate the trends value for 1993.

Year	82	83	84	85	86	87	88	89	90	91	92
Production in quintals	45	46	44	47	42	41	39	42	45	40	48

Solution. Let the line of best fit be y = ax + b

Take X = x - 1987 and Y = y - 42

Then the line of best fit become Y = aX + b

The normal equations are $\sum XY = a \sum X^2 + b \sum X$

$$\sum Y = a \sum X + nb \text{ when } n = 11$$

∴ From the table - 19 = 110 a Hence a = -19/110 = -0.17

17 = 11b Hence b = 17/11 = 1.55

∴The line of best fit is Y = -0.17X + 1.55

(i.e) y - 42 = -0.17(x - 1987) +1.55

y = -0.17 x + 1987 × 0.17 + 1.55 +42

∴ y = -0.17 x + 381.34 is the straight line trend.

NOTES

X	X = x - 1987	Y	Y = y - 42	XY	X^2
1982	-5	45	3	-15	25
83	-4	46	4	-16	16
84	-3	44	2	-6	9
85	-2	47	5	-10	4
86	-1	42	0	0	1
87	0	41	-1	0	0
88	1	39	-3	-3	1
89	2	42	0	0	4
90	3	45	3	9	9
91	4	40	-2	-8	16
1992	5	48	6	30	25
	0	**--**	**17**	**-19**	**110**

From the line trend, when x = 1982 , y = 44.4

x = 1983, y = 44.23	x =1984 , y = 44.06
x = 1985 , y =43.89	x = 1986 , y =43.72
x = 1987 , y =43.55	x = 1988 , y =43.38
x = 1989 , y =43.21	x = 1990 , y =43.04
x =1991 , y =42.87	x = 1992 , y =42.7

Thus the trend values are 44.4, 44.23, 44.06, 43.89, 43.72, 43.55, 43.38, 43.21, 43.04, 42.87, 42.7

Problem 2. Calculate the seasonal variation indices from the following data.

Month	**Monthly sales in lakhs Of Rs.**				**Total**	$\overline{x_1}$	**Seasonal indices** $\frac{\ddot{x}_1}{x} \times 100$
	I **1991**	**II** **1992**	**III** **1993**	**IV** **1994**			
January	10	11	11.5	13.5	46	11.5	$\frac{11.5}{12} \times 100 = 95.8$
February	8.5	8.5	9	10	36	9	$\frac{9}{12} \times 100 = 75$
March	10.5	12	11	12.5	46	11.5	$\frac{11.5}{12} \times 100=95.8$
April	12	14	16	18	60	15	$\frac{15}{12} \times 100 =125$
May	10	9	12	15	46	11.5	$\frac{11.5}{12} \times 100 =95.8$
June	10.5	10.5	11	14	46	11.5	$\frac{1.5}{12} \times 100 =95.8$
July	12	14	13	17	56	14	$\frac{4}{2} \times 100 =116.7$
August	9	8	11	16	44	11	$\frac{11}{12} \times 100 =91.7$
September	11	11	12.5	13.5	48	12	$\frac{12}{12} \times 100 =100$
October	10	9.5	11.5	13	44	11	$\frac{11}{12} \times 100 =91.7$
November	11	12.5	10.5	14	48	12	$\frac{12}{12} \times 100 =100$
December	12	13	15	16	56	14	$\frac{14}{12} \times 100=116.7$
Total						**144**	
Average						**12**	

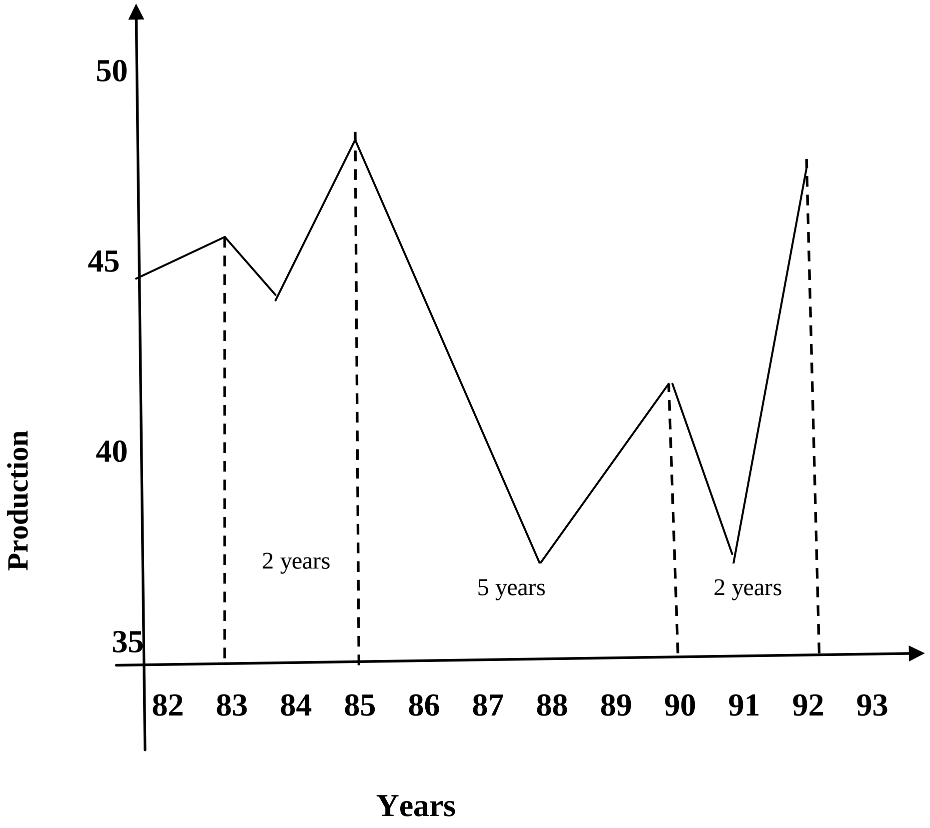
Problem 3. Compute the trend values by the method of 4 yearly moving average for the data given in problem 1.

I	II	III	IV	V	VI
Year	Production in quintals	4 yearly moving total	4 yearly moving average	2 period moving total	Trend Values (V)/2
1982	45	-	-	-	-
83	46	-	-	-	-
		182	45.50		
84	44			90.25	-
		179	44.75		
85	47				45.13
		174	43.50		
86	42			88.25	44.13
		169	42.25		
87	41			85.75	42.88
		164	41.00		
88	39			83.25	41.63
		167	41.75		
89	42			82.75	41.38
		166	41.50		
90	45			83.25	41.63
		175	43.75		
91	40			85.85	42.93
1992	48	-	-	-	-

NOTES

Problem 4. Determine the suitable period of moving average for the data given in problem

NOTES



We observe that the data has peaks at the following years 1983, 1985, 1990 and 1992 (refer the figure)

Thus the data shows 3 cycles with varying periods 2, 5, 2 respectively. Hence the suitable period of moving average is taken to be the A.M. of these periods.

Hence $\frac{2+5+2}{3} = 3$ is the period of the moving average.

Problem 5. Calculate (i) three yearly moving average (ii) short term flctuations for the data given in problem 1.

NOTES

I Year	II Production is quintals	III 3 yearly moving total	IV 3 yearly moving averages	Short term fluctuations (II - IV)
1982	45	-	-	-
83	46	135	45	1
84	44	137	45.7	-1.7
85	47	133	44.3	2.7
86	42	130	43.3	-1.3
87	41	122	40.7	0.3
88	39	122	40.7	-1.7
89	42	126	42	0
90	45	127	42.3	2.7
91	40	133	44.3	-4.3
1992	48	-	-	-

Trend values for the given time series are given in column IV.

Short term fluctuations are given in the last column.

Problem 6. Compute the seasonal indices for the following data by simple average method.

NOTES

Prices indifferent season	Season	1990	1991	1992	1993	1994
	Summer	68	70	68	65	60
	Monsoon	60	58	63	56	55
	Autumn	61	56	68	56	55
	Winter	63	60	67	55	58

Solution.

Year	**Summer**	**Monsoon**	**Autumn**	**Winter**	**Total**
1990	68	60	61	63	
1991	70	58	56	60	
1992	68	63	68	67	
1993	65	56	56	55	
1994	60	55	55	58	
Total	331	292	296	303	
Average	**66.2**	**58.4**	**59.2**	**60.6**	**244.4**
Seasonal Index	$\frac{66.2}{61.1} \times 100$ =**108.3**	$\frac{58.4}{61.1} \times 100$ =**95.6**	$\frac{59.2}{61.1} \times 100$ =**96.9**	$\frac{60.6}{61.1} \times 100$ =**99.2**	$\bar{x}$ = **61.1**

Exercises.

1. Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues what would be the predicted earnings for the year 1977 ?

Year	1970	1971	1972	1973	1974	1975	1976
Earnings in thousands	1.5	1.8	2.0	2.3	2.4	2.6	3.0

2. (i) Using three years moving average determine the trend. (ii) Also determine the short term fluctuations.

Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Production In lakhs of Tonnes	21	22	23	25	24	22	25	26	27	26

NOTES

<u>DISTANCE EDUCATION – CBCS-(2018-19 Academic Year Onwards).</u>

Question Paper Pattern(ESE)- Theory

(UG/PG/P.G Diploma Programmes)

Time : 3 hours Maximum :75 Marks

Part – A (10 X 2= 20 Marks)

Answer all questions

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

Part –B (5 X 5= 25 Marks)

Answer all questions choosing either (a) or (b)

11.a.

(or)

b.

12.a.

(or)

b.

13.a.

(or)

b.

14.a.

(or)

b.

15.a.

(or)

b.

Part –C (3 X 10 = 30 Marks)

(Answer any 3 out of 5 questions)

16.
17.
18.
19.
20.