MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI-627 012

DIRECTORATE OF DISTANCE AND CONTINUING EDUCATION

II B.Sc. MATHEMATICS

SEMESTER III

ELECTIVE III : STATISTICS I

Sub. Code: JEMA31

Prepared by

Dr. Leena Nelson S N

Associate Professor & Head, Department of Mathematics

Women's Christian College, Nagercoil – 1.

STATISTICS I

UNIT	DETAILS
I	Dispersion – Measures of Dispersion – Coefficient of Dispersion – Moments – Skewness - Kurtosis
II	Correlation – Scatter Diagram – Karl Pearson's coefficient of correlation – Probable error of Correlation Coefficient – Rank Correlation
III	Curve Fitting and Regression – Linear Regression – Curve linear Regression – Regression Curve.
IV	Theory of Attributes : Notations and Terminology – Classes and Class Frequency – Consistency of Data – Independence of Attributes – Association of Attributes.
V	Index Numbers – Consumer Price Index Numbers – Conversion of Chain Base Index number into Fixed Base conversely.

Recommended Text
S.G. Gupta and V.K. Kapoor, Fundamentals of Mathematical Statistics, 12th Edition, Sultan Chand & Sons, New Delhi, 2021.
S. Arumugam and A. Thangapandi Issac, Statistics, New Gamma Publishing House, 2016.

UNIT I

DISPERSION & MOMENTS

UNIT STRUCTURE

1.1 Measures of Dispersion

1.2 Coefficient of Dispersion

1.3 Moments

1.4 Skewness

1.5 Curtosis

INTRODUCTION

In this chapter we have described the method of classification and tabulation of data leading to frequency curve. This is to a certain extend helpful in reducing the bulk of the data. In this and following chapters we introduce several statistical constants which quantitatively describe some of the characteristics of 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 types viz.

(i) Measure of dispersion.

(ii) Measure of skewness.

(iii) Measure of kurtosis.

Here we introduce several commonly used measures of central tendencies.

Measures of central tendency given an idea of the concentration of the observations about the central part of the observations. However, these measures are inadequate to give a complete idea of the distribution.

For example, consider the two sets of observation.

I	**9**	**11**	**15**	**14**	**15**	**16**	**17**	**18**	**20**
II	**2**	**5**	**9**	**15**	**15**	**15**	**21**	**25**	**28**

We have 15 as the mean, median and mode for the two sets of observations. We observe that the individual items in the second setae scattered from the mean whereas in the first set they are closely packed. Thus, we cannot form a complete idea about the distributions from these averages. Hence averages must be supported and supplemented by some other measures. One such is measure of dispersion.

RELATIVE ADVANTAGES OF DIFFERENT MEASURES OF DISPERSION.

The qualities desired for an ideal measure of dispersion are the same as those for an ideal measure of central tendency.

(i) It should be rigidly defined and its value should be definite.

(ii) It should be based on the whole data under consideration.

(iii) It should be capable of being handled mathematically.

(iv) It should be easy to calculate and simple to follow.

(v) It should be least affected by fluctuations or sampling.

Range

Range is the most simple measure of dispersion. It is easily understood and computed but depends exclusively on the extreme values. Therefore, different samples of the same size from the same population may have widely different ranges and hence it is difficult to handle it mathematically. The range is widely used in individual quality control. Also, in many modern factories a running check is kept on the quality of the output by taking regular samples and noting both the mean and the range. Range is rarely used in advanced theory of statistics because of the mathematical difficulty of handling it.

Quartiles.

Definition. Consider a frequency distribution with a 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 $\boldsymbol{Q_1}$.

Similarly, the value of the variate for which the cumulative frequency is $3N/4$ is called the **third quartile** or **upper quartile** and it is denoted by $\boldsymbol{Q_3}$.

Clearly, median is the **second quartile** and it can also be denoted by $\boldsymbol{Q_2}$.

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

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

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 proceeding class.

Quartile deviation. (Semi inter quartile range)

The quartile deviation (Q.D) or semi inter quartile range isdefined by $Q.D = \frac{1}{2}(Q_3 - Q_1)$ where Q_1andQ_3 are the first and the thirdquartiles of the distribution.

Example. For the data in Table. 5,$Q_1 = 17$ and $Q_3 = 38.5$. (Referexampleofquartiles in section2.2).

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

Mean deviation. The mean deviation of a frequency distribution from anyaverage$A$ is defined by

$$M.D = \frac{\sum f_i |x_i - A|}{N} \; where \; N = \sum f_i.$$

Example. For the data given in *Table 4*,$\bar{x} = 27.3$.(Refer example under A.Min section 2.1) Now we find the mean deviation from the mean.

Mid x_i	f_i	$\lvert x_i - 27.3\rvert$	$f_i\lvert x_i - 27.3\rvert$
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
44.5	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 inmostcircumstances in statistics is the standard deviation.

Definition. The **standard deviation** σ of a frequency distributionis defined by

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

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**ofthe frequency distribution. Hence **variance**= $\boldsymbol{\sigma^2}$.

Note. If σ_x^2 is the variance of a sample of size$n$ the "best" estimate for the population variance $\sigma_x^2$is not σ_x^2 but $\left(\frac{N}{N-1}\right)\sigma_x^2$. For this reason,manyauthors define standard deviation by the formula

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

For large values of$N$ the two fomiulae for standard deviation arepractically indistinguishable. Throughout this book we use the first formulafor finding standard deviation of a frequency distribution. Both the formulaefor standard deviation find place in modem 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

$C.V = \frac{\sigma}{\bar{x}} \times 100.$

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.

Example 1. Consider the numbers $\mathbf{1, 2, 3, 4, 5, 6, 7}$.

The arithmetic mean $\bar{x} = 4$.

Now, $\sum(x_i - 4)^2 = (1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 28.$

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

Example 2: Calculate the mean and SD for the following table giving the age distribution of 542 numbers.

Age group	***20-30***	***30-40***	***40-50***	***50-60***	***60-70***	***70-80***	***80-90***
No. of members	***3***	***61***	***132***	***153***	***140***	***51***	***2***

Solution: Here we take $d = \frac{x-A}{h}$ here A is the middle number of 50-60.

$$\therefore A = 55$$

$$d = \frac{x-55}{10}$$

age group	mid x	f	d=(x-55)/10	fd	fd^2
20-30	25	3	-3	-9	27
30-40	35	61	-2	-122	244
40-50	45	132	-1	-132	132
50-60	55	153	0	0	0
60-70	65	140	1	140	140
70-80	75	51	2	102	204
80-90	85	2	3	6	18
Total		542	0	-15	765

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

$$= 55 + 10\left(-\frac{15}{542}\right)$$

$$= 54.72.$$

$$\sigma^2 = h^2[\frac{1}{N}\sum fd^2 - (\frac{1}{N}\sum fd)^2]$$

$$= 100\left[\frac{765}{542} - \left(-\frac{15}{542}\right)^2\right]$$

$$= 100 \times 1.4106$$

$$= 141.06.$$

Thus, $\sigma = \sqrt{141.06} = 11.9.$

Standard deviation satisfies almost all the properties for an ideal measure of dispersion except property (iv) that it is easy to calculate. This is the best and most powerful measure of dispersion. It gives greatest weight to extreme values of the observations. Mathematical treatment is also possible with standard deviation.

Theorem 1.1 . $\boldsymbol{\sigma^2 = s^2 - d^2}$***where*** $\boldsymbol{d = \bar{x} - A}$.

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

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

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

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

$= \sigma^2 + d^2$ (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 $\sigma^2 = s^2 - d^2$.

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

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

Theorem 1.2 . $\boldsymbol{\sigma = \left[\frac{\sum f_i x_i^2}{N} - \left(\frac{\sum f_i x_i}{N}\right)^2\right]^{1/2}}$

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

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

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

$= \frac{\sum f_i x_i^2}{N} - \bar{x}^2$

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

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

Theorem 1.3. The standard deviation σ is independent of change of origin and is dependent on change of scale.

Proof. We have $\sigma_x^2 = \left(\frac{1}{N}\right)\sum(x_i - \bar{x})^2$

Suppose we change the variable x_i to u_i. Where $u_i = x_i - A, A$ being an arbitrary origin.

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

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

Now, $\sigma_x^2 = \left(\frac{1}{N}\right)\sum(x_i - \bar{x})^2 = \sigma_x^2 = \left(\frac{1}{N}\right)\sum(u_i - \bar{u})^2$

$= \sigma_u^2.$

Hence σ is independent of change of origin.

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

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

$$\therefore v_i - \bar{v} = \left(\frac{1}{h}\right)(x_i - \bar{x})$$

Now, $\sigma_x^2 = \left(\frac{1}{N}\right)\sum f_i(x_i - \bar{x})^2 = \left(\frac{h^2}{N}\right)\sum f_i(v_i - \bar{v})^2$

$= h^2\sigma_v^2$

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

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 1.4. (Variance of combined set). Let the mean and standard deviation of two sets containing n_1 and n_2 members be $\overline{x_1}, \overline{x_2}$ 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 the set. Then$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right]$ where, $d_1 = \overline{x_1} - \bar{x}$ and $d_2 = \overline{x_2} - \bar{x}$.

Proof.

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

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

Now, $\sum_{i=1}^{n_1} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{n_1} f_i(x_i - \overline{x_1} + \overline{x_1} - \bar{x})^2$

$= \sum_{i=1}^{n_1} f_i[(x_i - \overline{x_1}) + d_1]^2$

$= \sum_{i=1}^{n_1} f_i(x_i - \overline{x_1})^2 + 2d\sum_{i=1}^{n_1} f_i\,(x_i - \overline{x_1}) + d_1^2\sum_{i=1}^{n_1} f_i$

$= n_1\sigma_1^2 + n_1 d_1^2$(Since $\sum f_i\,(x_i - \overline{x_1}) = 0$)

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

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)

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

Note. The above formula for σ^2 can also be written as

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1+n_2)^2}(\overline{x_1} - \overline{x_2})^2$$

We have $d_1 = \overline{x_1} - \bar{x}$ and $d_2 = \overline{x_2} - \bar{x}$

Also, we know that $\bar{x} = \frac{n_1\overline{x_1}+n_2\overline{x_2}}{n_1+n_2}$

$$\therefore d_1 = \overline{x_1} - \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2} = \frac{n_2(\overline{x_1} - \overline{x_2})}{n_1 + n_2}$$

Similarly, $d_2 = \frac{n_2(\overline{x_1}-\overline{x_2})}{n_1+n_2}$. From (1) we get,

$$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 n_2^2(\overline{x_1} - \overline{x_2})^2}{(n_1+n_2)^2} + \frac{n_2 n_1^2(\overline{x_2} - \overline{x_1})^2}{(n_1+n_2)^2}\right]$$

$$\therefore \sigma^2 = \frac{1}{n_1+n_2}\left[n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{x_1} - \overline{x_2})^2\right]$$

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.

$$\mathbf{20.22, 27, 30, 40, 48, 45, 32, 31, 35.}$$

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

(ii) Range = Maximum value – Minimum value

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

Here we have $\sum x_i^2 = 11652$ (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 $\boldsymbol{n}$ natural numbers is $\frac{\mathbf{1}}{\mathbf{12}}(\boldsymbol{n}^{\mathbf{2}}-\mathbf{1})$.

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

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

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

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

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

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

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

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

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

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.

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 $\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{u_i}$	$\boldsymbol{f_iu_i}$	$\boldsymbol{f_iu_i^2}$	$\boldsymbol{c.f}$
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**	**-**

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

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

$$= 300 - \frac{1550}{72}$$

$$= 300 - 21.53$$

$$= 278.47.$$

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

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

$= 211.36$

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

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

$= 314.29$

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

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

$$= 51.45.$$

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

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

$$\therefore \sigma = 88.52\ (verify)$$

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

$= 31.79.$

Problem 4. Find the arithmetic mean $\bar{x}$, standard deviation σ and percentage of cases 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***

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**

$$\bar{x} = \frac{\sum f_i\, x_i}{N} = \frac{374}{57} = 6.56$$

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

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

$$= \frac{2618 \times 57 - 374^2}{57^2}$$

$$= \frac{9350}{572}$$

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

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.

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

Now, $\bar{x} \pm 2\sigma = 6.56 \pm 3.4 = 9.96, 3.16$.

There are only 53 items $[3 + 7 + 12 + 15 + 11 + 5]$ which lie 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.

$$We\ have\ \sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$$

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

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

$$= 360.9375.$$

$$\therefore \sigma = 16\ (approximately).$$

C.V of marks of class A$= \frac{\sigma_1}{\overline{x_1}} \times 100 = \frac{15}{80} \times 100$

$= 18.75$

C.V of marks of class B$= \frac{\sigma_2}{\overline{x_2}} \times 100 = \frac{20}{85} \times 100$

$= 23.53.$

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

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 $\sigma^2 \geq m^2$.

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

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

$$\Leftrightarrow \left(\frac{1}{N}\right)[\sum f_i\,{z_i}^2 - (\sum f_i\,z_i)^2] \Leftrightarrow \sigma_2^2 \geq 0$$

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 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: $\bar{y} = \frac{380}{10} = 38.$

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**

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

Similarly,

$$\sigma_y = \left[(1/n)\sum(y_i - \bar{y})^2\right]^{1/2} = \left[\frac{10680}{10}\right]^{1/2} = \sqrt{1068} = 32.68.$$

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

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

Since, $\bar{x} > \bar{y}$ cricketer A is better run getter. $C.V.$ of $A < C.V.$ of B, criketer 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 $n = 200;\ \bar{x} = 60;\ \sigma = 20.$

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

$$\therefore \sum x_i = 12000$$

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

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

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

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

$$\text{After correction} \sum x_i^2 = 800000 - (3^2 + 6^2) + (13^2 + 17^2) = 795960.$$

$$\therefore \text{corrected}\sigma^2 = \frac{795960}{200} - (59.8)^2 = 3979.8 - 3576.04 = 403.76$$

$$\therefore \sigma = 20.09.$$

Problem 9. Find (i) the mean deviation from the mean (ii) variance of the arithmetic progression $a, a + d, a + 2d, \ldots, a + 2nd$.

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

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

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

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

$$= a + nd$$

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

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

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

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

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

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

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

1.2 COEFFICIENT OF DISPERSION

Whenever we want to compare the variability of the 2 series, which are differ widely in there averages or which are measured in different units, we do not calculate the measures of dispersion but we calculate the coefficient of dispersion which are pure numbers, independent of units of measurements.

The coefficient of dispersion based on different measures of dispersion is as follows:

(i) Coefficient of dispersion based upon range = $\frac{A-B}{A+B}$ where A and B are the greatest and the smallest items in the series.

(ii) Based upon quartile deviation =$\frac{Q_3-Q_1}{Q_3+Q_1}$.

(iii) Based upon mean deviation = $\frac{mean\ deviation}{average\ form\ which\ is\ calculated}$

(iv) Based upon standard deviation= $\frac{SD}{mean} = \frac{\sigma}{\bar{x}}$.

Coefficient of Variation (CV)

100 times the coefficient of dispersion based upon SD is called coefficient of variation.

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

According to Prof. Karl Pearson who suggested the measure, CV is the percentage variation in the mean. SD being consider as the total variation in the mean.

The series having the greater CV is said to be more variable than the other and the series having lesser CV is said to b more consistent than the other.

Problem 10: Calculate the C.V. for a series for which the following results are known: $\boldsymbol{n = 50, \sum d_i = -10, \sum d_i^2 = 400,\ where\ d_i = x_i - 75}$.

Solution : Given $n = 50, \sum d_i = -10, \sum d_i^2 = 400,\ d_i = x_i - 75.$

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

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum(d_i + 75)}{n} = 75 + \frac{\sum d_i}{n}$$

$$= 75 + \frac{(-10)}{50} = 75 - 0.2 = 74.8.$$

$$\sigma^2 = \frac{\sum d_i^2}{n} - \left(\frac{\sum d_i}{n}\right)^2$$

$$= \frac{400}{50} - \left(\frac{-10}{50}\right)^2$$

$$= 8 - 0.04 = 7.96.$$

$$\therefore \sigma = \sqrt{7.96} = 2.82$$

$$CV = \frac{2.82}{74.8} \times 100 = 3.8.$$

Problem 11: Given $\boldsymbol{\sum x_i = 99, n = 9, \sum(x_i - 10)^2 = 79.}$ ***Find*** $\boldsymbol{\sum x_i^2}$***, hence find*** $\boldsymbol{\sigma^2}$**.**

Solution : Given, $\sum x_i = 99, n = 9, \sum(x_i - 10)^2 = 79.$

$$\sum(x_i - 10)^2 = 79 \Rightarrow \sum x_i^2 - 20 \sum x_i + n \times 100 = 79.$$

$$\Rightarrow \sum x_i^2 - 20 \times 99 + 9 \times 100 = 79$$

$$\Rightarrow \sum x_i^2 = 79 - 900 + 1980 = 1159.$$

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

$$= \frac{1}{n}\sum (x_i - 11)^2 \qquad \left[since, \bar{x} = \frac{\sum x_i}{n} = \frac{99}{9} = 11\right]$$

$$\sum (x_i - 11)^2 = 70.$$

Thus $\sigma^2 = 1159.$

Problem 12: The means of two samples of sizes 50 and 100 respectively are 54.1 and 50.3 and the SD are 8 and 7. Obtain the mean and SD of sample of size 150, obtained by combining the two samples.

[Hint : $\overline{x_1} = 54.1, \overline{x_2} = 50.3, \sigma_1 = 8, \sigma_2 = 7;\ \bar{x} = \frac{n_1\overline{x_1}+n_2\overline{x_2}}{n_1+n_2} = 51.57; d_1 = \overline{x_1} - \bar{x}, d_2 = \overline{x_2} - \bar{x};\ \sigma^2 = \frac{1}{n_1+n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)] = 57.2$]

1.3 MOMENTS

In this section, we introduce some more statistical constants known as moments.

Definition. The $\boldsymbol{r^{th}}$ **moment about any point $\boldsymbol{A}$**, 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 $\mu'_r = \frac{\sum f_i x_i^r}{N}$ which is the r^{th} moment about the origin.

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

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

μ_r is also called the $\boldsymbol{r^{th}}$ **central moment.**

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

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

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 1.5

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

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 $d = \bar{x} - A$

$= (1/N)\left[\sum f_i\,(x_i - A)^r - r_{c_1} d \sum f_i\,(x_i - A)^{r-1} + r_{c_2} d^2 \sum f_i\,(x_i - A)^{r-2}\right]$

$- \cdots + r_{c_{r-1}}(-d)^{r-1}\sum f_i\,(x_i - A) + r_{c_r}(-d)^r \sum f_i$

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

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

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)$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$

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

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

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

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 h^r times the r^{th} moment of the variable u_i.

Definition. Karl Peterson's $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ coefficientsare defined as follows

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

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

The above four coefficients depend upon the first four central moments. They are pure numbers independent of units in which the variable x_i is expressed. Also, their values are not affected by change of origin and scale. These constants are used in section 4.2 in the study of skewness and kurtosis.

1.4 SKEWNESS

If the value 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 clubbed, they retain their signs and cancel on addition.

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

Thus, in the case of symmetrical distribution $\beta_1 = 0$. If adistribution fails to be symmetric(asymmetric) then we say that it is a 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 $\beta_1 > 0$and **negative skewness** if $\beta_1 < 0$.

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.

∴**Mean - Mode** and **Mean −Median** may be taken as measures of skewness.

These measures wen suggested by **Karl Pearson.**

Another measure of skewness due to **Bowley** is based on the fact that for a positively skewed distribution the third quartile is farther from the median than the first quartile so that $Q_3 -$ Median> Median $-Q_1$. Hence($Q_3 -$ Median) − (Median $-Q_1$) $= Q_3 + Q_1 - 2$Median may be taken as another measure of skewness.

The above measures of skewness are the **absolute 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 Pearson's coefficient of skewness.

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

are called Pearson's coefficients of skewness.

(ii) Bowley's coefficient ofskewness is given by

$$\frac{\boldsymbol{Q_3-Q_1-2\,Median}}{\boldsymbol{Q_3-Q_1}}$$

1.5 KURTOSIS

Definition. Kurtosis is the *degree of peakedness* of a distribution usually taken relative to a normal distribution. Thus, kurtosis enables us to have an idea about the *flatness or peakedness* of a frequency curve. It is measured by the coefficient β_2.

For a normal curve $\beta_3 = 3$ or $(\gamma_2 = 0)$**mesokurtic.**

For a curve which is flatter than the normal curve$\beta_3 < 3$ or $(\gamma_2 < 0)$and such a curve is known as **platykurtic.**

For a curve which is more peaked than the normal curve$\beta_3 > 3$ or $(\gamma_2 > 0)$and such a curve is known as **leptokurtic.**

Solved problems.

Problem 1. Calculate the first four central moments from the following datato find $\beta_1$and β_2 and discuss the nature of the distribution.

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

Solution. Here $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{300}{100} = 3.$

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

x_i	f_i	u_i	f_iu_i	$f_iu_i{}^2$	$f_iu_i{}^3$	$f_iu_i{}^4$
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**

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

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

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

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

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

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

Since, $\beta_1 > 0$ the distribution is *positively skewed.*

Since $\beta_1 = 2.237 < 3$ the distribution is *platykurtic.*

Problem 2. Calculate the values of β_1 and β_2 for the distribution given in table 4.

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

x_i	f_i	u_i	f_iu_i	$f_i{u_i}^2$	$f_i{u_i}^3$	$f_i{u_i}^4$
04.5	11	-2	-22	44	-88	176
14.5	20	-1	-20	20	-20	20
24.5	16	0	0	0	0	0
34.5	36	1	36	36	36	36
44.5	17	2	34	68	136	272
Total	**100**	**0**	**28**	**168**	**64**	**504**

Here we have chosen $A = 24.5$ and $h = 10$.

$$\mu'_1 = \left(\frac{1}{N}\right)\sum f_i\,(x_i - A) = \frac{1}{N}\sum f_iu_i \times h = \frac{28}{100} \times 10 = 2.8$$

$$\mu'_2 = \left(\frac{1}{N}\right)\sum f_i\,u_i^2 \times h^2 = \frac{168}{100} \times 10^2 = 168.$$

$$\mu'_3 = \left(\frac{1}{N}\right)\sum f_i\,u_i^3 \times h^3 = \frac{64}{100} \times 10^3 = 640.$$

$$\mu'_4 = \left(\frac{1}{N}\right)\sum f_i\,u_i^4 \times h^4 = \frac{504}{100} \times 10^4 = 50400.$$

Now, $\mu_1 = 0$.

$\mu_2 = \mu'_2 - \left(\mu'_1\right)^2 = 168 - (2.8)^2 = 160.16.$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_2 + 2\left(\mu'_1\right)^3 = 640 - 3 \times 168 \times 2.8 + (2.8)^3 = -727.296$

$\mu_4 = \mu'_4 - 4\mu'_3 + 6\mu'_2\left(\mu'_1\right)^2 - 3\left(\mu'_1\right)^4$

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

$= 50950.323$

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

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

Problem 3. The first 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 zero.

Solution. Given $\mu'_1 = 1;\ \mu'_2 = 2.5;\ \mu'_3 = 5.5;\ \mu'_4 = 16$ where $A = 2$.

(i) *Moments about mean.*

$\mu_1 = 0.$

$\mu_2 = \mu'_2 - \left(\mu'_1\right)^2 = 2.5 - 1 = 1.5$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\left(\mu'_1\right)^3 = 5.5 - 3 \times 2.5 + 2 = 0$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\left(\mu'_1\right)^2 - 3\left(\mu'_1\right)^4$

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

(ii) *Moments about zero.*

We have $\bar{x} = A + \mu'_1$

$= 2 + 1 = 3$

Now, the first moment about zero $\mu'_1 = \left(\frac{1}{N}\right)\sum f_i\,(x_i - 0)$

$= \bar{x} = 3.$

Now, $\mu'_2 = \mu_2 + \left(\mu'_1\right)^2 = 1.5 + 3^2 = 10.5$

$\mu'_3 = \mu_3 + 3\mu'_2\mu'_1 + \left(\mu'_1\right)^3 = 0 + 3 \times 1.5 \times 3 + 3^3 = 40.5$

$\mu'_4 = \mu_4 + 4\mu'_3\mu'_1 + 6\mu'_2\left(\mu'_1\right)^2 + 3\left(\mu'_1\right)^4$

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

Problem 4. The first three moments about the origin are given by $\boldsymbol{\mu'_1 = \frac{1}{2}(n+1)}$; $\boldsymbol{\mu'_2 = \frac{1}{6}(n+1)(2n+1)}$; $\boldsymbol{\mu'_3 = \frac{1}{4}n(n+1)^2}$. ***Examine the skewness of the distribution.***

Solution. $\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$

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

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

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

Hence, $\mu_3 = 0$.

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = \frac{1}{6}(n+1)(2n+1) - \left[\frac{1}{2}(n+1)\right]^2$$

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

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

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

$\therefore$ when $n > 1, \beta_1 = 0$.

Hence the distribution is symmetric.

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

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

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

Now, $\mu_4 - \mu_2^2 = \frac{\sum f_i(x_i-\bar{x})^4}{N} - \left[\frac{\sum f_i(x_i-\bar{x})^2}{N}\right]^2$

$= \frac{\sum f_i z_i{}^2}{N} - \left(\frac{\sum f_i z_i}{N}\right)^2$ 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 $\beta_2 \geq 1$

<u>EXERCISE PROBLEMS:</u>

1. for the following data calculate the Karl Pearson’s coefficient of skewness.

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

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	75	**Total**	**590**

3. Karl Pearson's coefficient of skewness of a distribution is 0.4, it's S.D is 8 and mean 30. Find 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
f	1	8	28	56	70	56	28	8	1

5. Find β_2 for the following frequency distribution.

Class	0-10	10-20	20-30	30-40
f	1	3	4	2

6. Calculate the first four central moments from the following data. Also calculate β_1 and β_2 and comment on the nature of the distribution.

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	20	15	10	5

7. Calculate the first four moments of the following distribution about $x = 4$ and hence find the moments about the mean of the distribution. Also find the values of β_1 and β_2.

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

8. The first four moments of a distribution about $x = 4$ are $-1.5, 17, -30$ and 108. Find the first four moments (i) about mean (ii) about the origin (iii) about $x = 2$ (iv) Also calculate β_1 and β_2.
9. The first three moments of a distribution about the value 3 of the variables are 2, 10 and 30 respectively. Obtain the first three moments about zero.
10. The first four moments of a distribution about $x = 5$ are 2, 20, 40 and 50. Show that the mean = 7, variance $= 16, \mu_3 = -64, \mu_4 = 162, \beta_1 = 1$ and $\beta_2 = 0.63$.
11. For a distribution the mean is 10, the variance is 16, γ_1 is 1 and β_2 is 4. Find the first four moments about the origin.
12. The marks of 250 students appeared for an examination give the following information's. $\bar{x} = 39.72, \sigma^2 = 97.8, \mu_3 = -114.18, \mu_4 = 28396.14$. It was later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moments.
13. The first four moments of a distribution about $x = 4$ are respectively 1, 4, 10, 45. Calculate the moments about the mean.
14. In calculating the moments of a frequency distribution based on 100 observations the following results were obtained $\bar{x} = 9, \sigma^2 = 19, \beta_1 = 0.7, \beta_2 = 4$. But later on it is noticed that one observation 12 was read 21. Obtain the correct values of the mean and the first four central moments.
15. In a frequency distribution the coefficient if skewness based upon the quartiles is 0.6. If the sum of the third and first quartiles is 100 and median is 38. Find the value of the upper and lower quartiles.
16. Find the C.V of a frequency distribution given that its mean is 120, mode is 123 and Karl Pearson's coefficient of skewness is -0.3.

UNIT II

CORRELATION AND REGRESSION

UNIT STRUCTURE

2.1 Correlation

2.2 Rank Correlation

INTRODUCTION

In statistics 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 different fields of activities giving n pairs of corresponding observations $(x_i, y_i); i = 1,2, \dots, n$.

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

There are two main problems involved in the relationship between x and y. The first is to find the measure of the degree of association or **correlation** between the values ox 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 other. This is the problem of **regression**.

2.1 CORRELATION

Definition. Consider a set of bivariate data $(x_i, y_i); i = 1,2, \dots, n$. If there is a change in one variable corresponding to a 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 v; pressure p of a perfect gas which obeys the law $pv = k$ where k is a constant, are examples of variables with negative correlation.

Definition. Karl Pearson's coefficient of 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

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

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***

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

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

$$= 17$$

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

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

Theorem 2.1.

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

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

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

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

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

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

$$= \frac{1}{n}[n\sum x_i y_i - \sum x_i \sum y_i]$$ ……………….(2)

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

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

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

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

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

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

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

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

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

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

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

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

Also, $\sigma_x = h\sigma_u$ and $\sigma_y = k\sigma_v$

$$\gamma_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} = \frac{hk\sum(u_i - \bar{u})(v_i - \bar{v})}{n(h\sigma_u)(k\sigma_v)}$$

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

$$= \gamma_{uv}.$$

Hence, $\gamma_{xy} = \gamma_{uv}$.

Theorem 2.3. $\boldsymbol{-1 \leq \gamma \leq 1}$.

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

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

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

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

By Schwartz inequality we have $(\sum a_i b_i)^2 \leq (\sum a_i^2)(\sum b_i^2)$

Hence, $\gamma_{xy}^2 \leq 1$

$\therefore |\gamma_{xy}| \leq 1$

$\therefore -1 \leq \gamma \leq 1.$

Note 1. 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 $\boldsymbol{cov\ (x, y) = 0}$.

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

Theorem 2.4 . $\boldsymbol{\gamma_{xy} = \frac{\sigma_x^2+\sigma_y^2-(\sigma_x-y)^2}{2\sigma_x\sigma_y}}$

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

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

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

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

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

Solved Problems.

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

$\boldsymbol{x_i}$	$\boldsymbol{u_i}$	$\boldsymbol{y_i}$	$\boldsymbol{v_i}$	$\boldsymbol{u_i^2}$	$\boldsymbol{v_i^2}$	$\boldsymbol{u_iv_i}$
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**	**500**

$\gamma_{xy} = \gamma_{uv}$ (by theorem 2.2)

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

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

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

$= 0.6.$

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 $v = cy + d$

$\therefore \bar{u} = a\bar{x} + b$ 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, $\sigma_v^2 = a^2\sigma_y^2$

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

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

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

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

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

$$= \frac{62100 - 63000}{(112500 - 90000)^{1/2}(58500 - 44100)^{1/2}}$$

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

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

$$= -\frac{1}{20}$$

$$= -0.05$$

Problem 4. If x, y and z 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 $\Longrightarrow \sum(x - \bar{x})(y - \bar{y}) = 0$

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

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

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

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

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

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

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

$= 2\sigma^2$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Problem 6. Show that if X', Y' are the derivations of the random variables X and Y fro 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 $X_i' = X_i - \bar{X}$ and $Y_i' = Y_i - \bar{Y}$

$$1 - \frac{1}{2N}\sum\left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 = 1 - \frac{1}{2N}\sum\left(\frac{X_i - \bar{X}}{\sigma_X} - \frac{Y_i - \bar{Y}}{\sigma_Y}\right)^2$$

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

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

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

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

$$= 1 - (1 - \gamma)$$

$$= \gamma$$

(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, $1 - \frac{1}{2N}\sum\left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 \leq 1$

$\therefore$ By (i) $\gamma \leq 1$ similarly by (ii) $-1 \leq \gamma$

Hence, $-1 \leq \gamma \leq 1$.

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 $\gamma_{uv} = 0$ (i.e.u and v are un correlated) find the value of k.

Solution. $u = x + ky \implies \bar{u} = \bar{x} + k\bar{y}$

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

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

Now, $\gamma_{uv} = 0 \implies cov\,(u, v) = 0$

$$\implies \sum(u - \bar{u})(v - \bar{v}) = 0$$

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

$$\implies \sum(x - \bar{x})^2 + k\left(\frac{\sigma_x}{\sigma_y}\right)\sum(y - \bar{y})^2 + k\sum(x - \bar{x})(y - \bar{y}) + \left(\frac{\sigma_x}{\sigma_y}\right)\sum(x - \bar{x})(y - \bar{y}) = 0$$

$$\implies n\sigma_x^2 + nk\left(\frac{\sigma_x}{\sigma_y}\right)\sigma_y^2 + \gamma_{xy}\sigma_x\sigma_y\left(k + \frac{\sigma_x}{\sigma_y}\right) = 0$$

$$\implies n\sigma_x\left[\sigma_x + k\sigma_y + \gamma_{xy}\left(k\sigma_y + \sigma_x\right)\right] = 0$$

$$\implies n\sigma_x\left[\left(\sigma_x + k\sigma_y\right)\left(1 + \gamma_{xy}\right)\right] = 0$$

$$\implies \sigma_x\left(\sigma_x + k\sigma_y\right)\left(1 + \gamma_{xy}\right) = 0$$

$$\implies \sigma_x + k\sigma_y = 0 \; or \; 1 + \gamma_{xy} = 0 \; or \; \sigma_x = 0$$

If $\gamma_{xy} \neq -1$ and $\sigma_x \neq 0$ we get $k = -\left(\frac{\sigma_x}{\sigma_y}\right)$.

EXERCISE QUESTIONS:

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

(iv)

Sales	50	50	55	60	65	65	65	60	60	60
Expenses	11	13	14	16	16	15	15	14	13	13

(v)

Length	3	4	6	7	10
Weight	9	11	14	15	16

(vi)

Marks in math	65	66	67	67	68	69	70	72
Marks in statistics	67	68	65	68	72	72	69	71

(vii)

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

(viii)

Marks in Maths	77	54	27	52	14	35	90	25	56	60
Marks in English	35	58	60	40	50	40	35	56	34	42

(ix)

Production	250	270	278	325	260	310	428	320	440	310
Price	84	50	62	75	90	170	136	65	72	58

(x)

Father's height	67	68	64	67	72	70	70	69	70
Son's height	65	66	67	68	68	69	71	72	73

(xi)

Saving deposits in lakh of Rs.	54	57	59	62	68	63	73	78
Strikes	36	42	41	34	31	21	11	9

2. If $\sum x = 71, \sum y = 70, \sum x^2 = 555, \sum y^2 = 526$ and $\sum xy = 527 n = 100$ find γ_{xy}.
3. Given $n = 1000$; $\bar{x} = 65$; $\bar{y} = 83$; $\sigma_x = 4.5$; $\sigma_y = 3.6$ and the sum of the products of deviations from the mean x and y is 4800. Find the correlation coefficient between x and y.
4. In the two sets of variables x and y with 50 observations each of the following data were observed. $\bar{x} = 10$; $\bar{y} = 6$; $\sigma_x = 3$; $\sigma_y = 2; \gamma = 0.3$. But on subsequent verification it was found that one value of $x = 10$ and one value of $y = 6$ were found

inaccurate and hence weeded out. With the remaining 49 pairs of values how is the original value of γ affected?

5. Two independent variables x and y have means 5 and 10 and variances 4 and 9 respectively. Obtain the correlation coefficient between u and v where $u = 3x + 4y$ and $v = 3x - y$.
6. If $z = ax + by$ and γ is the correlation coefficient between x and y show that $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 - 2ab\gamma\sigma_x\sigma_y$. Hence deduce that $\gamma = \frac{\sigma_x^2+\sigma_y^2-\left(\sigma_{x-y}\right)^2}{2\sigma_x\sigma_y}$.
7. Show that $\gamma = \frac{\left(\sigma_{x-y}\right)^2-\sigma_x^2-\sigma_y^2}{2\sigma_x\sigma_y}$.
8. If x and y are discrete variables and if$\sigma_x^2 = \sigma^2 = \sigma_y^2$ and $cov\,(x, y) = \frac{1}{2}\sigma^2$ find (i) σ_{2x-3y} (ii) $\gamma_{2x+3,2y-3}$
9. If the variates x and y have zero means, the same variance σ^2 and zero correlation then show that the variates $u = x\cos\alpha + y\sin\alpha$ and $v = x\sin\alpha + y\cos\alpha$ have the same variance σ^2 and zero correlation.
10. If a, b, c are positive consonants σ_1 and σ_2 are standard deviations of x and y and γ be the correlation between x and y show that the correlation between cx and $ax + by$ is $\frac{a\sigma_1+b\gamma\sigma_2}{\sqrt{a^2\sigma_1^2+b^2\sigma_2^2+2ab\gamma\sigma_1\sigma_2}}$. Deduce that if $\sigma_1 = \sigma_2$ then the correlation between x and $x + y$ is $\sqrt{(1+\gamma)/2}$.

2.2 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 values $1, 2, 3, \dots, n$ 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 characteristic. 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 2.5 . 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 the i^{th} individual in the two different rankings.

$$\therefore \bar{x} = \frac{1}{2}(n+1) = \bar{y} \; and \; \sigma_x^2 = \frac{1}{12}(n^2-1) = \sigma_y^2.$$

Now, $\sum(x-y)^2 = \sum[(x-\bar{x})-(y-\bar{y})]^2$ (since$\bar{x} = \bar{y}$)

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

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

$= 2n\sigma_x^2(1-\rho)$ (Since $\sigma_x^2 = \sigma_y^2$)

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

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

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

Note 1. This is 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 coefficient will not apply since in this case $\bar{x} \neq \bar{y}$. In such case we assign a common rank to the 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 items an item has repeated values. This correlation factor is added for each repeated rank of the variable x and y. For example, after assigning rank 2 if four items get the same rank 3 then these four 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 correlation factor to be added is $\frac{1}{12} \times 4 \times (4^2-1) = 5$.

Problem 1. Find the rank correlation coefficient between the height in c.m and the 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**	$x-y$	$(x-y)^2$
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 Px**	Q	**Rank in Qy**	$x-y$	$(x-y)^2$
35	8	50	6	2	4
56	2	35	8	-6	36
50	3.5	70	2	1.5	2.25
65	1	25	10	-9	81
44	5.5	35	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 corrected 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, $\rho = 1-\frac{6\sum(x-y)^2}{n(n^2-1)} = 1-\frac{6\times188}{10\times99} = 1-\frac{1128}{990}$

$= 1-1.139 = -0.139.$

Problem 3. Three judges assign the ranks to 8 entries in a beauty contest.

Judge Mr. X	***1***	***2***	***4***	***3***	***7***	***6***	***5***	***8***
Judge Mr. Y	***3***	***2***	***1***	***5***	***4***	***7***	***6***	***8***
Judge Mr. Z	***1***	***2***	***3***	***4***	***5***	***7***	***8***	***6***

Which pair of judges has the nearest approach to common task in beauty?

Solution. Table for rank correlation coefficients ρ_{xy}, ρ_{yz} and ρ_{zx}.

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{z}$	$(\boldsymbol{x-y})$	$(\boldsymbol{x-y})^2$	$(\boldsymbol{y-z})$	$(\boldsymbol{y-z})^2$	$(\boldsymbol{z-x})$	$(\boldsymbol{z-x})^2$
1	3	1	-2	4	2	4	0	0
2	2	2	0	0	0	0	0	0
4	1	3	5	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**

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

$$= 1 - \frac{6\times28}{8\times(8^2-1)}$$

$$= 1 - \frac{168}{504}$$

$$= 1 - 0.333$$

$$\rho_{yz} = 1 - \frac{6\times18}{8\times63}$$

$$= 1 - \frac{108}{504}$$

$$= 1 - 0.214$$

$$\rho_{zx} = 1 - \frac{6\times20}{8\times63}$$

$$= 1 - \frac{120}{504}$$

$$= 1 - 0.238$$

$$= 0.762$$

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 differences in marks 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. $\rho_{xy} = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$

Given $\rho_{xy} = 0.8$ and $n = 10$.

$\therefore 0.8 = 1 - \frac{6\sum(x-y)^2}{10(10^2-1)} = 1 - \frac{6\sum(x-y)^2}{990}$

$$\frac{6\sum(x-y)^2}{990} = 1 - 0.8 = 0.2$$

$$6\sum(x-y)^2 = 990 \times 0.2 = 198.$$

$$\therefore \sum(x-y)^2 = 33$$

Corrected $\sum(x-y)^2 = 33 - 5^2 + 8^2 = 72$

Now, after correction $\rho_{xy} = 1 - \frac{6\times72}{10\times(10^2-1)}$

$$= 1 - \frac{432}{990}$$

$$= 1 - 0.436$$

$$= 0.564$$

The corrected coefficient of rank correlation is 0.564.

Problem 5. Let $x_1, x_2, \dots, x_n$ be the ranks of n individuals according to a character A and $y_1, y_2, \dots, y_n$ the ranks of the same individuals according to another character B. It is given that $x_i + y_i = 1 + n\, for\ i = 1, 2, \dots, n$. Show that the rank of the correlation coefficient ρ between the character 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, $\rho = 1 - \frac{6\sum(x-y)^2}{n(n^2-1)}$

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

$= 1 - 2$

$= -1$

<u>EXERCISE QUESTIONS:</u>

1. Calculate the rank correlation coefficient for the following data.

(i)

$\boldsymbol{x}$	5	2	8	1	4	6	3	7
$\boldsymbol{y}$	4	5	7	3	2	8	1	6

(ii)

x	10	12	18	18	15	40
y	12	18	25	25	50	25

(iii)

x	20	25	60	45	80	25	55	65	25	75
y	52	50	55	50	60	70	72	78	80	63

(iv)

x	33	56	50	65	44	38	44	50	15	26
y	50	35	70	25	35	58	75	60	55	26

(v)

x	35	56	20	65	42	33	44	50	15	60
y	50	35	70	25	58	75	60	45	80	38

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 marks in two subjects.

Economics	78	65	36	98	25	75	82	90	62	39
Statistics	84	53	51	91	60	68	62	86	58	47

Calculate the rank correlation coefficient.

4. The following table shows how 10 students were ranked according to their achievements in the laboratory and lecture portions of a biology course. Find the coefficient of rank correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

5. The $I.Q^5$ of a group of 6 persons were measured and then they sat for certain examination. Their$I.Q^5$ and examination marks were as follows.

Exam marks	70	60	80	60	10	20
$\boldsymbol{I.Q}$	110	100	140	12	80	90

Compute the coefficients of correlation and rank correlation.

Why are the correlation figures obtained different?

6. Ten competitions in a beauty contest were ranked by three judges in the following order.

Judge I	1	6	5	10	3	2	4	9	7	8
Judge II	3	5	8	4	7	10	2	1	6	9
Judge III	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

7. Following are the marks obtained by 10 students in first three semester in three ancillary papers out of 75.

Semester I (Ancillary I)	60	55	75	45	69	45	72	39	35	45
Semester II (Ancillary II)	70	58	73	49	60	49	60	55	60	48
Semester III (Ancillary III)	55	61	68	40	58	60	50	38	50	60

Calculate the rank correlation coefficient between Ancillaries I and II; Ancillaries II and III; and Ancillaries III and I.

8. The coefficient of rank correlation of the marks obtained by 10 students in Physics andChemistry was found to be 0.5. It was later discovered that the difference in ranks

in the two subjects obtainedby one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

9. A computer while calculating while calculating the correlation coefficient between two variables x and y obtained the following constants.

   $n = 25$; $\sum x = 125$; $\sum x^2 = 650$; $\sum y = 100$; $\sum y^2 = 460$; and $\sum xy = 508$.

   It was later discovered at the time of checking that he had copied down two pairs of observations (x_i, y_i) as $(6, 14)$ and $(8, 6)$ instead of the correct values $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient between x and y.

10. The coefficient of rank correlation between marks in Statistics and Mathematics obtained by a certain group of students is 0.8. If the sum of the squares of the difference in the ranks is given to be 33 find the number of students in the group.

UNIT III

CURVE FITTING AND REGRESSION

UNIT STRUCTURE

3.1 Curve Fitting

3.2 Regression

INTRODUCTION

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 variables comparing the sample data.

Let x_i where $i = 1,2, \dots, n$ be the values of the independent variable x_i and y_i where $i = 1,2, \dots, n$ be the corresponding values of the dependent variables y_i. If the points $(x_i, y_i), i = 1,2, \dots, n$ are plotted on a graph paper and we obtain a diagram called *scatter diagram*. Hence of there is functional relationship between x_i and y_i. The points of the scatter diagram will be found to be concentrated round a certain curve. In this chapter we shall see how such a curve can be fitted to indicate a functional relationship between two variables 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 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.

PRINCIPLE OF LEAST SQUARES

Among the many methods available for curve fitting the most popular method is the principle of least squares. Let (x_i, y_i), where $i = 1,2, \dots, n$ be the observed set of values of the variable (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 value of y determined by the functional relation is called the **residuals.** The principle of the least squares states that the parameters involved in $f(x)$ should be chosen in such a way that $\sum d_i^2$ is minimum.

Fitting a straight line

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

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

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

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

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

$$\therefore a\sum x_i^2 + b\sum x_i = \sum x_i y_i \qquad \dots\dots\dots\dots(1)$$

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

$$\therefore a\sum x_i + nb = \sum y_i \qquad \dots\dots\dots\dots(2)$$

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

Fitting a second-degree parabola.

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

The residual d_i is given by $d_i = y_i - (a\,x_i^2 + bx_i + c)$

$$\therefore \sum d_i^2 = \sum(y_i - ax_i^2 - bx_i - c)^2 = R\ (say)$$

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

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

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

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

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

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

$$\therefore a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i\, y_i \dots\dots\dots\dots\dots(2)$$

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

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

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

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 from by some suitable transformations of variables. 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 delt with here in solved problems.

Solved Problems.

Problem 1. Fit a straight line to the following data.

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

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.

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

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

$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{x_i y_i}$	$\boldsymbol{x_i^2}$
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 10	**26.5**	**69.0**	**30**

Hence the normal equations are

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

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

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

$\therefore$ The straight line fitted for the data is $y = 1.6x + 2.1$.

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

Solution.

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

$$-3a + 6b = -12 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$19a - 3b = 67 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

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

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

$\therefore 5y - 110 = 3.49x - 52.35 - 1.30$

$\therefore 5y = 3.49x + 56.35$

$\therefore y = .698x + 11.27.$

Now for $x = 6$ the estimated value og y is $y = .698 \times 6 + 11.27 = 15.458$

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

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

Solution. Let the second-degree parabola to be fitted to the data be $y = ax^2 + bx + c$. 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 $y = 2.21\,x^2 + 0.26x + 1.42$.

Fitting of power curves

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. $y = bx^a$

∴ $\log y = a \log x + \log b$.

Let $\log y = Y$ and $\log x = X$

Then the curve is transformed into $Y = AX + B$ where $A = a$ and $B = \log b$. Hence 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**

∴ The normal equations are

$$2.9\,A + 6B = 14.9\ approximately$$

$$1.8A + 2.9B = 6.6\ approximately$$

$$\therefore A = -2.3\ and\ B = 3.6\ (verify)$$

$$\therefore A = a = -2.3\ and\ B = \log b = 3.6$$

$$\therefore a = -2.3\ and\ b = \text{antilog}\,3.6 = 3981$$

∴ 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)

$\therefore \log y = \log a + bx\log e$ ……………(2)

Let $Y = \log y\,; B = \log a\,; A = b\log e$

∴ (2) between $Y = Ax + B$

This is linear equation in x and Y whose normal equations are,

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

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

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

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

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

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

Problem 7. Fit the 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)

$\therefore \log y = \log a + x \log b$ …………..(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

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

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

$28A = 2.1491$ ……………….(1)

$7B = 17.9363$ ……………….(2)

Solving the above equations we get $A = 0.0768 \ \ B = 2.5623$

$\therefore b = \text{antilog}\, A = \text{antilog}\, 0.0768 = 1.19$ (approximately)

$a = \text{antilog}\, B = \text{antilog}\, 2.5623 = 365.01$ (approximately)

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

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

EXERCISE QUESTIONS:

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

(i)

$\boldsymbol{x}$	0	1	2	3	4
$\boldsymbol{y}$	1	1.8	3.3	4.5	6.3

(ii)

Years $\boldsymbol{x}$	1911	1921	1931	1941	1951
Production in tons $\boldsymbol{y}$	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

(iii)

x	10	15	20	25	30	35	40
y	11	13	16	20	27	34	41

(iv)

x	1	2	3	4	5	6	7
y	2.3	5.2	9.7	16.5	29.4	35.5	54.4

3. 11fit a curve $y = ae^{bx}$ 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	5.02	10	31.62

6. Fit the exponential curve $y = ae^{bx}$ for the following data.

No. of petals	5	6	7	8	9	10
No. of flowers	133	55	23	7	2	2

3.2 REGRESSION

If there is a functional group 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 a straight line, it is called a **line of regression** between the two variables.

Regression line of y on x

Definition. If we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance *parallel to the y-axis* 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.**

Theorem 3.2.1 . The equation of the regression line of y on x is given by

$$\boldsymbol{y-\bar{y}=\gamma\frac{\sigma_y}{\sigma_x}(x-\bar{x})}$$

Proof. Let $y \;=\; ax+b$ be the line of regression of y on x.

According to the principle of least squares the constants a and b are to be determined in such a way that $S=\sum[y_i-(ax_i+b)]^2$ is minimum.

$$\frac{\partial S}{\partial a}=0 \Longrightarrow -2\sum(y_i-ax_i-b)x_i=0$$

$$\Longrightarrow -2\sum x_iy_i=a\sum x_i^2+b\sum x_i \qquad \ldots\ldots\ldots\ldots(1)$$

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

$$\Longrightarrow \sum y_i=a\sum x_i+nb \qquad \ldots\ldots\ldots\ldots(2)$$

Equations (1) and (2) are called normal equations.

From (2) we obtain $\bar{y}=a\bar{x}+b$ ....(3)

$\therefore$ *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. $a\sum X_i^2=\sum X_iY_i$.

$$\therefore \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\gamma\sigma_x\sigma_y}{\sigma_x^2} = \gamma\frac{\sigma_y}{\sigma_x}$$

∴ the required regression line (4) becomes $Y = \left(\gamma\frac{\sigma_y}{\sigma_x}\right)Y$

$$\therefore y - \bar{y} = \gamma\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Regression line of x on y

Definition. It we fit a straight line by the principle of least squares to the points of the scatter diagram in such a way that the sum of the squares of the distance *parallel to the x-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 x on y.**

Theorem 3.2.2. The equation of regression line x on y is given by

$$\boldsymbol{x - \bar{x} = \gamma\frac{\sigma_x}{\sigma_y}(y - \bar{y})}$$

Proof. Proof is similar to that of theorem 3.2.1.

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 3.2.3. correlation coefficient is the geometric mean between the regression coefficients. (i.e) $\boldsymbol{\gamma = \pm\sqrt{b_{xy}b_{yx}}}$,

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.

Theorem 3.2.4. If one of the regression coefficients is greater than unity the other is less than unity.

Proof. We have$b_{xy}c = \gamma^2 \leq 1$ so that $b_{xy}b_{yx} \leq 1$.

Hence, $b_{xy} > 1 \Longrightarrow b_{yx} < 1$.

Hence, the theorem.

Theorem 3.2.5. Arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.

Proof. Let b_{xy} and b_{xy} be the regression coefficients.

We have to prove $\frac{1}{2}\left(b_{xy} + b_{yx}\right) \geq \gamma$.

Now, $\frac{1}{2}\left(b_{xy} + b_{yx}\right) \geq \gamma \Leftrightarrow b_{yx} + b_{xy} \geq 2\gamma$

$$\Leftrightarrow \gamma\frac{\sigma_y}{\sigma_x} + \gamma\frac{\sigma_x}{\sigma_y} \geq 2\gamma$$

$$\Leftrightarrow \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2$$

$$\Leftrightarrow \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y \geq 0$$

$$\Leftrightarrow \left(\sigma_x - \sigma_y\right)^2 \geq 0.$$

This is always true. Hence the theorem.

Theorem 3.2.6. Regression coefficients are independent of the change of the origin but dependent on change of scale.

Proof. Let $u_i = \frac{x_i - A}{h}$ 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, $b_{yx} = \gamma_{xy}\frac{\sigma_y}{\sigma_x}\left(\frac{k\sigma_v}{h\sigma_u}\right) = \frac{k}{h}b_{uv}$ ……….(1)

Similarly, $b_{xy} = (k/h)b_{uv}$ ……….(2)

From (1) and (2) we observe that b_{yx} and b_{xy} depend upon the scales h and k but not on the origins A and B.

Theorem 3.2.7. the angle between two regression lines is given by $\boldsymbol{\theta = tan^{-1}\left[\left(\frac{\gamma^2 - 1}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]}$

Proof. The equation of lines of regression of y on x and x on y respectively are

$$y-\bar{y}=\gamma\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$x-\bar{x}=\gamma\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

(2) can also be written as $y-\bar{y}=\frac{1}{\gamma}\frac{\sigma_y}{\sigma_x}(x-\bar{x})$ ……………….(3)

$\therefore$ slopes of the two lines (1) and (2) are $\gamma\frac{\sigma_y}{\sigma_x}$ and $\frac{\sigma_y}{\gamma\sigma_x}$.

Let θ be the acute angle between the two lines of regression.

$$\therefore \tan\theta=\frac{\frac{\gamma\sigma_y}{\sigma_x}-\frac{\sigma_y}{\gamma\sigma_x}}{1+\left(\frac{\gamma\sigma_y}{\sigma_x}\right)\left(\frac{\sigma_y}{\gamma\sigma_x}\right)}$$

$$=\frac{\gamma^2-1}{\gamma}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

$$=\frac{1-\gamma^2}{\gamma}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right) \text{ (since}\gamma^2\leq 1 \text{ and } \theta \text{ is acute)}$$

$$\theta=\tan^{-1}\left[\left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right]$$

Note 1. The obtuse angle between the regression lines is given by $\tan^{-1}\left[\left(\frac{\gamma^2-1}{\gamma}\right)\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right]$.

Note 2. If $\gamma=0$ then $\tan\theta=\infty$. Hence $\theta=\pi/2$. Thus if the *two variables are correlated then the lines of regression are perpendicular to each other.*

Note 3. If $\gamma=\pm1$ then $\tan\theta=0$.

Hence, $\theta=0$ or π.

$\therefore$ 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 examinations be denoted by x and y respectively.

We have $\bar{x} = \frac{1}{10}$; $\sum x_i = 35$ and $\bar{y} = \frac{1}{10}$;$\sum y_i = 29$.

For the calculation of regression, we have the following table.

x_i	$x_i - 35$	$(x_i - 35)^2$	y_i	$y_i - 29$	$(y_i - 29)^2$	$(x_i - 35)(y_i - 29)$
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**

$$\sigma_x^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{1}{10}\sum(x_i - 35)^2 = 35.8$$

$$\sigma_y^2 = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{1}{10}\sum(y_i - 29)^2 = 59.8$$

$\sigma_x = 5.98$ and $\sigma_y = 7.73$.

$$\therefore \gamma = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} = \frac{324}{10 \times 5.98 \times 7.73}$$

$= \frac{324}{462.254} = 0.7$ (approximately)

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 is $x - 35 = 0.542(y - 29)$

(i.e.) $x = 0.542y + 19.282$ …………………..(2)

(1) and (2) are the required regression equations.

(ii) the most likely interval mark for the university mark of 25 is got from the regression equation of x on y by putting $y = 25$.

$$\therefore x = 0.524 \times 25 + 19.282 = 32.83.$$

(iii) the most likely university mark for the internal mark of 30 is got from the regression equation of y on x by putting $x = 30$.

$$\therefore y = 0.905 \times 30 - 2.675 = 24.475.$$

Problem 2. For the solved problem 1 of 3.1 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 for the given value of $x = 61$.

The regression line of y on x is given by $y - \bar{y} = \gamma\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$$\bar{x} = A + h\bar{u} = 63 + \left(\frac{-70}{10}\right) = 56.$$

$$\bar{y} = B + h\bar{v} = 60 + \left(\frac{-12}{10}\right) = 58.8.$$

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

$$= \left[\frac{980}{10} - \left(\frac{-70}{10}\right)^2\right]^{1/2}$$

$$= 7.$$

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

$$= \left[\frac{994}{10} - \left(\frac{-12}{10}\right)^2\right]^{1/2}$$

$$= 9.898.$$

Since the scale factor for u_i and v_i is one we note $\sigma_x = \sigma_u$ and $\sigma_y = \sigma_v$.

We have $\gamma = 0.6$ for the solved problem1 of 6.1.

∴ Regression equation of y on x is $y - 58.8 = 0.6\left(\frac{9.898}{7}\right)(x - 56)$

$$\therefore 7y = 5.9388x + 79.0272$$

When $x = 61, y = 5.9388 \times 61 + 79.0272 = 441.294$

∴ $y = 63$ (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 equations can be written as $x = -2y + 5$ and $y = -\frac{2}{3}x + \frac{8}{3}$.

Hence, the two regression coefficients $b_{yx} = -\frac{2}{3}$ and $b_{yx} = -2$

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 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 coefficients as $b_{yx} = -\frac{3}{2}$ and $b_{xy} = -\frac{1}{6}$.

Now, $\gamma^2 = \left(-\frac{3}{2}\right) \times \left(-\frac{1}{6}\right) = \frac{1}{4}$.

$$\therefore\ \gamma = \pm\frac{1}{2}$$

Since, both the regression coefficients are negative we take $\gamma = -\frac{1}{2}$.

(iii) Given $\sigma_x = 5$

We have $b_{yx} = \frac{\gamma\sigma_y}{\sigma_x}$

$$\therefore -\frac{3}{2} = \left(-\frac{1}{2}\right)\left(\frac{\sigma_y}{5}\right)$$

$$\therefore \sigma_y = 15.$$

Problem 5. If $x = 4y + 5$ and $y = kx + 4$ are regression lines of x on y and y on x respectively (i) show that $0 \le k \le 1/4$ (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, $b_{xy}b_{yx} = \gamma^2 \Longrightarrow 4k = \gamma^2$.

Now, $0 \le \gamma^2 \le 1 \Longrightarrow 0 \le 4k \le 1$

$\Longrightarrow 0 \le k \le 1/4$.

(ii) Given $k = 1/8$. Hence $b_{yx} = 1/8$.

Hence, $\gamma = \frac{1}{\sqrt{2}}$. (Positive value of γ is taken since both regression coefficients are positive)

Let $\bar{x}$ and $\bar{y}$ be the means of the two variables x and y.

Since the regression lines pass through $(\bar{x}, \bar{y})$ we have $\bar{x} = 4\bar{y} + 5$ and $\bar{y} = \frac{1}{8}\bar{x} + 4$. (Taking k=1/8)

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 $y = -\frac{a}{b}x - \frac{c}{b}$ we get the regression coefficient of y on x is $b_{yx} = -\frac{a}{b}$.

Writing $ax + by + c = 0$ in the form $y = -\frac{b}{a}x - \frac{c}{a}$ we get the regression coefficient of x on y is $b_{xy} = -\frac{b}{a}$.

Now, $\gamma^2 = b_{yx}b_{xy} = \gamma^2 = -\left(\frac{a}{b}\right)\left(\frac{b}{a}\right)$

Suppose a and b are of same sign. Then $\gamma^2 = 1$

Hence $\gamma = -1$(sinceb_{yx} and b_{xy} are negative).

Suppose a and b are of opposite sign. Then $\gamma^2 = 1$

Hence $\gamma = 1$(since b_{yx} and b_{xy} are negative).

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 $\sigma_x^2 + \sigma_y^2 < 2\sigma_x\sigma_y$.

(i.e.) $\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y < 0$

(i.e.) $(\sigma_x - \sigma_y)^2 < 0$ 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 $\tan\theta \leq \left(\frac{1-\gamma^2}{\gamma}\right)\left(\frac{1}{2}\right)$

$\therefore \tan\theta \leq \left(\frac{1-\gamma^2}{\gamma}\right)$. Hence $\sin\theta \leq \left(\frac{1-\gamma^2}{1+\gamma^2}\right)$

$$\therefore \sin\theta \leq 1 - \gamma^2.$$

EXERCISE QUESTIONS:

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 the 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 womwn whose age is 45 years.

Age (x)	**Blood pressure (y)**	**Age (x)**	**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

Also, calculate regression line of y on x and predict the value of y when $x = 9$.

4. The following data relate to the ages of husbands and wives.

Ages of husband	26	29	31	33	35	34	38	39	41	45
Ages of wives	22	26	27	31	28	19	29	36	35	46

Obtain the regression equations and determine.
   (i) The most likely age of husband for age of wife 30 years.
   (ii) The most likely age of wife for age of husband 32 years.

5. Given the following data find the expected value of x when $y = 12$.
   $\bar{x} = 25;\ \bar{y} = 22;\ \sigma_x = 4; \sigma_y = 5$ and $\gamma = 0.8$.

6. The following data regarding the heights (x) and weights (y) of 100 college students are given as $\sum x = 15000; \sum x^2 = 2272500; \sum y = 6800; \sum y^2 = 463025$ and $\sum xy = 102250$. Find the equation to the regression line of height on weight.

7. Give the following results for the heights (x) and weights (y) of 1000 policemen of Tamil Nadu.

Mean height = 68 inches, mean weight = 150 lbs; $\sigma_x = 2.5$ inches and $\sigma_y = 30$ lbs. Estimate from the above data (i) the height of a particular policeman whose weight is 200 lbs. (ii) the weight of a particular policeman whose height is 5 feet.

8. The average daily wage for working class in madras is Rs. 12 and for that in Delhi is Rs.18; their respective standard deviations are Rs. 2 and Rs.3 and the coefficient of correlation is 0.67. find the most likely wage in Delhi corresponding to the wage of Rs. 20 in adras.
9. In a partially destroyed laboratory record of an analysis of correlation data the following results only are legible. Variance of $x = 25$. Regression equations: $3x + 2y - 26 = 0$ and $6x + y - 51 = 0$. Find (i) the man values of x and y. (ii) the standard deviation of x and y. (iii) the coefficient of correlation between x and y.
10. Given the equations of the two regression lines $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Decide which is the regression of y on x.
11. Find σ_y and γ from the following data. Regression lines $3x = y; 4y = 3x$ and $S.D$ of $x = 2$.
12. The equations of two regression lines obtained in a correlation analysis are $4x - 5y + 33 = 0$ and $20x - 9y - 107 = 0$. If the variance of y is 16. Find (i) the mean values of x and y. (ii) the correlation coefficient between x and y. (iii) the standard deviation of x.
13. From the regression lines $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$ find the value of (a) mean if $x's$ (b) mean of $y's$. (c) the correlation coefficient between x and y.
14. From a partially destroyed laboratory record only the following are available, $x = 4y + 5$ and $y = kx + 4$ are the regression lines of x on y and y on x respectively. Show that $0 \leq k \leq 1/4$. If $k = 1/16$ find the means of the two variables x and y and also find the correlation between them.
15. From the following regression equations find the mean values of x and y; $3x + 12y = 19; 3y + 9x = 46$.
16. The regression lines of two variables x and y are $x = 1.4y - 12.3$ and $y = 0.6x + 32.6$. find the arithmetic means of x and y and also the correlation coefficient between x and y.
17. The regression lines of y on x and x on y are respectively $y = ax + b$ and $x = cy + d$. Show that (i) $\bar{x} = \frac{bc+d}{1-ac}$ and $\bar{y} = \frac{ad+c}{1-ac}$ (ii) $\frac{\sigma_x}{\sigma_y} = \sqrt{\left(\frac{c}{a}\right)}$ (iii) $\gamma_{xy} = \sqrt{ac}$.

UNIT IV

THEORY OF ATTRIBUTES

UNIT STRUCTURE

4.1 Attributes

4.2 Consistency of Data

4.3 Independence and Association of Data

INTRODUCTION

Statistics chiefly deals with collection of data, classification of data based on certain characteristics, calculation of statistical constants such as mean, median, mode, standard deviation etc., and analysis of data based on the statistical constants. The characteristics used for classification of data may be quantitative 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 where as 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 qualities. 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.

4.1 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 attribute and the **negative class** denotes the absence of the attribute. Capital roman letters such as A, B, C, D, … 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 A Represents the attribute non-richness (poor).

The combinations of attributes are denoted by grouping together the letters concerned.

For example, if attribute 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 wealth; $\alpha\beta$ represents non-health and non-wealth. A convenient way of representing two attributes in a 2×2 table is as follows.

Attribute	$\boldsymbol{B}$	$\boldsymbol{\beta}$
$\boldsymbol{A}$	AB	$A\beta$
$\boldsymbol{\alpha}$	αB	$\alpha\beta$

A class represented by n attributes is called a **class** ofn^{th} order.

For example, $A, B, C, \alpha, \beta, \gamma$ are all set of first order; $AB, A\beta, \alpha B, \alpha\beta$ are of second order, and $ABC, A\beta\gamma, A\beta C, \alpha\beta\gamma$ are of the third order.

The number of individuals possessing the attributes in a class of n^{th} order is called a **class frequency** of order n and class frequencies are denoted by bracketing the attributes.

Thus, (A) stands for the frequency of A the number of individuals possessing the attribute A and $(A\beta)$ stands for the number of individuals possessing the attributes A and not B.

Note 1. Class frequencies of the type $(A), (AB), (ABC), \ldots$ are known as**positive class frequencies.**

Class frequencies of the type $(\alpha), (\beta), (\alpha\beta), (\alpha\beta\gamma), \ldots$ are known as **negative class frequencies.**

Class frequencies of the type $(\alpha B), (A\beta), (A\beta\gamma), (\alpha\beta C), \ldots$ are known as **contrary class 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 then 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.

Attribute	$\boldsymbol{B}$	$\boldsymbol{\beta}$	**Total**
$\boldsymbol{A}$	(AB)	$(A\beta)$	(A)
$\boldsymbol{\alpha}$	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

N denotes the total number in the population.

In a population of size N, the relation between the class frequencies of various orders are given below.

$$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma)\ etc.,$$

$$\left.\begin{aligned}(A) &= (AB) + (A\beta)\\ (B) &= (AB) + (\alpha B)\end{aligned}\right\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$\left.\begin{aligned}(\alpha) &= (\alpha B) + (\alpha\beta)\\ (\beta) &= (A\beta) + (\alpha\beta)\end{aligned}\right\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$\left.\begin{aligned}N &= (A) + (\alpha)\\ N &= (B) + (\beta)\end{aligned}\right\} \Rightarrow N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

For three attributes A, B and C we get similar results as shown below.

$(A) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma).$

$(B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma).$

$(C) = (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C).$

$(AB) = (ABC) + (AB\gamma)$

$(A\beta) = (A\beta C) + (A\beta\gamma).$

$(\alpha B) = (\alpha BC) + (\alpha B\gamma).$

$(\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma) etc$

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 frequencies upto 3 attributes.

Order	**Attribute**	**Class frequencies of all orders**	**Number in each order**	**Total number**
0		N		1
	A			
0		N	1	3
1		$(A),(\alpha)$	2	
	A,B			
0		N	1	
1		$(A),(B),(\alpha),(\beta)$	4	9
2		$(AB),(A\beta),(\alpha B),(\alpha\beta)$	4	
	A,B,C			
0		N	1	
1		$(A),(B),(C),(\alpha)(\beta),(\gamma)$	6	
2		$(AB),(A\beta),(\alpha B),(\alpha\beta)$		27
		$(AC),(A\gamma),(\alpha C),(\alpha\gamma)$	12	
		$(BC),(B\gamma),(\beta C),(\beta)$		
3		$(ABC),(AB\gamma),(A\beta C),(A\beta\gamma)$		
		$(\alpha BC),(\alpha B\gamma),(\alpha\beta C),(\alpha\beta\gamma)$	8	

The classes of highest order are called the **ultimate classes** and their frequencies are called the **ultimate frequencies.**

Theorem 4.1. 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 frequencies of order r is $\binom{n}{r}2^r$.

Thus the total number of frequencies (of all orders).

$$=\sum_{r=0}^{n}\binom{n}{r}2^r = 1+\binom{n}{1}2+\binom{n}{2}2^2+\cdots+\binom{n}{n}2^n$$

$$=(1+2)^n = 3^n.$$

(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)

$$=\sum_{r=0}^{n}\binom{n}{r} = 1+\binom{n}{1}+\binom{n}{2}+\cdots+\binom{n}{n}$$

$$=(1+1)^n = 2^n.$$

(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)

$$=\sum_{r=0}^{n}\binom{n}{r} = 2^n - 1.$$

Dichotomisationis the process of diving a collection of objects into two classes according to the possession or non-possession iof an attribute.

Suppose a population consists of N objects. If A is an attribute we have $N=(A)+(\alpha)$. Symbolically we write, $(A)=A.N$ and $(\alpha)=\alpha.N$.

Now, $N=(A)+(\alpha)$.

$\Longrightarrow N = A.N+\alpha.N$

$=(A+\alpha).N.$

$\Longrightarrow 1 = A+\alpha.$

Thus, in symbolic expression A can be replaced by $1-\alpha$ and α 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 (refer 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 example 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

$$N = (A) + (\alpha) = (B) + (\beta)$$

Hence, $N = (A) + (\alpha) = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$

And $N = (B) + (\beta) = (AB) + (\beta B) + (A\beta) + (\alpha\beta)$

If there are three attributes A, B, C we have $N = (A) + (\alpha)$.

$\Longrightarrow N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$. Thus

$$N = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$$

3. Consider two attributes A and B.

Now, $(\alpha\beta) = \alpha\beta. N = (1 - A)(1 - B). N.$

$= (1 - A - B + AB). N = N - A. N - B. N + AB. N$

$= N - (A) - (B) + (AB).$

Here negative class frequency has been expressed in terms of positive class frequencies.

4. $(AB) = AB. N.$

$$= (1 - \alpha)(1 - \beta). N = (1 - \alpha - \beta + \alpha\beta). N.$$

$$= N - \alpha. N - \beta. N + \alpha\beta. N = N - (\alpha) - (\beta) + (\alpha\beta).$$

Here positive class frequency has been expressed in terms of negative class frequencies.

5. $(\alpha\beta\gamma) = \alpha\beta\gamma. N = (1 - A)(1 - B)(1 - C). N.$

$$= N - A. N - B. N - C. N + AB. N + AC. N + BC. N - ABC. N.$$

$$= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC).$$

Note. $N = (A) + (B) + (C) - (AB) - (AC) - (BC) + (ABC) + (\alpha\beta\gamma).$

Solved Problems.

Problem 1. Given $(\boldsymbol{A}) = \mathbf{30}; (\boldsymbol{B}) = \mathbf{25}; (\boldsymbol{\alpha}) = \mathbf{30}; (\boldsymbol{\alpha\beta}) = \mathbf{20}.$

Find (i) N (ii) $(\boldsymbol{\beta})$ ***(iii)*** $(\boldsymbol{AB})$ ***(iv)*** $(\boldsymbol{A\beta})$ ***(v)*** $(\boldsymbol{\alpha B})$.

Solution. (i)$N = (A) + (\alpha) = 30 + 30 = 60.$

(ii) $(\beta) = N - (B) = 60 - 25 = 35.$

(iii) $(AB) = AB.N = (1-\alpha)(1-\beta).N.$

$= N - (\alpha) - (\beta) + (\alpha\beta)$

$= 60 - 30 - 35 + 20 = 15$

(iv) $(A\beta) = A\beta.N = A(1-B).N = (A) - (AB).$

$= 30 - 15 = 15.$

(v) $(\alpha B) = \alpha B.N. = (1-A)B.N = (B) - (AB).$

$= 25 - 15 = 10.$

Note. The result can also be got directly by completing the 2 × 2 contingency table for the attributes A and B.

	$(\boldsymbol{B})$	$(\boldsymbol{\beta})$	
$(\boldsymbol{A})$	-	-	30
$(\boldsymbol{\alpha})$	-	20	30
	25	-	-

Problem 2.

Given the following ultimate class frequencies of two attributes A and B. Find the frequencies of positive and negative class frequencies and the total number of observations,

$$(\boldsymbol{AB}) = \mathbf{975}; (\boldsymbol{\alpha B}) = \mathbf{100}; (\boldsymbol{A\beta}) = \mathbf{25}; (\boldsymbol{\alpha\beta}) = \mathbf{950}.$$

Solution. Positive class frequencies are (A) and (B).

$$(A) = (AB) + (A\beta) = 975 + 25 = 1000$$

$$(B) = (AB) + (\alpha B) = 975 + 100 = 1075$$

Negative class frequencies are (α) and (β).

$$(\alpha) = (\alpha B) + (\alpha\beta) = 100 + 950 = 1050$$

$$(\beta) = (A\beta) + (\alpha\beta) = 25 + 950 = 975$$

$$N = (A) + (\alpha) = (B) + (\beta)$$

Taking $N = (A) + (\alpha) = 1000 + 1050 = 2050.$

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 $\boldsymbol{N = 20; (A) = 9; (B) = 12; (C) = 8; (AB) = 6; (BC) = 4; (CA) = 4; (ABC) = 3}$.

Solution. There are three attributes A, B, C.

$\therefore$ the total number of class frequencies is $3^3 = 27$.

We are given only 8 class frequencies and we have to find the remaining 19 class frequencies. They are

Order 1. $(\alpha) = N - (A) = 20 - 9 = 11.$

$(\beta) = N - (B) = 20 - 12 = 8.$

$(\gamma) = N - (C) = 20 - 8 = 12.$

Order 2. $(A\beta) = A(1 - B).N = (A) - (AB) = 9 - 6 = 3.$

$(\alpha B) = (1 - A)B.N = (B) - (AB) = 12 - 6 = 6.$

$(A\gamma) = A(1 - C).N = (A) - (AC) = 9 - 4 = 5.$

$(\alpha C) = (1 - A)C.N = (C) - (AC) = 8 - 4 = 4.$

$(B\gamma) = B(1 - C).N = (B) - (BC) = 12 - 4 = 8.$

$(\beta C) = (1 - B)C.N = (C) - (BC) = 8 - 4 = 4.$

$(\alpha\beta) = (1 - A)(1 - B).N = N - (A) - (B) + (AB)$

$= 20 - 9 - 12 + 6 = 5.$

$(\beta\gamma) = (1 - B)(1 - C).N = N - (B) - (C) + (BC)$

$= 20 - 12 - 8 + 4 = 4.$

$(\alpha\gamma) = (1 - A)(1 - C).N = N - (A) - (C) + (AC)$

$= 20 - 9 - 8 + 4 = 7.$

Order 2. $(AB\gamma) = AB(1 - C).N = (AB) - (ABC) = 6 - 3 = 3.$

$(A\beta C) = A(1 - B)C.N = (AC) - (ABC) = 4 - 3 = 1.$

$(A\beta\gamma) = A(1 - B)(1 - C).N = (A) - (AC) - (AB) + (ABC)$

$= 6 - 3 = 3.$

$(\alpha BC) = (1 - A)BC.N = (BC) - (ABC)$

$= 4 - 3 = 1.$

$(\alpha B\gamma) = (1 - A)(1 - C)B.N = (B) - (BC) - (AB) + (ABC)$

$= 12 - 4 - 6 + 3 = 5.$

$(\alpha\beta C) = (1 - A)(1 - B)C.N = (C) - (AC) - (BC) + (ABC)$

$= 8 - 4 - 4 + 3 = 3.$

$(\alpha\beta\gamma) = (1-A)(1-B)(1-C).N = N-(A)-(B)-(C)$

$+(AB)+(BC)+(CA)-(ABC).$

$= 20-9-12-8+6+4+4-3 = 2.$

Problem 4. In a class test in which 135 candidates were examined for proficiency in English and Maths. It was discovered that 75 students failed in English, 90 failed in Maths and 50 failed in both. Find how many candidates (i) have passed in Maths. (ii) have passed in English, failed in Maths. (iii) have passed in both.

Solution. Let A denotes pass in English and B denotes pass in Maths.

$\therefore \alpha$ denotes fail in English and β denotes fail in Maths.

Given $(\alpha) = 75; (\beta) = 90; (\alpha\beta) = 50; N = 135.$

We have to find (i) (B) (ii) $(A\beta)$ (iii) (AB)

(i) $(B) = N - (\beta) = 135 - 90 = 45.$

(ii) Consider $(\beta) = (A\beta) + (\alpha\beta)$

$\Rightarrow (A\beta) = (\beta) - (\alpha\beta)$

$= 90 - 50 = 40.$

(iii) $(AB) = (1-\alpha)(1-\beta).N = N - (\alpha) - (\beta) + (\alpha\beta).$

$= 135 - 75 - 90 + 50 = 20.$

Problem 5. Given $\boldsymbol{N = 1200; (ABC) = 600; (\alpha\beta\gamma) = 50; (\gamma) = 270; (A\beta) = 36; (B\gamma) = 204;}$
$\boldsymbol{(A) - (\alpha) = 192; (B) - (\beta) = 620.}$ ***Find the remaining ultimate class frequencies.***

Solution. Since there are three attributes there are $2^3 (= 8)$ ultimate class frequencies. We are given two. Hence we have to find the remaining six. They are (i) $(AB\gamma)$ (ii) $(A\beta C)$ (iii) (αBC) (iv) $(A\beta\gamma)$(v) $(\alpha B\gamma)$ and (vi) $(\alpha\beta C)$

To find them we need the frequencies of the positive classes; $(A), (B), (C), (AB), (BC), (AC).$

First order: $(A) - (\alpha) = 192.$

$(A) + (\alpha) = 1200 (= N).$

Adding we get $2(A) = 1392.$

Hence $(A) = 696.$

$(B) - (\beta) = 620$

$(B)+(\beta)=1200(=N)$

Hence $(B)=910$.

Now, $(C)=N-(\gamma)$

$=1200-270$

$=930$.

Second order:

$(AB)=(A)-(A\beta)=696-36=660.$

$(BC)=(B)-(B\gamma)=910-204=706.$

We have $N=(A)+(B)+(C)-(AB)-(BC)-(AC)+(ABC)+(\alpha\beta\gamma)$

$\Longrightarrow (AC)=(A)+(B)+(C)-(AB)-(BC)+(ABC)+(\alpha\beta\gamma)$

$=696+910+930-660-706+600+50=620.$

Third order:

(i) $(AB\gamma)=AB(1-C).N=(AB)-(ABC)=660-600=60.$

(ii) $(A\beta C)=A(1-B)C.N=(AC)-(ABC)=620-600=20.$

(iii) $(\alpha BC)=(1-A)BC.N=(BC)-(ABC)=706-600=106.$

(iv) $(A\beta\gamma)=A(1-B)(1-C).N=(A)-(AC)-(AB)+(ABC)$

$=696-660-620+600=16.$

(v) $(\alpha B\gamma)=(1-A)(1-C)B.N=(B)-(BC)-(AB)+(ABC)$

$=910-660-706+600=144.$

(vi) $(\alpha\beta C)=(1-A)(1-B)C.N=(C)-(AC)-(BC)+(ABC)$

$=930-620-706+600=204.$

Problem 6. Given that $(\boldsymbol{A})=(\boldsymbol{\alpha})=(\boldsymbol{B})=(\boldsymbol{\beta})=\frac{\boldsymbol{N}}{\boldsymbol{2}}$.

Show that (i) $(\boldsymbol{AB})=(\boldsymbol{\alpha\beta})$ ***(ii)*** $(\boldsymbol{A\beta})=(\boldsymbol{\alpha B})$.

Solution. (i) $(AB)=AB.N=(1-\alpha)(1-\beta).N.$

$=N-(\alpha)-(\beta)+(\alpha\beta)$

$=N-\frac{N}{2}-\frac{N}{2}+(\alpha\beta)=(\alpha\beta)$

$\therefore (AB)=(\alpha\beta).$

(ii) $(A\beta)=A\beta.N=A(1-\alpha)(1-B).N$

$=N-(\alpha)-(B)+(\alpha B).$

$=N-\frac{N}{2}-\frac{N}{2}+(\alpha B)=(\alpha B)$

$$\therefore (A\beta) = (\alpha B).$$

Problem 7. Of 500 men in a locality exposed to cholera 172 in all were attacked; 178 were inoculated and of these 128 were attacked. Find the number of persons (i) not inoculated not attacked (ii) inoculated not attacked. (iii) not inoculated attacked.

Solution. Denote the attribute A as 'attacked' and the attribute B as 'inoculated'. Hence α denotes "not attacked"; β denotes "not inoculated".

Given $N = 500; (A) = 172; (B) = 178; (AB) = 128;$

To find (i) $(\alpha\beta)$ (ii) (αB) (iii) $(A\beta)$.

(i) $(\alpha\beta) = \alpha\beta.N = (1 - A)(1 - B).N.$

$= N - (A) - (B) + (AB)$

$= 500 - 172 - 178 + 128 = 278.$

(ii) $(\alpha B) = \alpha B.N = (1 - A)B.N.$

$= (B) - (AB) = 178 - 128 = 50.$

(iii) $(A\beta) = A\beta.N = A(1 - B).N.$

$= (A) - (AB)$

$= 172 - 128 = 44.$

Problem 8. There were 200 students in a college whose results in the first semester, second semester and third semester are as follows.

80 passed in the first semester; 75 passed in the second semester.

96 passed in the third semester; 25 passed in all the three semesters.

46 failed in all the three semesters; 29 passed the first two and failed in the third semester.

42 failed in the first two semester but passed in the third semester.

Find how many students passed in at least two semesters.

Solution. Denoting "pass in first semester" as A "pass in second semester" as B and "pass in third semester" as C we get

$N = 200; (A) = 80; (C) = 96; (ABC) = 25; (\alpha\beta\gamma) = 46; (AB\gamma) = 29; (\alpha\beta C) = 42$

We have to find $(AB\gamma) + (\alpha BC) + (A\beta C) = (ABC.)$

Consider $(C) = (AC) + (\alpha C).$

$= (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C).$

$$\therefore (ABC)+(A\beta C)+(\alpha BC)=(C)-(\alpha\beta C)=96-42=54.$$

$$\therefore (ABC)+(A\beta C)+(\alpha BC)+(AB\gamma)=54+29=83.$$

Thus the number of students who passed is atleast two semesters is 83.

Problem 9. Given $(\boldsymbol{ABC})=\mathbf{149}; (\boldsymbol{AB\gamma})=\mathbf{738}; (\boldsymbol{A\beta C})=\mathbf{225}; (\boldsymbol{A\beta\gamma})=\mathbf{1196}; (\boldsymbol{\alpha BC})=\mathbf{204}; (\boldsymbol{\alpha B\gamma})=\mathbf{1762}; (\boldsymbol{\alpha\beta C})=\mathbf{171}; (\boldsymbol{\alpha\beta\gamma})=\mathbf{21842}$.

Find $(\boldsymbol{A}),(\boldsymbol{B}),(\boldsymbol{C}),(\boldsymbol{AB}),(\boldsymbol{AC}),(\boldsymbol{BC})$ and $\boldsymbol{N}$.

Solution. $N=(ABC)+(AB\gamma)+(A\beta C)+(A\beta\gamma)+(\alpha BC)+(\alpha B\gamma)+(\alpha\beta C)+(\alpha\beta\gamma)$

$=149+738+225+1196+204+1762+11+21842$

$=26287$

$(A)=(ABC)+(AB\gamma)+(A\beta C)+(A\beta\gamma)$

$=149+738+225+1196$

$=2308$

$(B)=(ABC)+(AB\gamma)+(\alpha BC)+(\alpha B\gamma)$

$=149+738+204+1762$

$=2853$

$(C)=749$ (Verify)

$(AB)=(ABC)+(AB\gamma)=149+738=887$

$(AC)=(ABC)+(A\beta C)=149+225=374$

$(BC)=353$ (Verify)

Problem 10. Show that for $\boldsymbol{n}$ attributes $\boldsymbol{A_1,A_2,\ldots,A_n}$.

$$(\boldsymbol{A_1A_2\ldots.A_n})\geq(\boldsymbol{A_1})+(\boldsymbol{A_2})+\cdots+(\boldsymbol{A_n})-(\boldsymbol{n}-\mathbf{1})\boldsymbol{N}$$

Where N is the total number of observations.

Solution. We prove this by induction on n.

$(\alpha_1\alpha_2)=\alpha_1\alpha_2.N=(1-A_1)(1-A_2).N.$

$=N-(A_1)-(A_2)+(A_1A_2).$

Since $(\alpha_1\alpha_2)\geq 0$, we have $(A_1A_2)-A_1-A_2+N\geq 0$.

$\therefore (A_1A_2)\geq A_1+A_2-N$ ……………….(1)

Hence the result is true for $n=2$.

We now assume that the result is true for $n=k$(i.e., for k attributes) so that $(A_1A_2\ldots.A_n)\geq(A_1)+(A_2)+\cdots+(A_n)-(k-1)N$.

Replacing the attribute A_k by another compound attribute $A_k A_{k+1}$. We get

$(A_1 A_2 \ldots . A_k A_{k+1}) \geq (A_1) + (A_2) + \cdots + (A_{k+1}) + (A_k A_{k+1}) \ - (k-1)N$

$\geq (A_1) + (A_2) + \cdots + (A_{k-1}) + [(A_k) + (A_{k+1} - N)] - (k-1)N \ \ (by\ (1))$

$= (A_1) + (A_2) + \cdots + (A_k) + (A_{k+1}) - kN.$

Thus, $(A_1 A_2 \ldots . A_n) \geq (A_1) + (A_2) + \cdots + A_{k+1} - (\overline{k-1} - 1)N$.

Hence, the result is true for $n = k+1$ and by induction the result is true for all positive integers n.

Problem 11. In a very hotly fought battle 70% of the soldiers atleast lost an eye, 75% atleast lost an ear, 80% atleast lost an arm and 85% atleast lost a leg. How many at least must have lost all the four?

Solution. Denoting 'loosing an eye' by A, 'loosing an ear' by B, 'loosing an arm' by C and 'loosing a leg' by D we have

$$N = 100; (A) \geq 70\ (B) \geq 75 (C) \geq 80 (D) \geq 85.$$

To find the least value of $(ABCD)$.

By solving problem 10 $(ABCD) \geq (A) + (B) + (C) + (D) - 3N$

$$\geq 70 + 75 + 80 + 85 - 300$$

$$= 10.$$

$$\therefore (ABCD) \geq 10.$$

∴ At least 10% of the soldiers lost all the four.

Problem 12. A company produces tube lights and conducts a test on 5000 lights for production defects of frame (F), chokes (C), starters (S), and tubes (T). The following are the records of defects.

$$\boldsymbol{(F) = 130; (C) = 120; (S) = 115; (T) = 86; (FC) = 100;}$$

$$\boldsymbol{(CS) = 130; (ST) = 75; (FT) = 60; (CT) = 54; (FS) = 37;}$$

$$\boldsymbol{(FCS) = 90; (CST) = 85; (FST) = 112; (FCT) = 180; (FCST) = 5.}$$

Find the percentage of the tube lights which pass all the four tests.

Solution. Number of tube lights passing the four tests

$= (1-F)(1-C)(1-S)(1-T).N.$

$$= [1 - (F + C + S + T) + (FC + CS + ST + FT + CT + FS)$$

$$-(FCS + CST + STF + FCT) + FCST].N.$$

$$= N - [(F) + (C) + (S) + (T)] + [(FC) + (CS) + (ST) + (FT) + (CT)$$

$$+(FS)] - [(FCS) + (FCT) + (FST) + (CST)] + (FCST).$$

$$= 5000 - (130 + 120 + 115 + 86) + (100 + 130 + 75 + 60 + 54 + 37)$$

$$-(90 + 108 + 112 + 85) + 5.$$

$$= 5000 - 451 + 456 - 395 + 5$$

$$= 5461 - 846$$

$$= 4615.$$

∴ Out of 5000 tube lights 4615 pass the four tests for defects.

∴ Percentage of tube lights which pass the four tests $= \frac{4615}{5000} \times 100.$

$= 92.3$ %.

EXERCISE QUESTIONS:

1. Given the frequencies $(A) = 1150; (\alpha) = 1120; (AB) = 1075; (\alpha\beta) = 985.$ 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.

$$(AB) = 733; (A\beta) = 840; (\alpha B) = 699; (\alpha\beta) = 783.$$

3. Given the following data. Find the frequencies of

(i) The remaining positive classes (ii) ultimate classes

$$N = 1800; (A) = 850; (B) = 780; (C) = 326;$$

$$(AB\gamma) = 200; (A\beta C) = 94; (\alpha BC) = 72; (ABC) = 50.$$

4. Given the following ultimate class frequencies. Find the frequencies of the positive classes.

$$(ABC) = 298; (A\beta C) = 450; (\alpha BC) = 408; (\alpha\beta C) = 342;$$

$$(AB\gamma) = 1476; (A\beta\gamma) = 2292; (\alpha B\gamma) = 3524; (\alpha\beta\gamma) = 43684.$$

5. A survey reveals that out of 1000 people in a locality 800 like coffee; 700 like tea; 660 like both coffee and tea. Find how many people like neither coffee nor tea.
6. A social survey by parents association in Palayamkottai revealed the following results. 40% liked john's college, 39% liked hindu college, 48% liked Xavier's college; 10% liked all the three; 9% liked John's college and Hindu college and liked Xavier's College. Find how many percent liked at least two colleges?

7. 100 children took three examinations. 40 passed the first, 39 passed the second and 48 passed the third, 10 passed all three. 21 failed all three 9 passed the first two failed the third, 19 failed the first two and passed the third. Find how many children passed atleast two examinations?

8. A survey conducted among 1800 T.V. viewers in a city revealed the following results. 850 see Doordharshan T.V. programmess; 780 see star T.V. programmess; 326 see cable T.V. progremmess; 50 see all the three progremmess; 200 see Door Dharsan T.V. progremmess and Star T.V. progremmess but not cable T.V. progremmess; 110 do not see Door Dharsan and star T.V. progremmess but cable T.V. progremmes

   (i) Find how many people see Door Dharsan and star T.V progremmess.

   (ii) Find how many people see at least two T.V. progremmes.

9. Given that $(A) = (\alpha) = (B) = (\beta) = (C) = (\gamma) = N/2$ and $(ABC) = (\alpha\beta\gamma)$ show that $2(ABC) = (AB) + (AC) + (BC) - N/2$.

10. An examination result shows the following data. 56% at least failed in Part I Tamil, 76% atleast failed in Part II English, 82% at least failed in Major-Chemistry and 88% at least failed in Ancillary Maths. How many at least failed in all the four?

11. In a university examination 95% of the candidates passed Part I, 70% passed Part II, 65% passed Part III. Find how many at least should have passed the whole examination.

12. A certain factory produces and tests 700 motor cars per year. The possible defects are Body work Chasis$[C]$, Engine $[E]$, Instrument $[I]$. A years record of defects is shown below.

$$(B) = 120; (C) = 150; (E) = 185; (I) = 200; (BC) = 100; (CE) = 110; (EI) = 28;$$

$$(BI) = 80; (CI) = 90; (BE) = 80; (BCE) = 24; (CEI) = 10; (BEI) = 5; (BCI) = 15;$$

$$(BCEI) = 2.$$

   Show that 96.9% of the cars pass all the four tests.

13. If $(A) = (\alpha) = (B) = (\beta) = N/2$ show that $(AB) = (\alpha\beta)$; $(\alpha\beta) = (\alpha B)$

14. In a competitive examination at which 600 graduates appeared, boys outnumbered girls by 96. Those qualifying for interview exceeded in number those failing to qualify by 310. The number of science graduate boys interviewed was 300 while among the arts graduate girls there were 25 who failed to qualify for interview.

Altogether there were only 135 arts graduates and 33 among them failed to qualify. Boys who failed to qualify numbered 18.

Find (i) the number of boys who qualified for interview. (ii) the total number of science graduate boys appearing (iii) the number of science graduate girls who qualified.

[*Hint:*Denote 'boys' by A; 'qualifying for interviews' by B; 'offering science' by C. Given: $N = 600; (A) - (\alpha) = 96; (B) - (\beta) = 310; (ABC) = 300; (\alpha\beta\gamma) = 25;$
$(\gamma) = 135; (\beta\gamma) = 33; (A\beta) = 18.$ To find (i)(AB) (ii) (AC) (iii) (αBC)].

15. The following data of eye color in father and sons are given:

Fathers with dark eyes and son with dark eyes 50; fathers with dark eyes and sons with not dark eyes 79; fathers with not dark eyes and sons with dark eyes 89; fathers with not dark eyes and sons with not dark eyes 782. Find

(i) The number of fathers with dark eyes.

(ii) The number of sons with not dark eyes.

4.2 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, $(A\beta) = (A) - (AB) = -5$, which is negative. This motivates 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 ultimate class frequencies, it follows that a set of independent class frequencies is consistent if and only if no ultimate class frequency is negative.

We have the following set of criteria for testing the consistency in the case of single attribute, two attributes and three attributes.

Attributes	Condition of consistency	Equivalent positive class conditions	Number of conditions
A	$(A) \geq 0$ $(\alpha) \geq 0$	$(A) \geq 0$ $(A) \leq N$ (Since $(\alpha) = (1 - A).N \geq 0$)	2
A, B	$(AB) \geq 0$ $(A\beta) \geq 0$ $(\alpha B) \geq 0$ $(\alpha\beta) \geq 0$	$(AB) \geq 0$ $(AB) \leq (A)$ $(AB) \leq (B)$ $(AB) \geq (A) + (B) - N$	2^2
A, B, C	$(ABC) \geq 0$ $(AB\gamma) \geq 0$ $(A\beta C) \geq 0$ $(\alpha BC) \geq 0$ $(A\beta\gamma) \geq 0$ $(\alpha B\gamma) \geq 0$ $(\alpha\beta C) \geq 0$ $(\alpha\beta\gamma) \geq 0$	(i) $(ABC) \geq 0$ (ii) $(ABC) \leq (AB)$ (iii) $(ABC) \leq (AC)$ (iv) $(\alpha BC) \leq (BC)$ (v) $(ABC) \geq (AB) + (AC) - (A)$ (vi) $(ABC) \geq (AB) + (BC) - (B)$ (vii) $(ABC) \geq (AC) + (BC) - (C)$ (viii) $(ABC) \leq (AB) + (BC) + (AC)$ $-(A) - (B) - (C) + (N)$	2^3

Note 1. In the case of 3 attributes conditions

(i) and (viii) $\Longrightarrow (AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$ ……………………(ix)

Similarly,

(ii) and (vii) $\Longrightarrow (AC) + (BC) - (AB) \leq (C)$ ……………………(x)

(iii) and (vi) $\Longrightarrow (AB) + (BC) - (AC) \leq (B)$ ……………………(xi)

(iv) and (v) $\Longrightarrow (AB) + (AC) - (BC) \leq (A)$ ……………………(xii)

When the class frequencies of first and second order alone are known.

Note 2. If the given data are incomplete so that it is not possible to determine all the class frequencies then the conditions of consistency can be used to determine the limits in which an unknown class frequency can lie.

Solved Problems.

Problem 1. Find whether the following data are consistent

$$\boldsymbol{N = 600; (A) = 300; (B) = 400; (AB) = 50.}$$

Solution. We calculate the ultimate class frequencies $(\alpha\beta), (\alpha B)$ and $(A\beta)$.

$$(\alpha\beta) = \alpha\beta.N = (1-A)(1-B).N = N-(A)-(B)+(AB)$$

$$= 600-300-400+50$$

$$= -50.$$

Since $(\alpha\beta) < 0$, the data are inconsistent.

Problem 2. Show that there is some error in the following data: 50% of people are wealthy and healthy, 35% are wealthy but not healthy, 20% healthy but not wealthy.

Solution. Taking 'wealth' as A and 'health' as B we get the following data.

$$N = 100; (AB) = 50; (A\beta) = 35; (\alpha B) = 20.$$

To check the consistency of data we find $(\alpha\beta)$.

$$(\alpha\beta) = \alpha\beta.N = (1-A)(1-B).N.$$

$$= N-(A)-(B)+(AB).$$

But $(A) = (AB) + (A\beta) = 50 + 35 = 85.$

$(B) = (AB) + (\alpha B) = 50 + 20 = 70.$

$$\therefore (\alpha\beta) = 100-85-70+50 = -5$$

$$\therefore (\alpha\beta) < 0.$$

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, B, C denote the attributes of speaking Tamil, Hindi, English respectively.

$\therefore$ given $N = 2000; (A) = 1854; (B) = 1507; (C) = 575;$

$(AB) = 676; (AC) = 286; (BC) = 270; (ABC) = 114.$

Consider $(\alpha\beta\gamma) = \alpha\beta\gamma.N$

$$= (1-A)(1-B)(1-C).N$$

$$= N-(A)-(B)-(C)+(AB)+(BC)+(AC)-(ABC)$$

$$= 2000-1854-1507-572+676+270+286-144$$

$$= -815.$$

$\therefore (\alpha\beta\gamma) < 0$. Hence the data are inconsistent.

$\therefore$ the information is incorrect.

Problem 4. Find the limits of (BC) for the following available data:
$\boldsymbol{N = 125; (A) = 48; (B) = 62; (C) = 45; (A\beta) = 7}$ ***and*** $\boldsymbol{(A\gamma) = 18}$.

Solution. First of all we find (AB) and (AC).

$$(AB) = (A) - (A\beta) = 48 - 7 = 14.$$

$$(AC) = (A) - (A\gamma) = 48 - 18 = 30.$$

Now, by condition of consistency (ix)

$$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - N.$$

$$\Rightarrow 41 + (BC) + 30 \geq 48 + 62 + 45 - 125.$$

$$\therefore (BC) \geq -41 \qquad \ldots\ldots\ldots\ldots\ldots(i)$$

Also using (xii), $(AB) + (AC) - (BC) \leq (A)$.

$$\Rightarrow (BC) \geq (AB) + (AC) - (A) = 41 + 30 - 48 = 23.$$

$$\therefore (BC) \geq 23 \qquad \ldots\ldots\ldots\ldots\ldots(ii)$$

Using (xi), $(AB) + (BC) - (AC) \leq (B)$.

$$\Rightarrow (BC) \leq (B) + (AC) - (AB) = 62 + 30 - 41 = 51.$$

$$\therefore (BC) \geq 51 \qquad \ldots\ldots\ldots\ldots\ldots(iii)$$

Using (x), $(AC) + (BC) - (AB) \leq (C)$.

$$\Rightarrow (BC) \leq (C) + (AB) - (AC) = 45 + 41 - 30 = 56.$$

$$\therefore (BC) \leq 56 \qquad \ldots\ldots\ldots\ldots\ldots(iv)$$

From (i), (ii), (iii) and (iv) we find $23 \leq (BC) \leq 56$.

Problem 5. Find the greatest and least values of (ABC) if $(A) = 50; (B) = 60; (C) = 80;$ $(AB) = 35; (AC) = 45$ and $(BC) = 42$.

Solution. The problem involves three 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.

$$\left.\begin{array}{l} (ABC) \leq (AB) \Rightarrow (ABC) \leq 35 \\ (ABC) \leq (BC) \Rightarrow (ABC) \leq 42 \\ (ABC) \leq (AC) \Rightarrow (ABC) \leq 35 \\ \text{Using (v), (vi) and (vii.)} \end{array}\right\} \Rightarrow (ABC) \leq 35 \qquad \ldots\ldots\ldots.(1)$$

$$(ABC) \geq (AB) + (AC) - (A) \Rightarrow (ABC) \geq 35 + 45 - 50 = 30.$$

$$(ABC) \geq (AB) + (BC) - (B) \Rightarrow (ABC) \geq 35 + 42 - 60 = 17.$$

$(ABC) \geq (AC) + (BC) - (C) \Rightarrow (ABC) \geq 45 + 42 - 80 = 7.$

$$\text{Thus,}\quad \left.\begin{array}{l}(ABC) \geq 30.\\(ABC) \geq 17.\\(ABC) \geq 7.\end{array}\right\} \Rightarrow (ABC) \geq 30 \qquad \ldots\ldots\ldots\ldots.(2)$$

From (1) and (2) we get $30 \leq (ABC) \leq 35$.

∴ the least value of (ABC) is 30 and the greatest value of (ABC) is 35.

Problem 6. If $\boldsymbol{\frac{(A)}{N} = x, \frac{(B)}{N} = 2x, \frac{(C)}{N} = 3x}$, ***and*** $\boldsymbol{\frac{(AB)}{N} = \frac{(AC)}{N} = \frac{(BC)}{N} = y}$ ***prove that neither*** $\boldsymbol{x}$ ***nor*** $\boldsymbol{y}$ ***can exceed***$\mathbf{1/4}$.

Solution. We observe that x and y are obviously positive integers.

The condition of consistency $(AB) \leq (A) \Rightarrow \frac{(AB)}{N} \leq \frac{(A)}{N}$.

$$\Rightarrow y \leq x.$$

Similarly,$(BC) \leq (B) \Rightarrow y \leq 2x.$ $\qquad \Rightarrow y \leq x \qquad$ ……..(1)

$(AC) \leq (C) \Rightarrow y \leq 3x.$

Now, $(AB) \geq (A) + (B) - N \Rightarrow \frac{(AB)}{N} \geq \frac{(A)}{N} + \frac{(B)}{N} - 1.$

Thus, $(AB) \geq (A) + (B) - (N) \Rightarrow y \geq 3x - 1.$

Similarly, $(BC) \geq (B) + (C) - (N) \Rightarrow y \geq 5x - 1 \Rightarrow y \geq 5x - 1$ $\qquad$ ……(2)

$(AC) \geq (A) + (C) - (N) \Rightarrow y \geq 4x - 1.$

By (1) and (2) $5x - 1 \leq y \leq x$.

Taking $5x - 1 \leq x$ we get $x \leq \frac{1}{4}$.

Taking $y \leq x$ we get $y \leq \frac{1}{4}$.

∴ Neither x nor y can exceed $1/4$.

<u>EXERCISE QUESTIONS:</u>

1. Examine the consistency of data when

(i) $(A) = 800, (B) = 700, (AB) = 660, N = 1000.$

(ii) $(A) = 600, (B) = 500, (AB) = 50, N = 1000.$

(iii) $N = 2100, (A) = 1000, (B) = 1300, (AB) = 1100.$

(iv) $N = 100, (A) = 45, (B) = 55, (C) = 50, (AB) = 15, (BC) = 25,$
$(AC) = 20, (ABC) = 12.$

(v) $N = 1800, (A) = 850, (B) = 780, (C) = 326, (AB) = 250, (BC) = 122,$

$(AC) = 144, (ABC) = 50.$

2. A number of persons were observed for their food habits. It was found that 25 of them wee vegetarians and 20 of them liked boiled vegetables. Another 10 gave out themselves out as vegetarians and liking boiled vegetables. Show that data so returned as a result of enquiry from 30 persons is incorrect and that some of them have withheld true information.
3. A market investigator returns the following data. Of 2000 people consulted 1754 liked chocolates 1872 liked toffee and 572 liked biscuits, 676 liked chocolates and toffee, 286 liked chocolates and biscuits, 270 liked toffee and biscuits, 144 liked all the three. Show that the information as it stands must be incorrect.
4. The following summary appears in a report on a survey covering 1000 fields. Scrutinize the numbers and point out if there is any mistake or misprint in them.

Manured fields	……….510
Irrigated fields	……….490
Fields growing improved varieties	……….427
Fields both irrigated and manured	……….189
Fields both manured and growing improved varieties	……….140
Fields both irrigated and growing improved varieties	……….85.

(Hint: Assuming consistency of data, using $(\alpha, \beta, \gamma) \geq 0$ get a contradiction as $(ABC) \leq 0$).

5. If $(A) = 50, (B) = 60, (C) = 50, (A\beta) = 5, (A\gamma) = 20$ and $N = 100$. find the least and greatest values of (BC).
6. Given that $(A) = (B) = (C) = \frac{N}{2} = 50; (AB) = 30; (AC) = 25.$ find the limits within which (BC) will lie.
7. If $N = 120, (A) = 60, (B) = 90, (C) = 30, (BC) = 15, (AC) = 15.$ find the limits between which (AB) must lie.
8. Among the adult population of a certain town 50% of the population are male., 60% are wage earners and 50% are 45 years of age or over. 10% of the males are not wage earners and 40% of the male are under 45 years. Can we infer anything about what percentage of the population of 45 or over are wage-earners? (Hint: taking the

attributes, ‘male’, ‘wage earners’, ’45 years or over’ as A, B, C respectively find the limits of (BC))).

4.3 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 amongst $\beta's$. Or equivalently the proportion of $B's$ amongst $A's$ is the same as amoungst the $\alpha's$.

Thus, A and B are independent iff

$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$ ………..(i) (or) $\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$ ………..(ii)

From (i) we get

$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{N}$.

$\therefore (AB) = \frac{(A)(B)}{N}$ ……….(1) and $(A\beta) = \frac{(A)(\beta)}{N}$ … … …. (2)

Again from (i) we get $1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$.

$$\therefore \frac{(B)-(AB)}{(B)} = \frac{(\beta)-(A\beta)}{(\beta)}$$

$$\therefore \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}.$$

$$\therefore \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} = \therefore \frac{(\alpha\beta)+(\alpha B)}{(\beta)+(B)} = \frac{(\alpha)}{N}.$$

$\therefore (\alpha\beta) = \frac{(\alpha)(\beta)}{N}$ ………………….(3)

And $(\alpha B) = \frac{(\alpha)(B)}{N}$ ………………….(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 $(AB)(\alpha\beta) = (A\beta)(\alpha B)$.

For if A and B are independent attributes we have

$$(AB)(\alpha\beta) = \left[\frac{(A)(B)}{N}\right] . \left[\frac{(\alpha)(\beta)}{N}\right] = \left[\frac{(A)(\beta)}{N}\right]\left[\frac{(\alpha)(B)}{N}\right] = (A\beta)(\alpha B).$$

Thus A and B are independent if $(AB)(\alpha\beta) - (A\beta)(\alpha B) = 0$ ……….(5)

Association and coefficient of association.

If $(AB) \neq \frac{(A)(B)}{N}$ we say that A and B are **associated**. There are two probabilities. If $(AB) > \frac{(A)(B)}{N}$ we say that A and B are **positively associated** and if $(AB) < \frac{(A)(B)}{N}$say that A and B are **negatively associated.**

Notation.

Let us denote $\delta = (AB) - \frac{(A)(B)}{N}$.

Thus $\delta = (AB) - \frac{(A)(B)}{N} = \frac{1}{N}[N(AB) - (A)(B)]$.

$$= \frac{1}{N}[\{(AB) + (A\beta) + (\alpha B) + (\alpha\beta)\}(AB) - \{(AB) + (A\beta)\}\{(AB) + (\alpha B)\}]$$

$= \frac{1}{N}[(AB)(\alpha\beta) - (A\beta)(\alpha B)]$ ………………(6)

Note. From (5) A and B are independent if $\delta = 0. A$ and B are positively associated if $\delta > 0$. and negatively associated if $\delta < 0$.

Coefficient of association. There are several measures indicating the intensity of association between two attributes A and B.

The most commonly used measures are the **Yule's coefficient** of association Q and **coefficient of colligation**Y which are defined as follows.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \; from\ (6)$$

$$Y = \frac{\left[1 - \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}{\left[1 + \sqrt{\left\{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}\right\}}\right]}.$$

Note 1. We know that A and B are independent if $\delta = 0$

$\therefore A$ and B are independent if $Q = Y = 0$

Note 2. If A and B are positively associated then $(AB) = (A)$ hence $(A\beta) = 0$ or $(AB) = (B)$ hence $(\alpha B) = 0$. in either case $Q = 1 = Y$.

Note 3. If A and B are perfectly disassociated then either $(AB) = 0$ or $(\alpha\beta) = 0$ and in this case $Q = -1 = Y$.

Thus, we get $-1 \leq Q \leq 1$ and $-1 \leq Y \leq 1$.

Note. Yule's coefficient Qand the coefficient of colligation Y is related by the relation $Q = \frac{2Y}{1+Y^2}$.

Proof. Let $x = \frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}$. hence $Y = \frac{1-\sqrt{x}}{1+\sqrt{x}}$.

$$\therefore Y^2 = \frac{\left(1-\sqrt{x}\right)^2}{\left(1+\sqrt{x}\right)^2}$$

$$\therefore 1+Y^2 = 1+\frac{1+x-2\sqrt{x}}{\left(1+\sqrt{x}\right)^2}$$

$$\therefore 1+Y^2 = 1+\frac{1+x-2\sqrt{x}}{\left(1+\sqrt{x}\right)^2} = \frac{2(1+x)}{\left(1+\sqrt{x}\right)^2}$$

$$\therefore \frac{2Y}{1+Y^2} = \frac{2\left(\frac{1-\sqrt{x}}{1+\sqrt{x}}\right)}{\frac{2(1+x)}{\left(1+\sqrt{x}\right)^2}} = \frac{\left(1-\sqrt{x}\right)\left(1+\sqrt{x}\right)}{1+x} = \frac{1-x}{1+x}$$

$$= \frac{1-\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}{1+\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}} = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}.$$

$$= Q.$$

From the above relationship between Y and Q we infer the following.

$Q = 0 \Longrightarrow Y = 0; Q = -1 \Longrightarrow Y = -1$ and

$Q = 1 \Longrightarrow Y = 1$ and conversely.

Solved problems.

Problem 1. Check whether the attributes A and B are independent given that

(i) $\boldsymbol{(A) = 30; (B) = 60, (AB) = 12, N = 150}$.

(ii) $\boldsymbol{(AB) = 256, (\alpha B) = 768, (A\beta) = 48, (\alpha\beta) = 144}$.

Solution. (i) since the given class frequencies are of first order the condition for independence is $(AB) = \frac{(A)(B)}{N}$.

Consider $\frac{(A)(B)}{N} = \frac{30\times 60}{150} = 12 = (AB)$

$\therefore (AB) = \frac{(A)(B)}{N}$. hence, A and B are independent.

(ii) $(A) = (AB) + (A\beta) = 256 + 48 = 304.$

$$(B) = (AB) + (\alpha B) = 256 + 768 = 1024$$

$$(\alpha) = (\alpha B) + (\alpha\beta) = 768 + 144 = 912.$$

$$(\beta) = (A\beta) + (\alpha\beta) = 48 + 144 = 192.$$

$$N = (A) + (\alpha) = 304 + 912 = 1216.$$

Now $\frac{(A)(B)}{N} = \frac{304\times1024}{1216} = 256 = (AB)$

$$\therefore (AB) = \frac{(A)(B)}{N}.$$

Hence, A and B are independent.

Aliter. Applying the condition (5) for independence,

$$(AB)(\alpha\beta) - (A\beta)(\alpha B) = 256 \times 144 - 768 \times 48 = 36864 - 36864 = 0$$

$\therefore A$ and B are independent.

Note. By proving $Q = 0$ also we can conclude A and B are independent.

Problem 2. Ina class test in which 135 candidates were examined for proficiency in Physics and chemistry, it was discovered that 75 students failed in Physics, 90 failed in Chemistry and 50 failed in both. Find the magnitude of association and state if there is any association between failing in Physics and Chemistry.

Solution. Denoting 'fail in Physics' as A and 'fail in chemistry' as B we get

$$(A) = 75, (B) = 90, (AB) = 50, N = 135.$$

The magnitude of association is measured by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

We now get the ultimate class frequencies.

$$(\alpha) = N - (A) = 135 - 75 = 60$$

$$(\beta) = N - (B) = 135 - 90 = 45$$

$$(\alpha B) = (B) - (AB) = 90 - 50 = 40$$

$$(A\beta) = (A) - (AB) = 75 - 50 = 25.$$

$$(\alpha\beta) = (\alpha) - (\alpha B) = 60 - 40 = 20$$

$$\therefore Q = \frac{50 \times 20 - 25 \times 40}{50 \times 20 + 25 \times 40} = 0$$

$\therefore A$ and B are independent. Hence failure in Physics and Chemistry are completely independent of each other.

Problem 3. Show whether A and B are independent or positively associated or negatively associated in the following cases.

(i) $\boldsymbol{N = 930; (A) = 300, (B) = 400, (AB) = 230}$.

(ii) $\boldsymbol{(AB) = 327, (A\beta) = 545, (\alpha B) = 741, (\alpha\beta) = 235}$.

(iii) $\boldsymbol{(A) = 470, (AB) = 300, (\alpha) = 530, (\alpha B) = 150}$.

(iv) $\boldsymbol{(AB) = 66, (A\beta) = 88, (\alpha B) = 102, (\alpha\beta) = 136}$

Solution. (i) $\frac{(A)(B)}{N} = \frac{300\times400}{930} = 129.03.$

Now, $\delta = (AB) - \frac{(A)(B)}{N} = 230 - 129.03 = 100.97.$

Here, $\delta > 0$. Hence A and B are positively associated.

(ii) $Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)} = \frac{327\times225-545\times741}{327\times225+545\times741}$

$$= \frac{76854 - 403845}{76854 + 403845} = \frac{-32700}{480690} = -0.6803.$$

$\therefore Q < 0$. Hence, A and B are negatively associated.

(iii) $N = (A) + (\alpha) = 470 + 530 = 1000.$

$$(B) = (AB) + (\alpha B) = 300 + 150 = 450.$$

Now, $\frac{(A)(B)}{N} = \frac{470\times450}{1000} = 2115.$

$$\therefore \delta = (AB) - \frac{(A)(B)}{N} = 300 - 2115 = -1915.$$

$\therefore \delta < 0$. 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

$$(AB) = 200, (A\beta) = 50, (\alpha B) = 110, (\alpha\beta) = 600.$$

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{200 \times 600 - 50 \times 110}{200 \times 600 + 50 \times 110}$$

$$= 0.91235.$$

Since Q is positive it means that intelligent fathers are likely to have intelligent sons.

Problem 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 $(AB) = 25, (A\beta) = 220$, $(\alpha B) = 90$ and $(\alpha\beta) = 160$.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{25 \times 160 - 220 \times 90}{25 \times 160 + 220 \times 90}$$

$$= \frac{400 - 19800}{400 + 19800} = \frac{-15800}{23800} = -0.6638.$$

This shows that the attributes A and B have negative association.

(i.e.,) 'innoculation' and 'attack from small pox' are negatively associated.

Thus innoculationaganst small pox can be taken as the preventive measure.

Problem 6. From the following data compare the association between marks in Physics and Chemistry in Madurai Kamaraj University (MKU) and Manonmaniamsundaranar University (MSU).

University	***MSU***	***MKU***
Total number of candidates	***200***	***1600***
Pass in Physics	***80***	***320***
Pass in Chemistry	***40***	***90***
Pass in Physics and Chemistry	***20***	***30***

Solution. Denoting 'Pass in Physics' as A and 'Pass in Chemistry' as B we have the following data for KMU and MSU.

MKU	MSU
$N = 1600$	$N = 200$
$(A) = 320$	$(A) = 80$
$(B) = 90$	$(B) = 40$
$(AB) = 30$	$(AB) = 20$

From the above data we get the rest of the class frequencies for MKU and MSU.

MKU	MSU
$(A\beta) = (A) - (AB)$ $= 320 - 30$ $= 290.$ $(\alpha\beta) = (B) - (AB)$ $= 90 - 30$ $= 60$ $(\alpha\beta) = N - (A) - (B) + (AB)$ $= 1600 - 320 - 90 + 30$ $= 1220$	$(A\beta) = (A) - (AB)$ $= 80 - 20$ =60. $(\alpha\beta) = (B) - (AB)$ $= 40 - 20$ $= 20$ $(\alpha\beta) = N - (A) - (B) + (AB)$ $= 200 - 80 - 40 + 20$ $= 100$

We now find the coefficient of association between A and B for MKU and MSU respectively.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

MKU	MSU
$Q = \frac{30 \times 120 - 290 \times 60}{30 \times 120 + 290 \times 60}$ $= \frac{36600 - 17400}{36600 + 17400}$ $= \frac{19200}{54000}$ $= 0.3556$	$Q = \frac{20 \times 100 - 60 \times 20}{20 \times 100 + 60 \times 20}$ $= \frac{2000 - 1200}{2000 + 1200}$ $= \frac{800}{3200}$ $= 0.25$

$\therefore Q$ of MKU $> Q$ of MSU. Thus the association between the knowledge in Physics and Chemistry is greater in MKU than in MSU.

EXERCISE QUESTIONS:

1. Show whether A and B are independent, positively 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) $(AB) = 90, (A\beta) = 64, (\alpha B) = 260, (\alpha\beta) = 110$

   (v) $(A) = 245, (\alpha) = 285, (AB) = 147, (\alpha B) = 190.$

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. How far is the knowledge in the two subjects associated?

3. Calculate Yule's coefficient of association between marriage failure of students

	Passed	**Failed**	**Total**
Married	90	65	155
Unmarried	260	110	370
Total	350	175	525

4. Investigate if there is any association between extravagance in father and son from the following data.

extra vagent fathers with extra vagent sons	= 470
extra vagent fathers with miserly sons	= 315
miserly fathers with extra vagent sons	= 280
miserly fathers with miserly sons	= 1220

5. Of 500 students appeared for a competitive examination 350 were successful 280 had attended a coaching class and of these 220 came out successful. Does coaching help in success?

6. Investigate the association between darkness of eye color in father and son from the following data:

Father with dark eyes and sons with dark eyes	78
Father with dark eyes and sons with not dark eyes	122
Father with not dark eyes and sons with dark eyes	96

Father with not dark eyes and sons with not dark eyes 704

Can we infer eye color is hereditary?.

7. From the following informations from the table, discuss the association between the color of the skin and color of the eyes.

Color of the skin	**Color of the eyes**	
	Black	**Brown**
Black	25	10
Red	12	38

8. 160 plants are characterized as per the nature of the laves and the color of the flowers.

	Normal leaves	**Abnormal leaves**
White flowers	99	36
Red flowers	20	5

9. In anti-malerial campaign in a certain area quinine was administered 812 persons out of a total population of 3248. The number of fever cases is shown below.

	Fever	**No fever**
Quinine	20	792
No quinine	220	2216

Can we infer quinine checks malerialfever.

10. From the figures given in the following table compare the association between literacy and unemployment in rural and urban areas and give reasons for the difference if any

	Urban	**Rural**
Total adult males	25 lakhs	200 lakhs
Litrate males	10 lakhs	40 lakhs
Unemployed males	5 lakhs	12 lakhs
Literate and unemployed males	3 lakhs	4 lakhs

11. Find whether A and B are in dependent in the following case. $(AB) = 256, (\alpha B) = 768;$

$(A\beta) = 48$ and $(\alpha\beta) = 200.$

UNIT V

INDEX NUMBERS

UNIT STRUCTURE

3.1 Consumer Price Index Numbers

3.2 Conversion of Chain Base Index number into Fixed base index numbers

INTRODUCTION

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. The comparison can be expressed as the percentage of ratio 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, various standard measures of central tendencies such as arithmetic mean, geometric mean, harmonic mean can be used.

In order to compute an index number it is necessary to collect a mass of data on the items which are being compared, decide on the type of average to be used and the relative weighting of the different items in the group. Generally quantities consumed are assigned as weight when price index number is considered. We shall describe several methods of calculating such index numbers but we shall not enter into technique aspects and the practical details of how the data are obtained, how accurate they are, and what governs the choice of the items that make up the group.

In the computation of an index number, if the base year used for comparison is kept constant throughout, 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 comparison are

(i) Aggregate method.

(ii) 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 denotes the price of the current year

$$p_{01} = \frac{\sum p_1}{\sum p_0} \times 100,$$

Where, $\sum p_1$ is total of the current year and

$\sum p_0$ is the total of the base year.

This is the simplest method in which aggregate of the prices for the 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 192 Rs.
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	58
Flour	4.5	5.25

Solution. Construction of price index taking 1991 as base year.

Commodities	Price in 1991 Rs.	Price in 192 Rs.
Rice	7	8
Wheat	3.5	3.75
Oil	40	45
Gas	78	58
Flour	4.5	5.25
Total	**133.0**	**147.00**

∴ Aggregate index number $p_{01} = \frac{\sum p_1}{\sum p_0} \times 100.$

$$= \frac{147}{133} \times 100.$$

$$= 110.5.$$

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 the 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 $\boldsymbol{p_{01}} = \frac{\boldsymbol{p_1}}{\boldsymbol{p_0}} \times \mathbf{100}$.

In the average of price relatives method the average for price relatives for various items is calculated by using any one of the measures of central tendencies such as arithmetic mean, geometric mean, harmonic mean, etc., Arithmetic and geometric means are the very common averages used in this method.

(i) The arithmetic mean index number $\boldsymbol{p_{01}} = \frac{\Sigma\left(\frac{\boldsymbol{p_1}}{\boldsymbol{p_0}}\times\mathbf{100}\right)}{\boldsymbol{n}}$

(ii) Geometric mean index number $\boldsymbol{p_{01}} = \left[\prod\left(\frac{\boldsymbol{p_1}}{\boldsymbol{p_0}}\right)\right]^{\mathbf{1}/\boldsymbol{n}} \times \mathbf{100}$

Where Π denotes the product.

Hence, $\log p_{01} = \frac{\Sigma\left(\frac{\boldsymbol{p_1}}{\boldsymbol{p_0}}\times\mathbf{100}\right)}{\boldsymbol{n}}$.

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 1991 Rs.	Price in 192 Rs.	$\frac{p_1}{p_0} \times 100$	$\log\frac{p_1}{p_0} \times 100$
Rice	7	8	14.3	2.0580
Wheat	3.5	3.75	107.1	2.0298
Oil	40	45	112.5	2.0512
Gas	78	58	109.0	2.0374
Flour	4.5	5.25	116.7	2.0671
Total			**559.6**	**10.2435**

(i) Using arthimetic mean the index number $p_{01} = \frac{559.6}{5} = 111.92.$

(ii) Using geometric mean the index number $\log p_{01} = \frac{10.2435}{5} = 2.0487.$

$$\therefore p_{01} = \text{antilog of}(2.0487) = 111.87.$$

Solved Problems

Problem 1. From the following data of the whole sale price of rice for the five 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	100
1988	6.00	$\frac{6}{5} \times 100 = 120$
1989	6.50	$\frac{6.5}{5} \times 100 = 130$
1990	7.00	$\frac{7}{5} \times 100 = 140$
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 1989 to 1988 there is an increase of 20% in the price of rice per Kg; for 1987 to 1989 there is an 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	100
1991	7.5	$\frac{7.5}{7} \times 100 = 107.1$

1992	8	$\frac{8}{7} \times 100 = 114.$

Problem 2. Construct the whole sale price index number for 1991 and 19922 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 1990as 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 numbers for 1992 as base year 1990 is 110.5

Problem 3. From the following average prices of the three groups of commodities in rupees per unit find (i) fixed base index number (ii) chain base index numbers with 1988 as the base year and

Commodity	*1988*	*1989*	*1990*	*1991*	*1992*
A	***2***	***3***	***4***	***5***	***6***
B	***8***	***10***	***12***	***15***	***18***
C	***4***	***5***	***8***	***10***	***12***

Solution.(i) Fixed base index number.

Commodity	**1988**	**1989**	**1990**	**1991**	**1992**
A	100	$\frac{3}{2} \times 100 = 150$	$\frac{4}{2} \times 100 = 200$	$\frac{5}{2} \times 100 = 250$	$\frac{6}{2} \times 100 = 300$
B	100	$\frac{10}{8} \times 100$ $= 125$	$\frac{12}{8} \times 100$ $= 150$	$\frac{15}{8} \times 100$ $= 188$	$\frac{18}{8} \times 100$ $= 225$
C	100	$\frac{5}{4} \times 100 = 125$	$\frac{8}{4} \times 100 = 200$	$\frac{10}{4} \times 100$ $= 200$	$\frac{12}{4} \times 100$ $= 300$
Total	**300**	**400**	**550**	**688**	**825**
Index number (A.M.)	100	133.3	183.3	229.3	275

(ii) chain base index numbers

Commodity	**1988**	**1989**	**1990**	**1991**	**1992**
A	$\frac{2}{2} \times 100$ $= 100$	$\frac{3}{2} \times 100$ $= 150$	$\frac{4}{3} \times 100$ $= 133.3$	$\frac{5}{4} \times 100$ $= 125$	$\frac{6}{5} \times 100$ $= 120$
B	$\frac{8}{8} \times 100$ $= 100$	$\frac{10}{8} \times 100$ $= 125$	$\frac{12}{10} \times 100$ $= 120$	$\frac{15}{12} \times 100$ $= 125$	$\frac{18}{15} \times 100$ $= 120$

C	$\frac{4}{4} \times 100$ $= 100$	$\frac{5}{4} \times 100$ $= 125$	$\frac{8}{5} \times 100$ $= 160$	$\frac{10}{8} \times 100$ $= 125$	$\frac{12}{10} \times 100$ $= 120$
Total	**300**	**400**	**413.3**	**375**	**360**
Index number (A.M.)	100	133.3	137.8	125	120

EXERCISE QUESTIONS:

1. From the following data construct the aggregate index number for taking 1990 as the base:

Commodities	Prices in 1990 Rs.	Prices 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	1988	1989	1990	1991	1992
Price of wheat per Kg.	4	5	6	7	8	10	9	10	11

3. Find the index numbers of price relatives using

   (i) Arithemetic mean

   (ii) Geometric mean as averages taking 1990 as base year

   (iii) Also find simple aggregate index.

Commodities	Year	
	1990	1991
Rice	158	272

Cholam	168	326
Cambu	157	309
ragi	155	304

4. From the following data construct an index number for 1970 taking 1969 as base by the price relatives using

   (i) A.M.

   (ii) G.M. for averaging the relatives.

Commodities	Price in 1969 (Rs.)	Price in 1970 (Rs.)
A	150	170
B	40	60
C	80	90
D	100	120
E	20	25

5. From the following data find

   (i) Fixed base index numbers with 1988 as the base year

   (ii) chain base index numbers.

commodities	Price in rupees				
	1988	1989	1990	1991	1992
I	2	3	5	7	6
II	8	10	12	4	18
III	4	3	7	9	12

6. The following table gives the average whole sale prices of gold, coal, rice during the years 1934 to 1941. Find out the index number by reference to 1934 as base year.

Commodities	Average whole sale prices in rupees							
	1934	1935	1936	1937	1938	1939	1940	1941
Gold per tola	50.6	61.6	66.8	71.0	70.6	72	72	75.6
Coal per maund	6.8	6.4	5.6	6.2	6.4	7.8	6	6.8

Rice per maund	29.6	25.8	26.4	28.6	28.6	30.2	28	34.6

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. For, example, in the budget of a middle class family food items are much more essential than cosmetic and luxury items. Even among food items consumption of rice is greater than the consumption of wheat and flour. Hence due weightage is to be assigned to individual items under consideration. This leads to the concept of weighted index numbers. Generally we assign in such cases weights by taking the quantity of the commodities which is actually consume as the weights or by the values of commodities as the weights. When the quantities consumed are taken as the weight they may be of different units of measurements which may not be conductive to have an index number. Such situation will compel as to take the value of the commodities as the weights. Thus we notice that weighted index number gives due consideration to the relative importance of items and makes the index number more representative.

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.

Note. In what follows p_0, p_1 denotes the prices of the base year and current year respectively and q_0, q_1 denote the quantities consumed in the base year and current year rsectively.

(a) Laspeyre's index number. According to Laspeyr'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

$$\boldsymbol{L_{I_{01}} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100}.$$

(b) Paasches' index number. According to Paasches' method current year quantities are taken as weights and hence Paasches' index number is defined

$$\boldsymbol{P_{I_{01}}} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \mathbf{100}.$$

(c) Marshall – Edgeworth's index number. According to this method the weight is the sum of the quantities of the base period and current period. Hence Marshall-Edgeworths' formulae is defined by

$$\boldsymbol{M_{I_{01}}} = \frac{\sum p_1(q_0 + q_1)}{\sum p_0(q_0 + q_1)} \times \mathbf{100}$$

$$= \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times \mathbf{100}.$$

(d) Bowley's index number. The arithemetic mean of Laspeyre's an Paasches' index number is defined to be Bowley's index number. Hence Bowley's index number is given by

$$\boldsymbol{B_{I_{01}}} = \frac{L_{I_{01}} + P_{I_{01}}}{2}$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times \mathbf{100}.$$

(e) Fisher's index number. Prof. Irving Fisher, though suggested many index numbers, gives an 'ideal index number' as

$$\boldsymbol{I_{01}} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \mathbf{100}.$$

Wenotice$I_{01} = \sqrt{L_{I_{01}} \times P_{I_{01}}}$. That is Fisher's index number is the geometric mean of Laspeyre'sindexnumber and Paasche's index number.

(f) Kelley's index number. According to Kellly, weight may be taken as the quantities of the period which is not necessarily the base year or current year. The average of two quantity of two or more years may be taken as the weight. Hence Kelly's index number can be defined as

$$\boldsymbol{K_{I_{01}}} = \frac{\sum p_1 q}{\sum p_0 q} \times \mathbf{100}$$

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) Paasches' (iii) Fishers' 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	10	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 $\boldsymbol{L_{I_{01}}} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \mathbf{100}$

$$= \frac{433.5}{343} \times 100.$$

$$= 126.4$$

(ii) Paasche's index number $\boldsymbol{P_{I_{01}}} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \mathbf{100}.$

$$= \frac{466.5}{370} \times 100.$$

$$= 126.1$$

(iii) Fisher's ideal index number $\boldsymbol{I_{01}} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times \mathbf{100}.$

$$= \sqrt{\frac{433.5 \times 466.5}{343 \times 370}} \times 100$$

$$= 126.2.$$

(iv) Bowley's index numbers $\boldsymbol{B}_{I_{01}} = \frac{L_{I_{01}}+P_{I_{01}}}{2}$

$$= \frac{126.4 + 126.1}{2}$$

$$= 126.25.$$

(v) Edge-worth's index number $= \frac{\sum p_1q_0+\sum p_1q_1}{\sum p_0q_0+\sum p_0q_1} \times \mathbf{100}.$

$$= \frac{433.5 + 466.5}{343 + 370} \times 100.$$

$$= \frac{900}{713} \times 100$$

$$= 126.2$$

II-B Weighted average of price relatives method.

In this method 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_0q_0 then the index number $p_{01} = \frac{\sum PV}{\sum V}$.

Example.

Index numbers using weighted arithmetic mean of price relatives.

Commodity	**Price in 1990 p_0**	**Price in 1990 p_1**	**Quantity in 1990 q_0**	V p_0q_0	P $\frac{p_1}{p_0} \times 100$	PV
Coconut oil	Rs.50	Rs.54	15 lit	750	108	81000
Groundnut oil	Rs.45	Rs.48	25 lit	1125	106.7	120037.5
Gingilee 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**

∴ Weighted index number $= \frac{\sum PV}{\sum V} = \frac{651170.5}{5615} \approx 116.$

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 test.

(iii) The commodity reversal test.

(i) The time reversal test. Let $I_{(01)}$ denote the index 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

$I_{(01)} \times I_{(10)} = 1$, then we say tht the index number satisfies the time reversal test.

(ii) The factor reversal test. In this test the prices and quantities are interchanged, without considering the percentage, satisfying the following relation $I_{(pq)} \times I_{(qp)} = \frac{\Sigma \boldsymbol{p_1 q_1}}{\Sigma \boldsymbol{p_0 q_0}}$,where $I_{(pq)}$ is the price index of the current year relative to base year and $I_{(qp)}$ is the quantity index of the current year relative to the base year.

(iii) The commodity reversal test. 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 three tests for ideal number.

Fisher's index number is $I_{(01)} = \sqrt{\frac{\Sigma \boldsymbol{p_1 q_0}}{\Sigma \boldsymbol{p_0 q_0}} \times \frac{\Sigma \boldsymbol{p_1 q_1}}{\Sigma \boldsymbol{p_0 q_1}}}$

Time reversal test. Interchanging base year and current year

$$= I_{(10)} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

Now, $I_{(01)} \times I_{(10)} = \sqrt{\frac{\Sigma \boldsymbol{p_1 q_0}}{\Sigma \boldsymbol{p_0 q_0}} \times \frac{\Sigma \boldsymbol{p_1 q_1}}{\Sigma \boldsymbol{p_0 q_1}} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = 1$

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

$$I_{(qp)} = \sqrt{\frac{\Sigma q_1 p_1}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}.$$

Now, $I_{(pq)} \times I_{(qp)} = \sqrt{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right) \times \left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}\right)}$

$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$.

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 $p_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$ and $p_{10} = \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1}$

Now, $p_{01} \times p_{10} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \neq 1.$

Hence, Laspeyre's index number does not satisfy the time reversal test.

Remark 3. Paasche's index number does not satisfy time reversal test (verify)

Remark 4. Laspeyre's and Paasche's index number also does not satisfy factor reversal teast (verify)

Solved Problems.

Problem 1. Construct, with the help of data given below, Fisher's index number and show 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	50	110	480	440	600	550
D	3	30	4	35	90	105	120	140
Total					**1060**	**1145**	**1390**	**1510**

Fisher's index number is $I_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100.$

$$= \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}} \times 100.$$

Time’s reversal test.

$$\text{Now, } I_{(01)} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma\, p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma\, p_0 q_1}} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}}$$

$$\therefore I_{(10)} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma\, p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma\, p_1 q_0}} = \sqrt{\frac{1145}{1510} \times \frac{1060}{1390}}$$

$$\text{Now, } I_{(01)} \times I_{(10)} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145} \times \frac{1145}{1510} \times \frac{1060}{1390}} = 1.$$

Factor reversal test.

$$I_{(pq)} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma\, p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma\, p_0 q_1}} = \sqrt{\frac{1390}{1060} \times \frac{1510}{1145}}$$

Interchanging the factors,

$$I_{(qp)} = \sqrt{\frac{\Sigma q_1 p_1}{\Sigma\, q_0 p_0} \times \frac{\Sigma q_1\, p_1}{\Sigma\, q_0 p_1}} = \sqrt{\frac{1145}{1060} \times \frac{1510}{1390}}$$

Now,

$$I_{(pq)} \times I_{(qp)} = \sqrt{\frac{1390 \times 1510}{1060 \times 1145} \times \frac{1145 \times 1510}{1060 \times 1390}}$$

$$= \frac{1510}{1060}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma\, p_0 q_0}$$

Hence, the factor reverse test is also satisfied.

Problem 3.

Find the missing price in the following data if the ratio between Lapeyre’s and Paasche’s index number is $\mathbf{25{:}\,24}$.

Commodities	***Base year***		***Current year***	
	Price quantity		***Price quantity***	
A	***1***	***15***	***2***	***15***
B	***2***	***15***	***-***	***30***

Solution. Let the missing price be x.

Commodity	**Base year**	**Current year**	$\boldsymbol{p_0 q_0}$	$\boldsymbol{p_0 q_1}$	$\boldsymbol{p_1 q_0}$	$\boldsymbol{p_1 q_1}$

	p_0	q_0	p_1	q_1				
A	1	15	2	15	15	15	30	30
B	2	15	x	30	30	60	15x	30x
Total					**45**	**75**	**30+15x**	**30+30x**

Laspeyre’s index number $L_{I_{(01)}} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$= \frac{30+15x}{45} \times 100.$

Paache’s index number $P_{I_{(01)}} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100.$

$= \frac{30+30x}{75} \times 100.$

Given $L_{I_{(01)}} : P_{I_{(01)}} = 25:24.$

$$\therefore \left(\frac{30+15x}{45} \times 100\right) : \left(\frac{30+30x}{75} \times 100\right) = 25:24$$

$$\therefore 24\left(\frac{30+15x}{45}\right) = 25\left(\frac{30+30x}{75}\right)$$

$$72(30+15x) = 45(30+30x)$$

$$8(30+15x) = 5(30+30x)$$

$$8 \times 15(2+x) = 5 \times 30(1+x)$$

$$4(2+x) = 5(1+x)$$

$\therefore x = 3$. Hence the missing price is Rs. 3.

Study Learning Material Prepared by

Dr. S.N. LEENA NELSON M.Sc., M.Phil., Ph.D.
Associate Professor & Head, Department of Mathematics,
Women’s Christian College, Nagercoil – 629 001,
Kanyakumari District, Tamilnadu, India.