PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in a subject is introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation. I am happy to note that the university has been recently accredited by National Assesment and Accreditation Council of India (NAAC) with grade "A".

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing, and devising of a proper lay-out of the materials. Practically speaking, their role amounts to an involvement in 'invisible teaching'. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the, University.

Needless to add, a great deal of these efforts are still experimental—in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

D:\NSOU\DTP\Statistic for Economic-2 (Alok)
Preface\ 6th proof\(Demay Submit)

Netaji Subhas Open University

Post Graduate Degree Programme

MA in Economics

Course : Statistics for Economics

Code : PGEC – II

First Print : December, 2021

Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.

D:\NSOU\DTP\Statistic for Economic-2 (Alok)
Preface\ 6th proof\(Demay Submit)

Netaji Subhas Open University

Post Graduate Degree Programme

MA in Economics

Course : Statistics for Economics

Code : PGEC – II

: Board of Studies :
Members

: Course Writer :
Dr. Bibekananda Raychoudhuri,
Associate Professor of Economics
NSOU

: Course Editor :
Dr. Seikh Salim,
Associate Professor of Economics,
NSOU

: Format Editor :
Priyanti Bagchi
Assistant Professor of Economics, NSOU

Kishore Sengupta
Registrar

D:\NSOU\DTP\Statistic for Economic-2 (Alok)
Preface\ 6th proof\(Demay Submit)

Netaji Subhas
Open University

PG : Economics
(PGEC)

PGEC – II : Statistics for Economics

Unit 1	❒ **Measures of Central Tendency**	7
Unit 2	❒ **Measures of Dispersion**	39
Unit 3	❒ **Moments, Skewness and Kurtosis**	67
Unit 4	❒ **Correlation and Regression**	96
Unit 5	❒ **Index Numbers and their Applications**	126
Unit 6	❒ **Introduction to the Theory of Probability and Distribution**	160

D:\NSOU\DTP\Statistic for Economic-2 (Alok)
Preface\ 6th proof\(Demay Submit)

Unit 1 ❐ Measures of Central Tendency

Structure

1.1 Objectives

1.2 Introduction

1.3 Types of Averages

1.3.1 The Arithmatic Mean

1.3.1.1 Arithmatic Mean Computed from grouped data

1.3.1.2 Weighted Mean

1.3.1.3 Properties of Arithmatic Mean

1.3.2 The Median

1.3.2.1 Median for ungrouped data

1.3.2.2 Median for grouped data

1.3.2.3 Quartiles, Deciles and Percentiles

1.3.3 The Mode

1.3.3.1 Mode for ungrouped data

1.3.3.2 Mode for grouped data

1.3.4 Empirical relation between Mean, Median and Mode

1.3.5 The Geometric Mean

1.3.6 The Harmonic Mean

1.3.7 Relation between Arithmatic, Geometric and Harmonic Mean

1.4 Some Problems and Theorems on Central Tendency

1.5 Summary

1.6 Questions

1.7 References

1.1 Objectives

There are two main objectives of the study of averages

- To get single value that describes the characteristics of the entire group. For example it is impossible to remember the individual incomes of millions of income earners in a country. But if the average income is obtained by dividing the total national income by total population, we get one single value that represents the entire population. This will also reflect the standard of living.

- To facilitate comparison : We can compare the percentage results of students of different schools in the H.S. Exam for 2018 and thereby conclude which school is the best or we can compare the pass percentage of the same school for different time periods and thereby conclude as to whether the results are improving or deteriorating.

1.2 Introduction

One of the most important objectives of statistical analysis is to get one single value that describes the characteristics of the entire mass of a huge data set. Such value is termed as the central value or an average or the expected value of the variable. When we say 'he is an average student' we mean he is neither very good nor very bad— just a mediocre student. But in statistics the term average has a different meaning.

An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is also called a measure of central value.

1.3 Types of Averages

The following are the mostly used averages.

- Arithmatic mean (i) simple and (ii) weighted
- Median
- Mode
- Geometric mean
- Harmonic mean

1.3.1 The Arithmatic Mean

The arithmatic mean is the most commonly used and readily understood measure of central tendency. It is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observation. Symbolically it can be represented as

$$\bar{X} = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + ... + x_n}{N}$$

Where, Σx indicates the sum of the values of all the observations and N is the total number of observations.

If the distribution is discrete, that is, the variate is a whole number, but in the form of frequency distribution, eg.

x (Marks)	**f** (No. of Students)	**fx**
10	2	20
15	4	60
20	6	120
25	8	200
30	10	300
Total	30	700

In such a situation the arithmatic mean (through Direct method)

$$\bar{X} = \frac{\Sigma fx}{N} \qquad \text{where, } N = \Sigma f$$

$$= \frac{700}{30} = 23.33 \text{ Marks.}$$

Short cut Method

x	**f**	**dx = (x – A)**	**fdx**
		A = Assumed Mean = 20	
10	2	–10	–20
15	4	–5	–20
20	6	0	0
25	8	5	40
30	10	10	100
	$\Sigma f = 30$		$\Sigma fdx = 100$

$$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f} = 20 + \frac{100}{30} = 20 + 3{\cdot}33 = 23{\cdot}33$$

1.3.1.1 Arithmatic mean computed from Grouped Data

When the observations are classified into a frequency distribution, the mid point of the class interval would be treated as the representative average value of that class. Therefore, for grouped data, the arithmatic mean is defined as

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{\Sigma fx}{N} \text{ where } \Sigma f = N$$

Where, x is the mid point of the various classes, f is the frequency for corresponding classes and N is the total frequency ($N = \Sigma f$). In continuous series, the class intervals are replaced by the mid-points of various classes. The mid values of a class is obtained by adding the upper and lower limits of the class and dividing the sum by two. The mid values of various classes are taken as representatives of the classes and we assume that the frequencies in each class are centred at mid point. Arithmatic mean in continuous series with grouped data can be obtained by any of the following methods (i) Direct method, (ii) Short-cut method, (iii) Step deviation method.

(i) Direct method $\quad \bar{X} = \dfrac{\Sigma fx}{\Sigma f}$

(ii) Short cut method $\quad \bar{X} = A + \dfrac{\Sigma fd}{\Sigma f}$

where, A is the assumed mean and

d is the deviation from assumed mean ie. $d = (X - A)$

(iii) Step Deviation method $\quad \bar{X} = A + \dfrac{\Sigma fd'}{\Sigma f} \times C$

where, C is the size of the class interval

$$d' = \frac{X - A}{C} = \frac{d}{C}$$

$$\rightarrow \quad fd' = \frac{1}{C} f\left(X - A\right)$$

$$\therefore \quad \Sigma fd' = \frac{1}{C} \Sigma f\left(X - A\right) \quad \text{or,} \quad \Sigma fd = C\Sigma fd'$$

Hence $\bar{X} = A + \dfrac{C\Sigma fd'}{\Sigma f}$

Example : Direct method to obtain the Arithmatic Mean

Marks :	0–10	10–20	20–30	30–40	40–50	50–60
No. of Students :	5	10	25	30	20	10

Marks	Mid Point	No. of Students	fx
	(x)	(f)	
0–10	5	5	25
10–20	15	10	150
20–30	25	25	625
30–40	35	30	1050
40–50	45	20	900
50–60	55	10	550
		N = 100	Σfx = 3,300

$$\therefore \quad \bar{X} = \frac{\Sigma fx}{N} = \frac{3300}{100} = 33$$

Short Cut method to obtain Arithmatic mean

Marks	Mid Point	No. of Students	(X–35)	fd
	(x)	(f)	(d)	
0–10	5	5	–30	–150
10–20	15	10	–20	–200
20–30	25	25	–10	–250
30–40	35	30	0	0
40–50	45	20	+10	+200
50–60	55	10	+20	+200
		N = 100		Σfd = – 200

$$\therefore \quad \bar{X} = A + \frac{\Sigma fd}{N} = 35 - \frac{200}{100} = 35 - 2 = 33$$

[Here, the assumed mean is A = 35]

Step deviation method to obtain Arithmatic Mean

Marks	Mid Point	No. of Students	(X–35)	(X–35) / 10	fd′
	(x)	(f)	d = X – A	(d')	
0–10	5	5	–30	–3	–15
10–20	15	10	–20	–2	–20
20–30	25	25	–10	–1	–25
30–40	35	30	0	0	0
40–50	45	20	+10	+1	+20
50–60	55	10	+20	+2	+20
		N = 100			$\Sigma fd' = -20$

$$\therefore \quad \bar{X} = A + \frac{\Sigma fd'}{N} \times C = 35 - \frac{20}{100} \times 10 = 35 - 2 = 33$$

1.3.1.2 Weighted Mean

One of the limitations of the arithmatic mean is that it gives equal importance to all the items. But there are cases where the relative importance of the different items is not the same.

The formula for computing weighted arithmatic mean is

$$\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$$

Where, $\bar{X}_w$ is the weighted arithmatic mean; X represents the variable values i.e. X_1, X_2, ...X_n; W represents the weights attached to variable values i.e. W_1, W_2,....., W_n respectively.

(i) Multiply the weights by the variable X and obtain the total ΣWX.

(ii) Divide this total by the sum of the weights i.e. ΣW.

In case of frequency distribution, if f_1, f_2,.....f_n are the frequencies of the variable values X_1, X_2,.....X_n respectively then the weighted arithmatic mean is given by

$$\bar{X}_w = \frac{\Sigma W(fX)}{\Sigma W}$$

In the expanded form it will be

$$\bar{X}_w = \frac{W_1(f_1X_1) + W_2(f_2X_2) + ... + W_n(f_nX_n)}{W_1 + W_2 + ... + W_n}$$

1.3.1.3 Properties of Arithmatic Mean

1. The greatest merit of the arithmatic mean is its simplicity— easy to calculate and easy to understand.
2. The sum of the deviations of the observations from the arithmatic mean is always zero

   $$\sum\left(X - \bar{X}\right) = 0$$

   So arithmatic mean is characterised as a point of balance, i.e., sum of the positive deviations from mean is equal to the sum of negative deviations from mean.
3. The sum of the squared deviations of the observations from the mean is minimum

   $\sum\left(X - \bar{X}\right)^2$ is minimum

   i.e. $\sum\left(X - \bar{X}\right)^2 < \sum\left(X - A\right)^2$ will be always true where A is not the arithmatic mean.
4. Since $\bar{X} = \frac{\Sigma X}{N}$ $\therefore$ $N\bar{X} = \Sigma X$

   If we replace each item in the series by the mean, then the sum of these substitutions will be equal to the sum of the individual items.

   This property of the arithmatic mean has great practical value. For example if we know the average wage in a factory, say Rs. 1000 and the number of workers employed, say 200, we can compute the total wage bill from the equation $N\bar{X} = \Sigma X$. The total wage bill in this case will be 200 × 1000 = 2,00,000/-.
5. If we have the arithmatic mean and number of items of two groups given to us, we can compute the combined average of these groups by using the following formula

   $$\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

   where, $\bar{X}_{12}$ = Combined mean of the two groups
   $\bar{X}_1$ = arithmatic mean of the first group
   $\bar{X}_2$ = arithmatic mean of the second group
   n_1 = No. of items in the first group
   n_2 = No. of items in the second group.

1.3.2 The Median

The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other part comprising all values less than the median. If the number of observations is odd, then the median is equal to one of the original observations which is the middle-most. If the number of observations is even, then the median is the arithmatic mean of the two middle-most observations.

Comparing the AM and Median as alternative measures of central tendency, we find that the former is the centre of gravity of the set of values and it balances observations on its left and right in terms of their magnitude, where as the median is the middle most value and has equal number of observations on both sides irrespective of their magnitude.

For example, if the income of 5 employees are Rs. 900, Rs. 950, Rs. 1020, Rs, 1200 and Rs. 1280, the median would be Rs. 1020 as this is the middle most (3rd) value of the variable.

Suppose, the income series is even for 6 employees : 900, 950, 1020, 1200, 1280 and 1300. So the median here will be the arithmatic mean of the two middle most numbers i.e., the AM of the 3rd of the 4th values.

$$\text{Median} = \frac{1020 + 1200}{2} = \frac{2220}{2} = 1110$$

1.3.2.1 Median for ungrouped data

In ungrouped series, first of all we find the cumulative frequencies and then we have to calculate the measure of $\frac{N+1}{2}$th item. The value can be easily located from the table. In the cumulative frequency column we have to find that total which is either equal to $\frac{N+1}{2}$ or next higher to that and median will be the value of the variable corresponding to it.

Example : From the following data find the value of Median

Income (Rs.) :	1000	1500	800	2000	2500	1800
No. of Persons :	24	26	16	20	6	30

Income arranged in ascending order	No. of Persons	Cumulative frequency
800	16	16
1000	24	40
1500	26	66
1800	30	96
2000	20	116
2500	6	122

Median = Size of $\frac{N+1}{2}$ th item $= \frac{122+1}{2}$ = 61.5th item size of 61.5th item = 1500.

1.3.2.2 Median for grouped data

First we have to determine the particular class in which the value of median lies. We use $\frac{N}{2}$ to find out the median class. Once the median class is ascertained, we use the following formula to find out the median.

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

Where, L = Lower limit of the median class, i.e. the class in which the middle-most item of the distribution lies.
c.f. = Cumulative frequency of the class preceding the median class
f = Simple frequency of the median class
i = The class interval of the median class.

Example : Calculate the median for the following grouped data

Marks	No. of Students
45–50	10
40–45	15
35–40	26
30–35	30
25–30	42
20–25	31
15–20	24
10–15	15
5–10	7

Ans. First we have to arrange the data in ascending order.

Marks	f	cf
5–10	7	7
10–15	15	22
15–20	24	46
20–25	31	77
25–30	42	119
30–35	30	149
35–40	26	175
40–45	15	190
45–50	10	200

Median = Size of $\frac{N}{2}$ item = $\frac{200}{2} = 100$ th item

$\therefore$ Median lies in the class 25–30

$$\therefore \quad \text{Median} = L + \frac{N/2 - c.f}{f} \times i$$

Here, L = 25, $N/2$ = 100, c.f. = 77, f = 42 and i = 5

$$\therefore \quad \text{Median} = 25 + \frac{100 - 77}{42} \times 5 = 25 + 2{\cdot}74 = 27{\cdot}74$$

1.3.2.3 Quartiles, Deciles and percentiles

The procedure for computing quartiles, deciles and percentiles is the same as that of the median. While computing these values in discrete cases (ungrouped data) we add 1 to N whereas in continuous cases (grouped data) we do not add 1 to N.

Thus, Q_1 (first quartile) = Size of $\frac{N+1}{4}$ th item for ungrouped data.

Q_1 (first quartile) = Size of $\frac{N}{4}$ th item for grouped data.

Q_3 (3rd quartile) = Size of $\frac{3(N+1)}{4}$ th item for ungrouped data.

Q_3 (3rd quartile) = Size of $\frac{3N}{4}$ th item for grouped data.

D_4 (4th decile) = Size of $\frac{4(N+1)}{10}$ th item for ungrouped data.

D_4 (4th decile) = Size of $\frac{4N}{10}$ th item for grouped data.

P_{60} (60th percentile) = Size of $\frac{60(N+1)}{100}$ th item for ungrouped data.

P_{60} (60th percentile) = Size of $\frac{60N}{100}$ th item for grouped data.

Example : Calculate the first quartile and 3rd decile of 20th percentile from the following data.

Mid-point :	2.5	7.5	12.5	17.5	22.5
Frequency :	7	18	25	30	20

Ans. : Since we are given the mid-points, we will first find the lower & upper limits of the various classes. The procedure is to take the difference between the two mid points and divide it by 2. Now we have to deduct the value so obtained from the mid-point to get the lower class limit and again to add the value so obtained to the mid point to obtain the upper class limit.

In the given case, $\frac{7{\cdot}5 - 2{\cdot}5}{2} = \frac{5}{2} = 2{\cdot}5$

Now we deduct 2·5 from the first mid-point 2·5 to get the lower class limit of the first class. ∴ 2·5 – 2·5 = 0 is the lower limit. The upper limit of the first class will be 2·5 + 2·5 = 5. So the first class is (0–5).

Class boundaries	**f**	**c.f**
0–5	7	7
5–10	18	25
10–15	25	50
15–20	30	80
20–25	20	100
	N = 100	

First Quartile : Q_1 = size of $\frac{N}{4}$ th item $= \frac{100}{4} = 25$ th item

$\therefore$ Q_1 lies in the class (5–10)

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

L = 5, $N/4$ = 25, c.f. = 7, f = 18 and i = 5

$$\therefore \quad Q_1 = 5 + \frac{25 - 7}{18} \times 5 = 5 + 5 = 10 \text{ (1st Quartile)}$$

Third decile : D_3 = Size of $\frac{3N}{10}$ th item

$$= \frac{3 \times 100}{10} = 30 \text{ th item}$$

$\therefore$ D_3 lies in the class (10–15)

$$D_3 = L + \frac{3N/10 - c.f}{f} \times i$$

L = 10, $3N/10 = 30$, c.f. = 25, f = 25, i = 5

$$\therefore \quad D_3 = 10 + \frac{30 - 25}{25} \times 5 = 10 + 1 = 11 \text{ (3rd Decile)}$$

20th Percentile : P_{20} = Size of $\frac{20N}{100}$ th item

$$= \frac{20 \times 100}{100} = 20 \text{ th item}$$

$\therefore$ P_{20} lies in the class (5–10)

$$P_{20} = L + \frac{20N/100 - c.f}{f} \times i$$

L = 5, $20N/100 = 20$, c.f. = 7, f = 18, i = 5

$$\therefore \quad P_{20} = 5 + \frac{20 - 7}{18} \times 5 = 5 + 3{\cdot}61 = 8{\cdot}61 \text{ (20th Percentile)}$$

Graphical determination of Median :

Example : Marks obtained by 100 students in statistics in a college is given below. Find out the Median and the quartiles.

Marks	No. of Students	Marks	No. of Students
0–5	4	20–25	25
5–10	6	25–30	22
10–15	10	30–35	18
15–20	10	35–40	5

First we draw the cumulative frequency curve of 'less than' type and calculate the median and the quartiles.

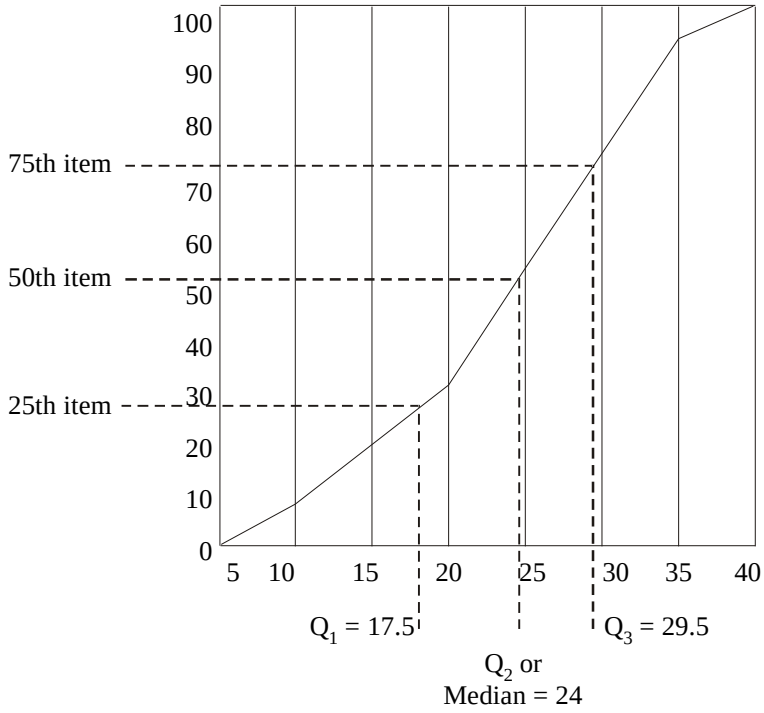
Marks	No. of Students (Cumulative fr)
less than 5	4
less than 10	10
less than 15	20
less than 20	30
less than 25	55
less than 30	77
less than 35	95
less than 40	100

Median = Size of $N/2$ th item $= 100/2 = 50$ th item

Q_1 = Size of $N/4$ th item $= 100/4 = 25$ th item

Q_3 = Size of $3N/4$ th item $= 3 \cdot 100/4 = 75$ th item

Locating Median, Q_1 and Q_3 graphically



So, graphically Q_1 = 17.5, Median = 24 and Q_3 = 29.5

1.3.3 The Mode

Mode is that value of the variate which has the maximum concentration arround it, For example the modal shirt size is that which is worn by more persons than any oher single size.

A distribution is said to bc unimodal, bi-modal or multi-modal according as it has unique mode or two-modes or more than two modes. Suppose, we have to find out the mode from the following data series of marks :

4, 6, 5, 7, 9, 8, 10, 4, 7, 6, 5, 9, 8, 7, 7

$\therefore$ Array = 4, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 10

Mode = 7 (frequency = 4 maximum)

1.3.3.1 Mode for ungrouped data

Normally the data series is arranged is an array and then we have to find out which value of the variate is occuring the maximum number of times. That value is the mode of the series. This is called the method of inspection.

Another method is called the method of grouping. We make 6 columns as per directions given below.

Column I – Write the original given frequency

Column II – Sum the frequencies taking two at a time starting from the 1st frequency of Col.I.

Column III – Sum the frequencies taking two at a time starting from the 2nd frequency in Col. I.

Column IV – Sum the frequencies taking three at a time starting from the 1st frequency in Col. I.

Column V – Sum the frequencies taking three at a time starting from the 2nd frequency in Col. I.

Column VI – Sum the frequencies taking three at a time starting from the 3rd frequency is Col. I.

Height in inches	f I	II	II	IV	V	VI
66	1	3				
67	2		⑥	⑦		
68	④	⑦			⑨	
69	3		5			⑨
70	2	3		6		
71	1		3		5	
72	2	3				4
73	1		2	4		
74	1	2			3	
75	1					

Analysis Table	
Columns	**Size of items having maximum f**
I	68
II	68, 69
III	67, 68
IV	66, 67, 68
V	67, 68, 69
VI	68, 69, 70

The item 68 occurs maximum number of times ie. 6 times. Hence Mode = 68

1.3.3.2 Mode for grouped data

Mode for the grouped data can be calculated with the help of the following formula.

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Where, L is the lower limit of the modal class

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class.

i is the class interval of the modal class

Graphical location of Mode

The value of the Mode can be determined graphically in a frequency distribution. Following are the steps for locating mode on graph.

1. Prepare a Histogram of the given data
2. The highest rectangle will indicate the modal class.
3. Draw two straight lines diagonally in the inside of the modal class rectangle from top corners of it to the upper corner of the adjacent bar.
4. From the point of intersection of these lines draw a perpendicular on the X-axis which gives the modal value.

Example : Find the Mode from the following frequency distribution.

Weekly wages (in Rs.) :	10–15	15–20	20–25	25–30	30–40	40–60	60–80
No. of workers :	7	19	27	15	12	12	8

Ans. Rearrange the frequency distribution

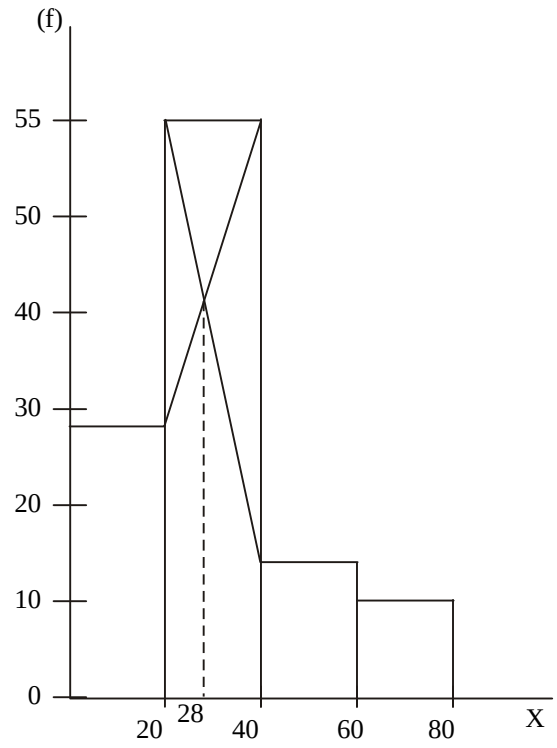
0–20	26
20–40	54
40–60	12
60–80	8

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 20 + \frac{54 - 26}{108 - 26 - 12} \times 20$$

$$= 20 + 8 = 28$$

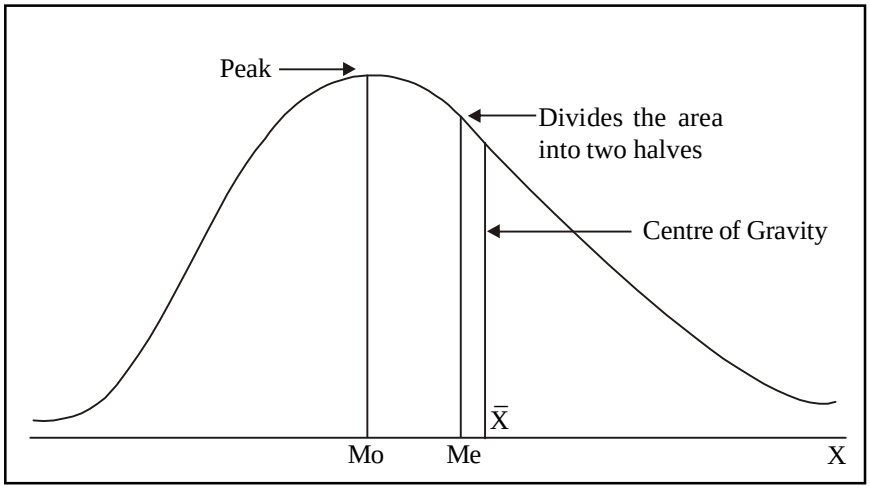
Graphically



Graphically the Mode is 28

1.3.4 Empirical relation between Mean, Median and Mode

A distribution in which the values of mean, median and mode coincide (i.e, mean = median = mode) is known as a symmetrical distribution. Conversely, when the values of mean, median and mode are not equal, the distribution is known as asymmetrical or skewed. In moderately skewed distributions a very important relationship exists between mean, median and mode. In such distributions the distance between the mean and the median is about one-third the distance between the mean and the mode as will be clear from the diagram below.



Karl Person has expressed this relationship as follows

Mode = Mean – 3 [Mean – Median]

$\therefore$ Mode = 3 Median – 2 Mean

and Median = Mode + $\frac{2}{3}$ [Mean – Mode]

If we know any of the two values out of the three, we can compute the third from these relationships.

1.3.5 The Geometric Mean

The geometric mean like the arithmatic mean is a calculated average. The geometric mean GM of a series of numbers x_1, x_2,....., x_n is defined as

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot .....x_n}$$

or the nth root of the product of n observations. When the number of observation is three or more, the task of computation becomes quite tedious. A transformation

into logarithm is useful to simplify calculations. If we take log of both sides, then the formula for GM becomes

$$\log GM = \frac{1}{n}\left(\log x_1 + \log x_2 + ... + \log x_n\right)$$

$$\therefore \quad GM = \text{Antilog}\left[\frac{\Sigma \log x}{n}\right]$$

For grouped data, the geometric mean is calculated with the following formula

$$GM = \text{Antilog}\left(\frac{\Sigma f \log x}{n}\right) \text{ where } n = \Sigma f$$

Geometric mean is specially useful in the construction of index numbers. This average is also useful in measuring the growth of population. GM is very helpful in averaging rates and percentages. However GM cannot be computed if any observation has either a value zero or negative.

Since the population increases according to compound rate, one may also obtain the average growth rate by using the compound interest formula.

$$P_n = P_0\left(1 + r\right)^n$$

Where, P_0 = Population in the begining of the period.

P_n = Population at the end of the period

r = Compound rate of growth

n = no. of periods.

1.3.6 The Harmonic Mean

The Harmonic mean is a measure of central tendency for data expressed as rates such as kilometers per hour, tonnes per day, km per litre etc. The harmonic mean is defined as the reciprocal of the arithmatic mean of the reciprocals of the individual observations. If X_1, X_2, ..., X_n are n boservations, then harmonic mean can be calculated using the following formula.

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

For grouped data, $HM = \dfrac{n}{\Sigma\left(\frac{f}{x}\right)}$ where $n = \Sigma f$

The Harmonic mean is useful for computing the average rate of increase of profits, or average speed at which a journey has been performed, or the average price at which an article has been sold. The HM is Computed from each and every observation. It is specially useful for averaging rates. However, HM cannot be computed when one or more observations have zero values, and when there are positive and negative observations. In dealing with business problems HM is rarely used.

Example : Find the GM and HM of the following distribution

x	f	f Log x	f/x
3	2	0.9542	0.6667
4	5	3.0105	1.2500
5	9	6.2910	1.8000
6	14	10.8948	2.3333
7	15	12.6765	2.1428
8	8	7.2248	1.0000
9	6	5.7252	0.6667
10	3	3.0000	0.3000
11	1	1.0414	0.0909
Total	**63**	**50.8184**	**10.2504**

$$GM = \text{antilog } \frac{50.8184}{63} = 6.406$$

$$\frac{1}{HM} = \frac{10.2504}{63} = 0.1627$$

$$\therefore \quad HM = 6.146$$

1.3.7 Relation among AM, GM and HM

In any distribution when the original items differ in size the value of AM, GM and HM would also differ and will be related as follows :

$$AM \geq GM \geq HM$$

i.e, arithmatic mean is greater than geometric mean and geometric mean is greater than harmonic mean. The equality Sign holds only if all the numbers x_1, x_2, ..., x_n are identical.

Suppose, a and b are two positive quantities such that $a \neq b$.

Then AM, GM and HM of these two quantities are

$$AM = \frac{a+b}{2} \quad ; \quad GM = \sqrt{a \times b} \text{ and } HM = \frac{2}{\frac{1}{a}+\frac{1}{b}} = \frac{2ab}{a+b}$$

We have to prove that, AM > GM > HM.

We know that the square of any real quantity is non-negative

Hence, $\left(\sqrt{a} - \sqrt{b}\right)^2 \geq 0$

or, $a + b - 2\sqrt{ab} \geq 0$

or, $a + b \geq 2\sqrt{ab}$

or, $\frac{a+b}{2} \geq \sqrt{ab}$.....(1)

$\therefore$ $AM \geq GM$ for n = 2 ..... (2)

Now multiplying both sides of (1) by $\frac{2\sqrt{ab}}{a+b}$

$$\frac{a+b}{2} \cdot \frac{2\sqrt{ab}}{a+b} \geq \sqrt{ab} \cdot \frac{2\sqrt{ab}}{a+b}$$

or, $\sqrt{ab} \geq \frac{2ab}{(a+b)}$

or, GM ≥ HM ..... (3)

Now combining (2) and (3) we may write

AM ≥ GM ≥ HM

The equality sign will hold when two observations are equal or same.

Multiplying AM and HM we get,

$AM \times HM$

$= \frac{a+b}{2} \times \frac{2ab}{a+b}$ Where, $AM = \frac{a+b}{2}$ and $HM = \frac{2ab}{a+b}$

$= ab$

$= \left(\sqrt{ab}\right)^2$

$= (GM)^2$

$\therefore \quad AM \times HM = (GM)^2$

1.4 Some Problems / Theorems on Central Tendency

Problem : **Find the AM, GM and HM of a, ar, ar^2,....ar^{n-1} Also check that A.M. × H.M. = $G.M^2$.**

Solution : Let x : a, ar, ar^2, ..., ar^{n-1}

$$\therefore \quad \Sigma x = a + ar + ar^2 + ... + ar^{n-1} = a \cdot \frac{r^n - 1}{r - 1}$$

$$\therefore \quad \bar{x} = \frac{\Sigma x}{n} = \frac{a}{n} \cdot \frac{r^n - 1}{r - 1} = A.M \quad \text{(Ans.)}$$

$$\text{or,} \quad \bar{x} = \frac{a}{n} \cdot \frac{1 - r^n}{1 - r}$$

$$GM = \left(a.ar.ar^2. .....ar^{n-1}\right)^{\frac{1}{n}}$$

$$= \left(a^n . r^{1+2+.....+(n-1)}\right)^{\frac{1}{n}} = \left(a^n . r^{\frac{(n-1)n}{2}}\right)^{\frac{1}{n}} = ar^{\frac{n-1}{2}} \text{ (Ans.)}$$

$$\text{H.M.} = \frac{n}{\frac{1}{a} + \frac{1}{ar} + \frac{1}{ar^2} + ... + \frac{1}{ar^{n-1}}}$$

$$= \frac{n}{\frac{r^{n-1} + r^{n-2} + ... + r + 1}{ar^{n-1}}}$$

$$= \frac{n.ar^{n-1}}{1 + r + ... + r^{n-2} + r^{n-1}} = \frac{n.ar^{n-1}}{\frac{1\left(r^n - 1\right)}{r - 1}}$$

$$= \frac{n.ar^{n-1}(r - 1)}{r^n - 1} \quad \text{(Ans.)}$$

$$\text{Now, A.M. × H.M.} = \frac{a\left(r^n - 1\right)}{n(r - 1)} \times \frac{n\,a\,r^{n-1}(r - 1)}{\left(r^n - 1\right)}$$

$$= a^2 r^{n-1} = \left(a.r^{\frac{n-1}{2}}\right)^2 = G.M^2 \quad \textbf{(Proved)}$$

Therem 2 : Relation maong AM, GM and HM is given by : AM ≥ GM ≥ HM

Relation (1) : For two observations only, $\frac{AM}{GM} = \frac{GM}{HM}$

or, $GM^2 = AM \times HM$

or, $GM = \sqrt{AM \times HM}$ i.e. for two observations, GM is the GM of AM & HM.

Proof : Let there be two observations, x_1 and x_2.

$$AM = \frac{x_1 + x_2}{2}, \quad GM = \sqrt{x_1 x_2}, \quad HM = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} = \frac{2x_1 x_2}{x_1 + x_2}$$

Now, $$AM \times HM = \frac{x_1 + x_2}{2} \times \frac{2x_1 x_2}{x_1 + x_2} = x_1 x_2 = \left(\sqrt{x_1 x_2}\right)^2 = GM^2$$

or, $GM = \sqrt{AM \times HM}$.

Thus, for two observations only, GM is the geometric mean of AM and HM.

Relation (2) : For any number of observations, n, $AM \geq GM \geq HM$

Proof : We know, $\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$

or, $x_1 + x_2 \geq 2\sqrt{x_1 x_2}$ or, $\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$ or, $AM \geq GM$ for n = 2

Similarly, $\frac{x_3 + x_4}{2} \geq \sqrt{x_3 x_4}$

Now, taking two quantities $\sqrt{\frac{x_1 + x_2}{2}}$ and $\sqrt{\frac{x_3 + x_4}{2}}$,

We have, $$\frac{\frac{x_1 + x_2}{2} + \frac{x_3 + x_4}{2}}{2} \geq \sqrt{\left(\frac{x_1 + x_2}{2}\right)\left(\frac{x_3 + x_4}{2}\right)}$$

From our previous result, $\frac{x_1 + x_2 + x_3 + x_4}{4} \geq \sqrt{\sqrt{x_1 x_2}\ \sqrt{x_3 x_4}}$

or, $\frac{x_1 + x_2 + x_3 + x_4}{4} \geq \sqrt[4]{x_1 x_2 x_3 x_4}$ i.e., $AM \geq GM$ for n = 4

In general, $AM \geq GM$ for N = 2^k values. But we have to prove the relation for any value of n.

Let us suppose, $2^{k-1} < n < 2^k$

We consider $2^k (= N)$ values of x. Among them, n observations are x_1, x_2, ..., x_n and the remaining (N – n) observations are all requal to A.

Let $A = \frac{x_1 + x_2 + ... + x_n}{n} = AM$ of n values.

Now, AM of N values $= \frac{x_1 + x_2 + ... + x_n + A + A + ...(N-n)\text{terms}}{N}$

$$= \frac{nA + (N-n)A}{N} = \frac{NA}{N} = A$$

GM of N values $= \left[x_1.x_2. .....x_n.A.A.....(N-n)\text{terms}\right]^{\frac{1}{N}} = \left[GM^n . AM^{N-n}\right]^{\frac{1}{N}}$

Since $AM \geq GM$ for N = 2K values, so, $AM \geq \left[GM^n AM^{N-n}\right]^{\frac{1}{N}}$

Raising both sides to the power N,

$AM^N \geq GM^n AM^{N-n}$ or, $AM^n \geq GM^n$ or, $AM \geq GM$ .....(1)

Let us consider the relation $GM \geq HM$. To prove this, we consider the values

$$\frac{1}{x_1}, \frac{1}{x_2}, ..., \frac{1}{x_n}$$

AM of these n values $= \frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{n} = \frac{1}{HM}$

GM of these n values $= \left(\frac{1}{x_1} \cdot \frac{1}{x_2} ..... \frac{1}{x_n}\right)^{\frac{1}{n}} = \frac{1}{\left(x_1.x_2.....x_n\right)^{\frac{1}{n}}} = \frac{1}{GM}$

We have already proved, $AM \geq GM$.

So, $\frac{1}{HM} \geq \frac{1}{GM}$, or, $GM \geq HM$ .....(2)

Combining (1) & (2) we have, $AM \geq GM \geq HM$

The equality sign will hold when all observations are same, say, equal to K.

Then $AM = \frac{K + K + ...n \text{ terms}}{n} = \frac{nK}{n} = K$

$$GM = \left(K.K. .....n \text{ terms}\right)^{\frac{1}{n}} = \left(K^n\right)^{\frac{1}{n}} = K$$

$$HM = \frac{n}{\frac{1}{K} + \frac{1}{K} + ...n \text{ terms}} = \frac{n}{\frac{n}{K}} = K$$

Thus, AM = GM = HM if all observations are same.

Therem 3 : For any set of observations, mean absolute deviation about median is minimum, i.e. $\frac{1}{n}\sum|x_i - A|$ = minimum if A = median.

Proof : $\frac{1}{n}\sum|x_i - A|$ is minimum if $\sum|x_i - A|$ = minimum.

Let us arrange the given observations $x_1, x_2, ..., x_n$ in increasing order of magnitude : $x_1, x_2, ..., x_n$. Thus, $x_1, x_2, ..., x_n$ is only a different arrangement of the same observations. So. $\sum|x_i - A| = \sum|x_i - A|$

Now, When n is odd, say, n = 2K + 1, we have

$$\sum_{i=1}^{n}|x_i - A| = \sum_{i=1}^{2K+1}|X_i - A| = |X_1 - A| + |X_2 - A| + ... + |X_K - A|$$

$$+|X_{K+1} - A| + ... + |X_{2K} - A| + |X_{2K+1} - A|$$

Now we make pairs taking the first term and the last term, the second term and the last but one term and so on. The first pair $|x_1 - A| + |x_{2K+1} - A|$ has the minimum value if A lies between x_1 and x_{2k+1}. So with the other pairs. As the number of terms is 2_{K+1}, there will be k pairs and a remaining term $|x_{K+1} - A|$. This has the minimum value zero only when $A = x_{K+1}$. But by definition, x_{K+1} is median as it occupies the middlemost position in the ordered series $x_1, x_2,....., x_{2k+1}$. So, A = median. Thus, $\sum_{i=1}^{n}|x_i - A|$ = minimum if A = median. This holds when n = odd.

When n is even, say, n = 2K,

$$\sum_{i=1}^{n}|x_i - A| = \sum_{i=1}^{2K}|X_i - A| = |X_1 - A| + |X_2 - A| + ... + |X_K - A|$$

$$+|X_{K+1} - A| + ... + |X_{2K-1} - A| + |X_{2K} - A|$$

Again we make pairs in the same way. We first take (K–1) pairs i.e., 2K–2 terms. There will be two middle most terms $|X_K - A|$ and $|X_{K+1} - A|$

Now, $|x_K - A| + |x_{K+1} - A|$ is minimum of A lies between two middle-most values X_K and X_{K+1}. If we take $A = \frac{X_K + X_{K+1}}{2}$, then also the RHS is minimum. But then, A is, by definition, median. Hence our theorem is proved i.e., $\sum_{i=1}^{n}|x_i - A|$ = minimum if A = median.

Theorem (4) : $S = f_K + (f_K + f_{K-1}) + (f_K + f_{K-1} + f_{K-2})$

$$+....+ (f_K + f_{K-1} + ... + f_2)$$

where $f_1, f_2, ..., f_k$ are the class frequencies, then $AM = x_1 + \frac{hS}{N}$ where x_1 = mid-point of the first class, h = class interval, N = total frequency.

Proof :

Mid-point	**Class fr**	**c.f. (>)**	$\mathbf{y = \frac{x - x_1}{h}}$	**fy**
x_1	f_1	$F_1 = f_1 + f_2 + ... + f_k$	0	0
x_2	f_2	$F_2 = f_2 + f_3 + ... + f_k$	1	f_2
x_3	f_3	$F_3 = f_3 + f_4 + ... + f_k$	2	$2f_3$
⋮	⋮	⋮	⋮	⋮
x_{K-1}	f_{K-1}	$F_{K-1} = f_{K-1} + f_K$	(K–2)	$(K-2)f_{K-1}$
x_K	f_K	$F_K = f_K$	(K–1)	$(K-1)f_K$
Total	$\Sigma f = N$	—	—	Σfy

$$\Sigma fy = f_2 + 2f_3 + 3f_4 + ... + (K-2)f_{k-1} + (k-1)f_k$$

$$= (f_2 + f_3 + ... + f_k) + (f_3 + f_4 + ... + f_k) + ... + (f_{k-1} + f_k) + f_k = S$$

Now, we know that if $y = \frac{x - x_1}{h}$, then $\bar{x} = x_1 + h\bar{y} = x_1 + h\frac{\Sigma fy}{N}$

$\therefore \quad \bar{x} = x_1 + \frac{hS}{N}$ [Proved]

Theorem (5) : For the variable x taking the values 1, 2,....., k with the frequencies f_1, f_2,....., f_k respectively, then $\bar{x} = \sum_{i=1}^{k} Fi/N$ where F_i is the cumulative frequency of greater than type and N is the total frequency.

Proof : We construct the frequency table to calculate $\bar{x}$.

$$\Sigma fx = f_1 + 2f_2 + ... + (K-1)f_{k-1} + Kf_k$$

$$= (f_1 + f_2 + ... + f_k) + (f_2 + f_3 + ... + f_k) + ... + (f_{k-1} + f_k)f_k$$

$$= F_1 + F_2 + ... + F_k$$

$$= \sum_{i=1}^{k} Fi$$

Now, $\bar{x} = \frac{\Sigma fx}{N} = \sum_{i=1}^{k} Fi/N$ [Proved]

x	f	c.f. (>)	fx
1	f_1	$F_1 = f_1 + f_2 + ... + f_k$	f_1
2	f_2	$F_2 = f_2 + f_3 + ... + f_k$	$2f_2$
3	f_3	$F_3 = f_3 + f_4 + ... + f_k$	$3f_3$
⋮	⋮	⋮	⋮
K–1	f_{K-1}	$F_{K-1} = f_{K-1} + f_K$	$(K-1)f_{K-1}$
K	f_K	$F_K = f_K$	Kf_K
	$\Sigma f = N$	—	Σfx

Theorem (6) : If M be the mode of a variable x and if y = a + bx, then prove that mode of y (= My) = a + bM.

Proof : Mode of $x = M = l_1 + \frac{d_1}{d_1 + d_2} \times w$

where l_1 = lower boundary of the modal class of x

w = width of the modal class of x.

$d_1 = f_0 - f_{-1}$ = fr of the modal class – fr. of the previous class.

$d_2 = f_0 - f_{+1}$ = fr of the modal class – fr. of the next class.

Now, we have the relation, y = a + bx.

So, lower boundary of the modal class of $y = a + bl_1$

and upper boundary of the modal class of $y = a + bl_2$

So, width of the modal class of $y = a + bl_2 - a - bl_1 = b(l_2 - l_1) = bw$.

Again, d_1 & d_2 will remain unchanged.

So, mode of $y = M_y = \left(a + bl_1\right) + \frac{d_1}{d_1 + d_2} \times bw$

$$= a + b\left[l_1 + \frac{d_1}{d_1 + d_2} \times w\right] = a + bM \quad \text{[Proved]}$$

The same argument will hold in the case of median also.

Thus, median of $y = \left(a + bl_1\right) + \frac{N/2 - F}{f_m} \times bw$

$$= a + b\left[l_1 + \frac{N/2 - F}{f_m} \times w\right] = a + bM_{e_x}$$

Thus, $M_{e_y} = a + bM_{e_x}$

Sum : The AM, GM and HM of 3 observations are 4, 3.63 and 3.27 respectively. What are the observations?

Solution : Let the three observations be x, y & z

So, $x + y + z = 3 \times 4 = 12 \quad \therefore x + y = 12 - z$ .....(1)

Again, $HM = \frac{3}{\frac{1}{x} + \frac{1}{y} + \frac{1}{z}} = 3{\cdot}27$

or, $\frac{1}{x}+\frac{1}{y}+\frac{1}{z}=\frac{3}{3\cdot27}$; or, $\frac{yz+zx+xy}{xyz}=\frac{3}{3\cdot27}$ .....(2)

Again, $GM=(xyz)^{\frac{1}{3}}=3\cdot63$ $\therefore$ $xyz=(3\cdot63)^3\simeq 48$

$\therefore$ $xy=\frac{48}{z}$ .....(3)

From (1), (2) and (3), $\frac{(x+y)z+xy}{xyz}=\frac{(12-z)z+\frac{48}{z}}{48}=\frac{3}{3\cdot27}$

or, $12z-z^2+\frac{48}{z}=\frac{48\times3}{3\cdot27}\simeq 44$

or, $12z^2-z^3+48=44z$

or, $z^3-12z^2+44z-48=0$

or, $z^3-2z^2-10z^2+20z+24z-48=0$

or, $z^2(z-2)-10z(z-2)+24(z-2)=0$

or, $(z-2)(z^2-10z+24)=0$

or, $(z-2)(z-4)(z-6)=0$

$\therefore$ z = 2, 4, 6, Then x = 6, 2, 4 and y = 4, 6, 2.

So, the observations are 2, 4 & 6.

1.5 Summary

At this stage it must be clear that no one average can be regarded as best for all circumstances.

If the data is badly skewed arithmatic mean should be avoided.

If the data is gappy around the middle, median should not be used.

If the class interval is unequal, mode should be avoided as a measure of central tendency. If we want to compare consumer preferences for different kinds of products or different kinds of advertising we can compare the modal preferences expressed by different groups of people.

Geometric mean is useful in averaging ratios and percentages and in computing average rates of increase or decrease.

Harmonic mean is useful in computing the average rate of increase in profits of a concern or average speed at which a journey has been performed or the average price at which an article has been sold.

1.6 Questions

1. What are the properties of arithmatic mean?
2. Compare the arithmatic mean, median, mode, geometric mean and Harmonic mean as a measure of central tendency.
3. How AM, GM and HM are related?
4. Write short note on quartiles, Deciles and percentiles.
5. Determine the mean, the median and the mode from the following figures :

   25, 15, 23, 40, 27, 25, 23, 25 and 20
6. From the table find out the median and Quartiles.

Size :	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
Frequency :	7	10	13	26	35	22	11	5

A. Select the correct Answer :

(i) For a moderately asymmetic distribution the mean and median are respectively 24.5 and 24.3. The mode is equal to—

(a) 24.3 (b) 24.1 (c) 23.9 (d) 24.2

(ii) The HM of 7 values 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$ and $\frac{1}{7}$ is

(a) 0.25 (b) 4 (c) 3.8 (d) None of these

(iii) Which one of the following is not a measure of location?

(a) mean (b) range (c) mode (d) median

(iv) If $2x = 7y$ be the relation between the variables x and y and GM of y is 1, then GM of x is—

(a) 3 (b) 3.5 (c) 7 (d) None of these

(v) The AM of 1, 3, 5,....., 29 is equal to

(a) 15 (b) 21 (c) 24 (d) 7.5

(vi) The marks of 5 students in a test in statistics are 10, 8, 68, 12 and 15. A suitable average of these marks is

(a) mean (b) median (c) mode (d) none of these

(vii) If the relation between two variables x and y be $5x + 7y = 28$ and median of y be 3, then the median of x is

(a) 3 (b) –2.4 (c) 4.2 (d) 1.4

(viii) The AM of 1, 2, 2^2,.....2^9 is

(a) 102.4 (b) 102.3 (c) 1024 (d) 1023

(ix) Mode depends on change of—

(a) origin (b) scale

(c) both origin and scale (d) neither origin nor scale

(x) A train ran at x km per hour from A to B and returned from B to A at y km per hour. The average speed (in km per hour) is

(a) $\frac{x+y}{2}$ (b) $\sqrt{xy}$ (c) $\frac{2xy}{x+y}$ (d) None of these

(xi) The weighted HM of first 11 natural numbers whose weights are equal to the corresponding numbers is

(a) 11 (b) 12 (c) $\frac{23}{3}$ (d) 6

(xii) The GM of the observations 2, 4, 8, 32 and 16 is

(a) 4 (b) 8 (c) 32 (d) none of these

B. Fill in the blanks :

(i) The AM of a set of 10 values is 5. If 2 is added to each value, the new AM of these 10 values is _____ .

(ii) The GM of the values 17, 8, 0, 5, 3 is _____ .

(iii) _____ quartile is meadian.

(iv) In a bimodal distribution _____ modes are available.

(v) In usual symbols, $M_0 = 3$ Med.– a.$\bar{x}$, where a = _____ .

(vi) The AM of first n natural numbers is _____ .

(vii) The abscissa of the point of intersection of less-than and more-than ogives gives the _____ .

(viii) There are _____ deciles and _____ percentiles.

1.7 References

1. Introduction to statistical calculations (1961)—J. Mounsey
   The English Universities Press.
2. Applied General Statistics by Croxton and Cowden, Prentice Hall
3. Mathematics of Statistics (1962) Vol. I by Kenney and Keeping, Van Nostrand Co.

Unit 2 ❐ Measures of Dispersion

Structure

2.1 Objectives

2.2 Introduction

2.3 Various measures of dispersion

2.3.1 The Range

2.3.2 Quartile Deviation or Semi-inter quartile Range

2.3.3 Average or Mean Deviation

2.3.4 Standard Deviation

2.3.5 The co-efficient of variation

2.3.6 The Lorenz Curve

2.4 Some Problems and Theorems on Dispersion

2.5 Summary

2.6 Questions

2.7 References

2.1 Objectives

The objectives of studying dispersion are the following :

(i) Measuring the variability determines the reliability of an average by pointing out as to how far an average is representative of the entire data.

(ii) Another objective of measuring variability is to determine the nature and cause of variation in order to control the variation itself.

(iii) Measures of variation enable comparison of two or more distributions with regard to their variability.

(iv) Measuring variability is of great importance to advanced statistical analysis. For example, sampling or statistical inference is essentially a problem in measuring variability.

A good measure of variation should possess, as far as possible, the same properties as those of a good measure of central tendency. The following are the main properties—

(i) It should be easy to understand

D:\NSOU\DTP\Statistic for Economic-2 (Alok)
Unit-2\ 6th proof\(Demay Submit)

(ii) It should be simple to compute.

(iii) It should be based on all observations.

(iv) It should be uniquely defined.

(v) It should be capable of further algebraic treatment

(vi) It should not be unduly affected by extreme values.

2.2 Introduction

Let us consider the marks of two groups of students

Group A	57	58	62	63	64	64	65	66	70	71
Group B	5	32	50	55	60	68	80	90	100	100

Here the mean marks for both the groups A and B is the same, that is 64. Also the median for both the groups is the same that is 64. Yet the marks of the Students in the two groups are quite different in the sense that the marks in group B are very much scattered from the mean marks. Thus a measure of central tendency alone is not sufficient to give a complete idea of the distribution and therefore to draw valid conclusions from the distribution we need some additional measures. One such measure is Dispersion. Dispersion literally means scatteredness. The study of dispersion enables us to know whether the distribution is homogeneous (as the marks in group A) or the distribution is non-homogeneous (as the marks in group B).

A measure of dispersion (or variation) describes the spread or scattering of the individual values around the central value. To illustrate the concept of variation, let us consider the data given below in five different series :

A	B	C	D	E
50	0	20	60	55
50	100	80	40	45

Since average for series A, B, C, D and E is the same, we are likely to conclude on the basis of the value of arithmatic average, that the distribution pattern of marks in all the series is the same. But a closer inspection reveals that the degree of variation is zero in series A, and the degree of variation is very high in series B. In series D the variation is comparatively low and in series E it is still lower. Therefore, different sets of data may have the same measure of central tendency but differ greatly in terms of variation. So the knowledge of central value is not enough to appreciate the nature of distribution of values. Hence a measure of variation (or dispersion) is absolutely necessary.

2.3 Various Measures of Dispersion

Following are the well known measures of variation which provide a numerical index of the variability of the given data.

1. Range
2. Average or Mean Deviation
3. Quartile Deviation or Semi-Interquartile Range
4. Standard Deviation
5. Lorenz curve.

Absolute and Relative Measures of Variation

Measures of variation may be either absolute or relative, Measures of absolute variation are expressed in terms of original data. In case two sets of data are expressed in different units of measurement, then the absolute measures of variation are not comparable. In such cases, measures of relative variation should be used.

A measure of relative variation is the ratio of a measure of absolute variation to an appropriate average. It is a numerical co-efficient free of unit. It should be noted that while computing the relative variation, the average used as base should be the same one from which the absolute deviations were measured. This means that arithmatic mean should be used with the standard deviation and either the arithmatic mean or median with the mean deviation.

2.3.1 The Range

The difference between the highest and the lowest value of the variate is called the range of the distribution. In symbols this may be indicated as

$$R = H - L$$

where, R = The Range, H = Highest value and L = Lowest value

The five series referred earlier

A	B	C	D	E
50	0	20	60	55
50	100	80	40	45

Calculation of Range

Group A Range = 50 – 50 = 0

Group B Range = 100 – 0 = 100

Group C Range = 80 – 20 = 60

Group D Range = 60 – 40 = 20

Group E Range = 55 – 45 = 10

The range is very easy to calculate and it gives some idea about the variability of the data. Hence it is one of the simplest measures of variability. However, the range is a crude measure of variation, since it uses only two extreme values, which are subject to chance fluctuations. Hence it is not a very reliable measure of dispersion. However, the concept of range is extensively used in statistical quality control. Range is helpful in studying the variations in the prices of shares and debentures and other commodities that are very sensitive to price changes from one period to another. For meteorological departments, the range is a good indicator for weather forcast.

For grouped data, the range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class. The relative measure corresponding to range (known as co-efficient of range) is obtained by applying the following formula

$$\text{Co-efficient of Range} = \frac{H - L}{H + L}$$

In fact, range is a crude mesure of variability and should be used carefully only where that data are fairly continuous and not irregular.

Calculation of the co-efficient of range

Class Interval	f
30–40	12
40–50	18
50–60	20
60–70	19
70–80	13
80–90	8

Range = H – L = 90 – 30 = 60

$$\text{Co-efficient of range} = \frac{H - L}{H + L} = \frac{90 - 30}{90 + 30} = \frac{60}{120} = \frac{1}{2} = 0{\cdot}5$$

2.3.2 Quartile Deviation or Semi-Interquartile Range

Another measure of dispersion which is fairly commonly used is called the Quartile Deviation. As the name suggests the quartile deviation is measured in terms of the quartiles of a given distribution. We know that the median of the distribution divides the total number of values into two equal halves. The median is that value of the variable which is right in the middle in the sense that the number of values less than

the median is equal to the number greater than the median. Similarly the quartiles of a distribution divide the distribution of values into four equal parts. Three points divide the series into four equal parts and these points are Q_1, Q_2 and Q_3. 25% of the total number of values will be less than Q_1, 25% will lie between Q_1 and Q_2, 25% will be between Q_2 and Q_3 and 25% will lie above Q_3. 50% of values are less than Q_2 (median) and 50% are greater than Q_2 (median). Q_1 and Q_3 are called lower and upper quartiles. It means 50% of the total number of values lie betwen Q_1 and Q_3. $(Q_3 - Q_1)$ is known as 'Inter Quartile Range'.

$$\text{Quartile Deviation (QD)} = \frac{Q_3 - Q_1}{2} = \text{Semi–Interquartile Range}$$

Although this measure avoids the effect of extreme values, it is not based on all the values in a series. It may be pointed out that in a symmetrical distribution, the value of the median lies midway between the two quartiles. Therefore, the lower and the upper quartiles should lie within quartile deviation distance from the median or is other words Median $\pm$ (Quartile Deviation) should equal Q_3 and Q_1. If however, a series is non-symmetrical as it is generally true of most economic series $\pm$ (QD) distance from the median would not give us the value of the two quartiles. However, we may expect to cover approximately 50% items, unless the skewness is very high.

Quartile Deviation is definitely a better measure of variation than the range, but still it does not take into account all the observations. So it cannot be regarded as a very reliable measure of dispersion.

Relative measure known as

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Quartile deviation is based on 50% observations, while range is based on two extreme values. Another advantage of quartile deviation is that it is the only mesure of variability which can be used for open-end distribution. The disadvantage of quartile deviation is that it ignores the first and the last 25% observations.

Example : Calculation of the co-efficient of Q.D.

Marks :	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
No. of students :	7	10	13	26	35	22	11	5

Series made exclusive :

Marks	No. of Students (f)	Cummulative Fequency
10·5–15·5	7	7
15·5–20·5	10	17
20·5–25·5	13	30
(25·5–30·5)	26	56
30·5–35·5	35	91
35·5–40·5	22	113
40·5–45·5	11	124
45·5–50·5	5	129

Q_1 = Measure of $= \frac{129}{4}$ th item

= 32·25th item which lies in the class (25·5 – 30·5)

$$Q_1 = l_1 + \frac{q_1 - C}{f} \times i$$

$$= 25{\cdot}5 + \frac{(32{\cdot}25 - 30)}{26} \times 5$$

$$= 25{\cdot}5 + \frac{2{\cdot}25 \times 5}{26}$$

$$= 25{\cdot}5 + \frac{11{\cdot}25}{26}$$

$$= 25{\cdot}5 + 0{\cdot}432 = 25{\cdot}9$$

Q_3 = Measure of $\frac{3N}{4}$ th item

= 96.75 which lies in the class (35.5 – 40.5)

$$= 35{\cdot}5 + \frac{96{\cdot}75 - 91}{22} \times 5$$

$$= 35{\cdot}5 + \frac{5{\cdot}75 \times 5}{22}$$

$$= 35{\cdot}5 + \frac{28{\cdot}75}{22}$$

$$= 35{\cdot}5 + 1{\cdot}3 = 36{\cdot}8$$

$$\text{Q.D} = \frac{Q_3 - Q_1}{2} = \frac{36{\cdot}8 - 25{\cdot}9}{2} = \frac{10{\cdot}9}{2} = 5{\cdot}45$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{36{\cdot}8 - 25{\cdot}9}{36{\cdot}8 + 25{\cdot}9} = \frac{10{\cdot}9}{62{\cdot}7} = 0{\cdot}174$$

2.3.3 Average or Mean Deviation

The mean deviation is an improvement over Range and Quartile Deviation, because, it considers all observations in the given set of data. This measure is computed as the mean of deviations from arthmatic mean or median.

All the deviations are treated as positive regardless of sign. If X_1, X_2, ..., X_n denote the values of the variate X, then Average or Mean Deviation is

$\text{AD} = \frac{\sum|X_i - \bar{X}|}{n}$ or, $\frac{\sum|X_i - \text{Median}|}{n}$. It is also called Mean absolute Deviation (MAD)

Theoretically speaking, there is an advantage in taking the deviations from median because the sum of the absulute deviations (ignoring positive and negative signs) from median is minimum. However, in actual practice arithmatic mean is more popularly used in the computation of mean deviation.

For grouped data, the formula to be used is

$$\text{Average Deviation (AD)} = \frac{\sum f|X_i - \bar{X}|}{\sum f} = \frac{\sum f|X_i - \bar{X}|}{N} \quad \text{where } N = \Sigma f.$$

Property of Average Deviation

Average Deviation from the median is less than that measured from any other value.

Let the variate be arranged in ascending order of magnitude. Let X be any arbitrary number.

Let us assume that m values of the variate X_1, X_2....X_m are less than X and the next n values of the variate X_{m+1}, X_{m+2},...., X_{m+n} are greater than X.

Let S be the sum of the absolute deviations of the variate from X.

$$\therefore \quad S = \sum_{i=1}^{m+n} \left| X_i - X \right|$$

$$\text{or,} \quad S = \sum_{i=1}^{m} \left(X - X_i \right) + \sum_{i=m+1}^{m+n} \left(X_i - X \right)$$

$$\therefore \quad \frac{\partial S}{\partial X} = \left(1 + 1 + ... + 1, \text{ m terms} \right) - \left(1 + 1 + ... + 1, \text{ n terms} \right)$$

$$= m - n$$

$\therefore \frac{\partial S}{\partial X}$ is negative if $m < n$

$\frac{\partial S}{\partial X}$ is positive if $m > n$

$\frac{\partial S}{\partial X}$ is zero if $m = n$

So we may say S is minimum when $m = n$

$\therefore$ X is the median because it is the middle most vlaue of the variate ($\because m = n$), while arranged in ascending or descending order of magnitude.

$\therefore$ The sum of the absolute deviations from the median will be the least.

The relative measure corresponding to the average deviation is called the co-efficient of average deviation. It is obtained by dividing average deviation by the particular average used in computing the average deviation. Thus if the average deviation has been computed from median, the co-efficient of average deviation shall be obtained by dividing the average deviation by the median.

Co-efficient of A.D. $= \frac{AD}{Median}$ or, $\frac{AD}{Mean}$

If one desires to measure and compare the variability among several sets of data, the average deviation may be used.

Example : Calculation of Mean Deviation

Height (in inches)	No. of Students	cf	| d |	f | d |
58	15	15	3	45
59	20	35	2	40
60	32	67	1	32
61	35	102	0	0
62	33	135	1	33
63	22	157	2	44
64	20	177	3	60
65	10	187	4	40
66	8	195	5	40
				334

Where, | d | is the absolute value of the deviation from Median.

Median = Measure of $\frac{N+1}{2}$ th item

= Measure of $\frac{195+1}{2}$ th item

= $\frac{196}{2}$ th item

= 98th item

= 61 inches

Mean Deviation $= \frac{\Sigma f \left| d \right|}{N} = \frac{334}{195}$ = 1·71 inches.

2.3.4 Standard Deviation

Standard Deviation satisfies most of the properties of a good measure of dispersion. It is also known as root-mean-square-deviation for the reason that it is the square root of the mean of the squared deviation from the arithmatic mean.

If X is the variate, $\bar{X}$ is the mean the standard deviation (σ) is

$$\sigma = \sqrt{\frac{1}{n}\sum\left(x_i - \bar{x}\right)^2} \quad .$$

For frequency data, $\sigma = \sqrt{\frac{1}{N}\sum f_i \left(X_i - \bar{X}\right)^2}$

where, $N = \Sigma f_i$

The square of the standard deviation is called the variance i.e, $\sigma^2 = \frac{1}{n}\sum \left(x_i - \bar{x}\right)^2$

or, for frequency data, $\sigma^2 = \frac{1}{N}\sum f_i \left(X_i - \bar{X}\right)^2$

Methods to calculate Standard Deviation

(i) Direct Method :

For simple series,

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

$$= \frac{1}{n}\sum \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) = \frac{\Sigma x_1^2}{n} - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{\Sigma x_i^2}{n} - \bar{x}^2 \quad = \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2$$

$$SD = \sigma = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

For frequency data :

$$\sigma = \sqrt{\frac{1}{N}\sum f_i \left(X_i - \bar{X}\right)^2}$$

$$= \sqrt{\frac{1}{N}\sum f_i \left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)} \quad = \sqrt{\frac{1}{N}\sum f_i X_i^2 - \frac{2\bar{X}}{N}\sum f_i X_i + \bar{X}^2 \cdot \frac{\Sigma f_i}{N}}$$

$$= \sqrt{\frac{1}{N}\sum f_i X_i^2 - 2\bar{X}\cdot\bar{X} + \bar{X}^2} \quad = \sqrt{\frac{1}{N}\sum f_i X_i^2 - \bar{X}^2}$$

$$= \sqrt{\frac{1}{N}\sum f_i X_i^2 - \left(\frac{1}{N}\sum f_i X_i\right)^2}$$

Short cut Method

$$\sigma = \sqrt{\frac{\Sigma f d_x^2}{\Sigma f} - \left(\frac{\Sigma f d_x}{\Sigma f}\right)^2}$$ where, $d_x = X - A$; A being the assumed mean

or, $$\sigma = \sqrt{\frac{\Sigma f d'^2_x}{\Sigma f} - \left(\frac{\Sigma f d'_x}{\Sigma f}\right)^2}$$

where, $$d'_x = \frac{X - A}{i} = \frac{d_x}{i}$$

i being the class interval.

Example : Calculation of Standard Deviation

Marks	No. of Students
More than 0	100
More than 10	90
More than 20	75
More than 30	50
More than 40	25
More than 50	25
More than 60	5
More than 70	0

Given the data we have to Calculate Standard Deviation

Class Interval	Mid Value (m)	f	dx	d′x	fd′x	fd′x²
0–10	5	10	–30	–3	–30	+90
10–20	15	15	–20	–2	–30	+60
20–30	25	25	–10	–1	–25	+25
30–40	35	25	0	0	0	0
40–50	45	0	10	1	0	0
50–60	55	20	20	2	40	80
60–70	65	5	30	3	15	45
		100			–30	300

Assumed Mean = 35 and class interval (i) = 10

$dx = m - 35;\ d'_x = \frac{m-35}{10};\ fd'^{2}_x = fd'_x.d'_x$

$$\sigma = \sqrt{\frac{\Sigma fd'^{2}_x}{N} - \left(\frac{\Sigma fd'_x}{N}\right)^2} \times i$$

$$= \sqrt{\frac{300}{100} - \left(\frac{-30}{100}\right)^2} \times 10$$

$$= \sqrt{3 - 0{\cdot}09} \times 10 \ = \sqrt{2{\cdot}91} \times 10 \ = 1{\cdot}7 \times 10 = 17$$

Example : A collar manufacturer is considering the production of a new style of collars to attract young men. The following statistics of neck circumferences are available based on measurements of a typical group.

Mid value (inches) :	12·5	13·0	13·5	14·0	14·5	15·0	15·5	16·0	16·5
No. of students :	4	19	30	63	66	29	18	1	1

Compute the standard Deviation and use the criteria $\bar{X} \pm 3\sigma$ to obtain the largest and smallest size of collar he should make in order to meet the need of practically all his customers having in mind that collars are worn on an average $\frac{3}{4}$ inches larger than neck size.

Assumed Mean = 14.5

x	**f**	$\mathbf{d_x}$	$\mathbf{fd_x}$	$\mathbf{fd_x^2}$
12·5	4	–2	–8	16·00
13·0	19	–1·5	–28·5	42·75
13·5	30	–1·0	–30·0	30·00
14·0	63	–0·5	–31·5	15·75
14·5	66	0	0	0·00
15·0	29	0·5	14·5	7·25
15·5	18	1·0	18·0	18·00
16·0	1	1·5	1·5	2·25
16·5	1	2·0	2·0	4·00
	231		–62·0	136·00

$$\bar{X} = A + \frac{\Sigma fd_x}{N} = 14{\cdot}5 + \frac{-62}{231} = 14{\cdot}5 - 0{\cdot}268 = 14.232$$

$$\sigma = \sqrt{\frac{\Sigma fd_x^2}{N} - \left(\frac{\Sigma fd_x}{N}\right)^2}$$

$$= \sqrt{\frac{136}{231} - \left(\frac{-62}{231}\right)^2}$$

$$= \sqrt{0{\cdot}517} = 0{\cdot}72$$

Biggest size $= \bar{X} + 3\sigma$

$= 14{\cdot}232 + 3 \times 0{\cdot}72$

$= 14{\cdot}232 + 2{\cdot}16$

$= 16{\cdot}392$

Smallest size $= \bar{X} - 3\sigma$

$= 14{\cdot}232 - 3 \times 0{\cdot}72$

$= 14{\cdot}232 - 2{\cdot}16$

$= 12{\cdot}072$

$\frac{3}{4}''$ Larger

$\therefore$ Smallest Collar $= 12{\cdot}072 + 0{\cdot}75 = 12{\cdot}82$

Biggest Collar $= 16{\cdot}392 + 0{\cdot}75 = 17{\cdot}14$

Combined Standard Deviation :

It is passible to calculate the combined standard deviation of two or more groups. Combined standard deviation for two groups is denoted by σ_{12} and is computed as follows :

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

Where, σ_1 = Standard deviation of the first group

σ_2 = Standard deviation of the second group

$d_1 = \bar{X}_1 - \bar{X}_{12}$

$d_2 = \bar{X}_2 - \bar{X}_{12}$

Example : Given the particulars of the distribution of the weight of boys and girls in a class find the standard deviation of the combined data :

	Boys	**Girls**
Number	100	50
Mean weight	60 kg	45 kg
Variance	9	4

Ans : For finding combined standard deviation, we have to calculate the combined mean first.

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$= \frac{100(60) + 50(45)}{100 + 50} = \frac{6000 + 2250}{150} = 55$$

Given data : $N_1 = 100$, $\sigma_1^2 = 9$, $N_2 = 50$, $\sigma_2^2 = 4$,

$$d_1 = \left|\bar{X}_1 - \bar{X}_{12}\right| = 60 - 55 = 5$$

$$d_2 = \left|\bar{X}_2 - \bar{X}_{12}\right| = \left|45 - 55\right| = 10$$

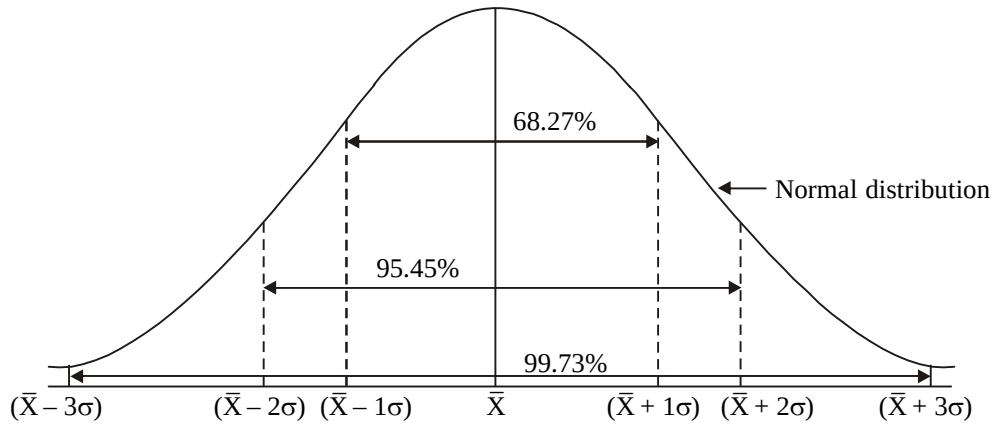
Substituting these values in

$$\text{Combined SD} = \sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100(9) + 50(4) + 100(5)^2 + 50(10)^2}{150}}$$

$$= \sqrt{\frac{900 + 200 + 2500 + 5000}{150}} = \sqrt{\frac{8600}{150}} = 7{\cdot}57$$

Distribution of the variate values in terms of mean and standard deviation in a frequency distribution which is normal.



The standard deviation enables us to determine, with a great deal of accuracy, where the values of the variate of a normal frequency distribution are located. According to Tchebycheff's theorem, for a symmetrical distribution, the following relationship will hold good.

(i) 68.27% of the total values of the variate will lie within $\left(\bar{X} \pm 1\sigma\right)$

(ii) 95.45% of the total values of the variate will lie within $\left(\bar{X} \pm 2\sigma\right)$

(iii) 99.73% of the total values of the variate will lie within $\left(\bar{X} \pm 3\sigma\right)$

Relation between Measures of Dispersion

In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. Quartile deviation is smallest, the mean deviation is next and the standard deviation is largest, in the following propertions :

$$\text{Q.D} = \frac{2}{3}\sigma \qquad \text{or,} \quad \sigma = \frac{3}{2}\text{Q.D}$$

$$\text{M.D} = \frac{4}{5}\sigma \qquad \text{or,} \quad \sigma = \frac{5}{4}\text{M.D}$$

Example : The mean and standard Deviation of normal distribution are 60 and 5 respectively.

Find the quartile Deviation and Mean Deviation of the distribution. Also find the inter-quartile range.

Ans. Given $\bar{X} = 60$ and $\sigma = 5$

$$\text{M.D} = \frac{4}{5}\sigma = \frac{4}{5} \times 5 = 4$$

$$\text{Q.D} = \frac{2}{3}\sigma = \frac{2}{3} \times 5 = \frac{10}{3}$$

$$\therefore \quad \text{Q.D} = \frac{Q_3 - Q_1}{2} = \frac{10}{3}$$

$$\text{or,} \quad Q_3 - Q_1 = \frac{20}{3} = 6{\cdot}67$$

$(Q_3 - Q_1)$ is the inter-quartile range where as

$\text{Q.D}\left(= \frac{Q_3 - Q_1}{2}\right)$ is the semi-inter-quartile range.

2.3.5 The co-efficient of Variation

The standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the co-efficient of variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the co-efficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which co-efficient of variation is less is said to be less variable or more consistent, more uniform, more stable or more homogeneous.

Co-efficient of variation or $C.V = \frac{\sigma}{\bar{X}} \times 100\,\%$

where, σ is standard deviation and $\bar{X}$ = arithmatic mean.

Example : From the prices of shares X and Y, find out which share is more stable.

X	35	54	52	53	56	58	52	50	51	49
Y	108	105	105	105	106	107	104	103	104	101

Calculation of co-efficient of variation

X	$\mathbf{(X - \bar{X})}$ **x**	$\mathbf{x^2}$	**Y**	$\mathbf{(Y - \bar{Y})}$ **y**	$\mathbf{y^2}$
35	–16	256	108	+3	9
54	+3	9	107	+2	4
52	+1	1	105	0	0
53	+2	4	105	0	0
56	+5	25	106	+1	1
58	+7	49	107	+2	4
52	+1	1	104	–1	1
50	–1	1	103	–2	4
51	0	0	104	–1	1
49	–2	4	101	–4	16
$\Sigma X = 510$	$\Sigma x = 0$	$\Sigma x^2 = 350$	$\Sigma Y = 1050$	$\Sigma y = 0$	$\Sigma y^2 = 40$

Co-efficient of Variation of X : $C.V = \frac{\sigma}{\overline{X}} \times 100$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{510}{10} = 51$$

$$\sigma = \frac{\sqrt{\Sigma x^2}}{N} = \frac{\sqrt{350}}{10} = 5.916$$

$$\therefore \quad CV = \frac{5.916}{51} \times 100 = 11.6\%$$

Co-efficient of Variation of Y : $CV = \frac{\sigma}{\overline{Y}} \times 100$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{1050}{10} = 105$$

$$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{40}{10}} = 2$$

$$\therefore \quad CV = \frac{2}{105} \times 100 = 1.905$$

Since co-efficient of variation is much less in case of Shares Y, hence they are more stable in value.

2.3.6 Lorenz Curve

Initially the Lorenz curve was used to measure the distribution of wealth and income. Now the curve is also used to study the distribution of profits, wages, turnover etc. However, still the most common use of this curve is in the study of the degreee of inequality in the distribution of income and wealth between countries or between different periods of time. It is a cumulative percentage curve in which the percentage of items is combined with the percentage of other things as wealth, profits, turnover, etc.

While drawing the Lorenz curve the following steps are adopted :

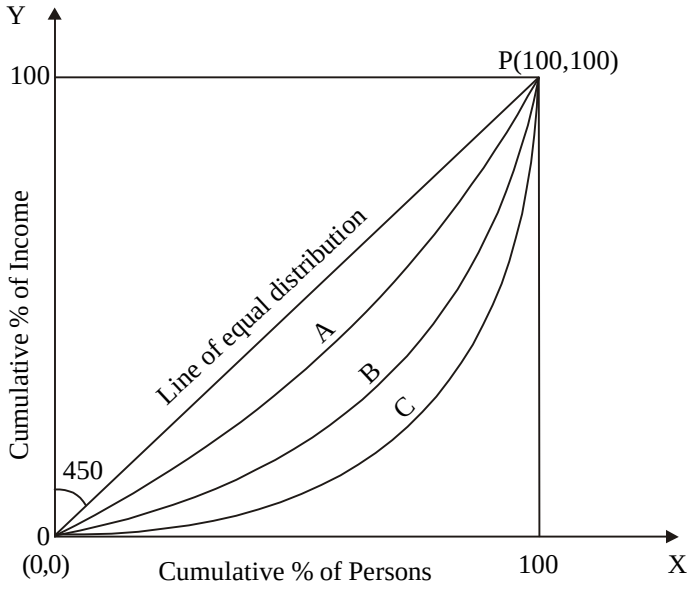
(i) The size of items (Variable values) and frequencies are both cumulated. Taking grand total for each as 100, percentages are obtained for these various cumulative values.

(ii) On the x-axis Start from 0 to 100 and take the percent of cumulative frequencies.

(iii) On the Y - axis Start from 0 to 100 and take the percent of cumulated values of the variable.

(iv) Draw a diagonal line joining 0 (o,o) with the point P (100,100) as shown in the diagram below. The line OP will make an angle of 45º with the Y - axis and is called the line of equal distribution. Any pint on this diagonal shows that same percent on X as on Y axes.

(v) Plot the percentages of the cumulated values of the variable (Y) against the percentages of the corresponding cumulated frequencies (X) for the given distribution and join these points with a smooth freehand curve. For any given distribution this will never cross the line of equal distribution OP. It will always lie below OP unless the distribution is uniform in which case it will coincide with OP. The greater the variability, the greater is the distance of the curve from OP.



Lonez Curve

In the above diagram OP is the line of equal distribution. The points lying on the curve OAP indicate a less degree of variability as compared to the points lying on the curve OBP. When the points lie on the curve OCP, variability is still greater. Thus a measure of variability of the distribution is provided by the distance of the curve of the cumulated percentages of the given distribution from the line of equal distribution.

2.4 Some Problems and Theorems on Dispersion

PROPERTIES OF SD :

1. If the given values of x are all equal, then its SD is zero.

Proof : Let $x_i = c$ for i = 1, 2, ..., n

$$\therefore \quad \bar{x} = \frac{\Sigma x_i}{n} = \frac{nc}{n} = c$$

$$\therefore \quad \left(x_i - \bar{x}\right) = 0 \text{ for each i.}$$

$$\therefore \quad \Sigma\left(x_i - \bar{x}\right)^2 = 0 \text{ So, } SD = \sqrt{\frac{1}{n}\sum\left(x_i - \bar{x}\right)^2} = 0$$

2. SD is independent of origin but not of scale.

Proof : Let $y_i = \frac{x_i - c}{d}$ where c & d are arbitriary constants.

Then, $x_i = c + dy_i$

$\therefore \quad \bar{x} = c + d\bar{y}$

Substracting, $\left(x_i - \bar{x}\right) = d\left(y_i - \bar{y}\right)$

Squarring both sides and summing for all i from 1 to n and then dividing by n, we get,

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

$$\therefore \quad \sqrt{\frac{1}{n}\Sigma\left(x_i - \bar{x}\right)^2} = \sqrt{d^2 \cdot \frac{1}{n}\Sigma\left(y_i - \bar{y}\right)^2}$$

$$\therefore \quad \sigma_x = |d|\sigma_y .$$

Thus, SD is independent of changes in origin but not of scale.

3. Let there be two sets of values of x with n_1 and n_2 values. Let $\bar{x}_1$ and $\bar{x}_2$ be their means and σ_1, σ_2 be their standard deviations respectively. Then SD of x for the two sets pooled together (σ) is given by, $\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$

or, $N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2$ where

$\therefore \quad d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $N = n_1 + n_2$

Proof : Let x_{1i} (i = 1, 2, ..., n_1) and x_{2j} (j = 1, 2, ..., n_2) be the observations of the two sets respectively.

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

Now, $$\sum_{i=1}^{n_1}\left(x_{1i}-\bar{x}\right)^2=\sum_{i=1}^{n_1}\left[\left(x_{1i}-\bar{x}_1\right)+\left(\bar{x}_1-\bar{x}\right)\right]^2$$

$$=\sum_{i=1}^{n_1}\left(x_{1i}-\bar{x}_1\right)^2+2\left(\bar{x}_1-\bar{x}\right)\sum_{i=1}^{n_1}\left(x_{1i}-\bar{x}_1\right)+n_1\left(\bar{x}_1-\bar{x}\right)^2$$

$$=n_1\sigma_1^2+n_1d_1^2 \text{ as } \sum_{i=1}^{n_1}\left(x_{1i}-\bar{x}_1\right)=0$$

Similarly, $\sum_{j=1}^{n_2}\left(x_{2j}-\bar{x}\right)^2=n_2\sigma_2^2+n_2d_2^2$

Putting these values, we get,

$$\sigma^2=\frac{n_1\sigma_1^2+n_2d_1^2}{n_1+n_2}+\frac{n_2\sigma_2^2+n_2d_2^2}{n_1+n_2}$$

$$=\frac{n_1\sigma_1^2+n_1d_1^2+n_2\sigma_2^2+n_2d_2^2}{n_1+n_2}$$

$\therefore$ $N.\sigma^2=n_1\sigma_1^2+n_2\sigma_2^2+n_1d_1^2+n_2d_2^2$ where $N=n_1+n_2$.

The theorem can be generalised for any number of groups of observations. That is, for k number of groups of observations,

$$N.\sigma^2=n_1\sigma_1^2+n_2\sigma_2^2+.....+n_k\sigma_k^2+n_1d_1^2+n_2d_2^2+.....+n_kd_k^2$$

where $N=n_1+n_2+.....+n_k$.

4. SD is the minimum-root mean-square deviation

i.e, $\sqrt{\frac{1}{n}\Sigma\left(x_i-\bar{x}\right)^2}\le\sqrt{\frac{1}{n}\Sigma\left(x_i-A\right)^2}$ whatever be the value of A.

Proof. : $\Sigma\left(x_i-A\right)^2=\Sigma\left[\left(x_i-\bar{x}\right)+\left(\bar{x}-A\right)\right]^2$

$$=\Sigma\left(x_i-\bar{x}\right)^2+2\left(\bar{x}-A\right)\Sigma\left(x_i-\bar{x}\right)+n\left(\bar{x}-A\right)^2$$

$$=\Sigma\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-A\right)^2 \text{ as } \Sigma\left(x_i-\bar{x}\right)=0$$

Now, as $n\left(\bar{x}-A\right)^2\ge 0$, we may write,

$$\Sigma\left(x_i - A\right)^2 \geq \left(x_i - \bar{x}\right)^2$$

or, $$\frac{1}{n}\Sigma\left(x_i - A\right)^2 \geq \frac{1}{n}\Sigma\left(x_i - \bar{x}\right)^2$$

$$\therefore \quad \sqrt{\frac{1}{n}\Sigma\left(x_i - A\right)^2} \geq \sqrt{\frac{1}{n}\Sigma\left(x_i - \bar{x}\right)^2}$$

Thus SD is the minimum root-mean square deviation.

Theorem 1 : Show that $\frac{1}{n}\Sigma\left(x_i - a\right)^2$ minimum if $a = \bar{x}$.

Proof : $\frac{1}{n}\Sigma\left(x_i - a\right)^2$ = Min. if $\Sigma\left(x_i - a\right)^2$ = minimum.

Now, $\Sigma\left(x_i - a\right)^2 = \Sigma\left[\left(x_i - \bar{x}\right) + \left(\bar{x} - a\right)\right]^2$

$$= \Sigma\left(x_i - \bar{x}\right)^2 + 2\left(\bar{x} - a\right)\Sigma\left(x_i - \bar{x}\right) + n\left(\bar{x} - a\right)^2$$

$$= \Sigma\left(x_i - \bar{x}\right)^2 + n\left(\bar{x} - a\right)^2 \text{ as } \Sigma\left(x_i - \bar{x}\right) = 0$$

Thus, $\Sigma\left(x_i - a\right)^2$ is the sum of two square quantities.

Thus expression has the minumum value if

$$n\left(\bar{x} - a\right)^2 = 0 \quad \text{or, if } \bar{x} - a = 0 \text{ or, if } a = \bar{x}$$

Thus, $\frac{1}{n}\Sigma\left(x_i - a\right)^2$ = minimum if $a = \bar{x}$

Theorem 2 : Show that the combined SD of two distributions pooled together is given by, $N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 n_2}{N}\left(\bar{x}_1 - \bar{x}_2\right)^2$

Proof : Here n_1, σ_1 and $\bar{x}_1$ represent the number of observations, SD and mean of the first group; n_2, σ_2 and $\bar{x}_2$ are those of the second group; N and σ represent, the total number of observations and SD of the combined group, respectively. Clearly, $N = n_1 + n_2$.

We know, $N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2$ .....(1)

Here $d_1 = \bar{x}_1 - \bar{x} = \bar{x}_1 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$= \frac{n_2\bar{x}_1 - n_2\bar{x}_2}{n_1 + n_2} = \frac{n_2\left(\bar{x}_1 - \bar{x}_2\right)}{n_1 + n_2}$$

Similarly, $d_2 = \dfrac{n_1\left(\bar{x}_2 - \bar{x}_1\right)}{n_1 + n_2}$

$$\therefore \quad n_1 d_1^2 + n_2 d_2^2 = \frac{n_1 n_2^2\left(\bar{x}_1 - \bar{x}_2\right)^2}{\left(n_1 + n_2\right)^2} + \frac{n_2 n_1^2\left(\bar{x}_1 - \bar{x}_2\right)^2}{\left(n_1 + n_2\right)^2}$$

$$= \frac{n_1 n_2^2\left(\bar{x}_1 - \bar{x}_2\right)^2 + n_2 n_1^2\left(\bar{x}_1 - \bar{x}_2\right)^2}{\left(n_1 + n_2\right)^2}$$

$$= \frac{n_1 n_2\left(n_1 + n_2\right)\left(\bar{x}_1 - \bar{x}_2\right)^2}{\left(n_1 + n_2\right)^2} = \frac{n_1 n_2\left(\bar{x}_1 - \bar{x}_2\right)^2}{N}$$

Putting this value in (1), we get,

$$N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 n_2}{N}\left(\bar{x}_1 - \bar{x}_2\right)^2 \text{ (Proved)}$$

Theorem 3 : MD Cannot exceed SD.

Proof : Variance of a set of objervations $x_i (i = 1, 2, ..., n)$ is given by,

$$\sigma^2 = \frac{1}{n}\Sigma\left(\bar{x}_i - \bar{x}\right)^2 \geq 0$$

$$\text{or,} \quad \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2 \geq 0 \quad \text{or,} \quad \frac{\Sigma x_i^2}{n} \geq \left(\frac{\Sigma x_i}{n}\right)^2$$

This inequality holds for any given set of real numbers, $x_1, x_2, ... x_n$. In particular we put $x_i = \left|Y_i - \bar{Y}\right|$

$$\therefore \quad \frac{\Sigma\left\{\left|Y_i - \bar{Y}\right|\right\}^2}{n} \geq \left[\frac{\Sigma\left|Y_i - \bar{Y}\right|}{n}\right]^2$$

$$\text{or,} \quad \frac{\Sigma\left(Y_i - \bar{Y}\right)^2}{n} \geq \left[\frac{\Sigma\left|Y_i - \bar{Y}\right|}{n}\right]^2 \quad \text{or,} \quad SD_Y^2 \geq MD_Y \quad \text{or,} \quad SD \geq MD$$

Thus, MD cannot exceed SD. The equality sign will hold when all observations are equal.

In that case, MD = SD = 0.

Theorem 4 : Prove that $n\left(x_1^2 + x_2^2 + ... + x_n^2\right) \geq \left(x_1 + x_2 + ... + x_n\right)^2$

Proof : We know hat variance can never by negetive

$$\therefore \quad \frac{1}{n}\Sigma\left(x_i - \bar{x}\right) \geq 0$$

$$\text{or,} \quad \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2 \geq 0$$

$$\text{or,} \quad \frac{\Sigma x_i^2}{n} \geq \left(\frac{\Sigma x_i}{n}\right)^2$$

Multiplying both sides by n^2, we get,

$$n\Sigma x_i^2 \geq \left(\Sigma x_i\right)^2$$

$$\text{or,} \quad n\left(x_1^2 + x_2^2 + ... + x_n^2\right) \geq \left(x_1 + x_2 + ... + x_n\right)^2 \text{ (Proved)}$$

Theorem 5 : Show that $MD_A = \frac{1}{n}\left[\left(S_2 - S_1\right) + \left(n_1 - n_2\right)A\right]$ where S_1 is the sum of observations that are less than A, n_1 is the number of such observations. S_2 is the sum of observations that are greater than A and n_2 is the number of such observations and n is the total number of observations.

Proof : $MD_A = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - A\right|$

Let us arrange the observations x_1, x_2,.....x_n in ascending order of magnitude :

$$X_1, X_2, ..., X_{n_1}, X_{n_1+1}, ..., X_{n_1+n_2}\left(= X_n\right)$$

$$\text{So,} \quad \sum_{i=1}^{n}\left|x_i - A\right| = \sum_{i=1}^{n}\left|X_i - A\right| = \sum_{X_i<A}^{n_1}{}_1\left|X_i - A\right| + \sum_{X_i>A}^{n_1+n_2}{}_{n_1+1}\left|X_i - A\right|$$

$$= \sum_{i=1}^{n_1}\left(A - X_i\right) + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i - A\right)$$

$$= n_1A - S_1 + S_2 + n_2A = \left(S_2 - S_1\right) + \left(n_1 - n_2\right)A$$

Now, $MD_A = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$

$\therefore \quad MD_A = \frac{1}{n}\left[(S_2 - S_1) + (n_1 - n_2)A\right]$ **[Proved]**

■ **Calculate Mean and S.D. of standard natural numbers.**

Ans. Here x : 1, 2, 3, ..., n

$\therefore \sum x = 1 + 2 + 3 + ... + n = \frac{n(n+1)}{2} \quad \therefore$ Mean $= \frac{\sum x}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$

$\sum x^2 = 1^2 + 2^2 + 3^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$

$\therefore$ Variance $= \sigma^2 = \frac{x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2$

$= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{n} = \frac{2(n+1)(2n+1) - 3(n+1)^2}{12}$

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

$\therefore \ SD = \sigma = \sqrt{\frac{n^2-1}{12}}$

■ **Calculate mean and S.D for first n odd numbers.**

Ans. Here x : 1, 3, 5, ..., (2n – 1),

Now, $\sum x = 1 + 3 + 5 + ... + (2n-1) = \sum_{r=1}^{n}(2r-1)$

$= 2\sum_{r-1}^{n} r - n = \frac{2n(n+1)}{2} - n = n(n+1) - n = n^2$

$\therefore \ \bar{x} = \frac{1}{n}\sum x = \frac{1}{n}.n^2 = n$

$\sum x^2 = 1^2 + 3^2 + 5^2 + ... + (2n-1)^2$

$$= \sum_{r=1}^{n}(2r-1)^2 = \sum_{r=1}^{n}(4r^2-4r+1) = 4\sum_{r=1}^{n}r^2 - 4\sum_{r=1}^{n}r + n$$

$$= \frac{4n(n+1)(2n+1)}{6} - \frac{4n(n+1)}{2} + n = \frac{2n(n+1)(2n+1)}{3} - 2n(n+1) + n$$

$$\therefore \ \frac{\sum x^2}{n} = \frac{2(n+1)(2n+1)}{3} - 2(n+1) + 1$$

$$= \frac{2(n+1)(2n+1) - 6(n+1) + 3}{3} = \frac{(n+1)(4n+2-6)+3}{3}$$

$$= \frac{4(n+1)(n-1)+3}{3} = \frac{4n^2-4+3}{3} = \frac{4n^2-1}{3}$$

$$\therefore \ \sigma^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{4n^2-1}{3} - n^2 = \frac{4n^2-1-3n^2}{3} = \frac{n^2-1}{3}$$

$$\therefore \ SD = \sqrt{\frac{n^2-1}{3}}$$

■ **Calculate mean and S.D of First n even numbers**

Here x : 2, 4, 6, ..., 2n

$$\therefore \sum x = 2+4+6+...+2n = 2(1+2+3+..+n) = \frac{2n(n+1)}{2} = n(n+1)$$

$$\therefore \ \bar{x} = \frac{\sum x}{n} = (n+1)$$

$$\sum x^2 = 2^2 + 4^2 + 6^2 + ... + (2n)^2$$

$$= 2^2.1^2 + 2^2.2^2 + 2^2.3^2 + ... + 2^2.n^2$$

$$= 2^2(1^2+2^2+3^2+ ...+n^2) = \frac{4.n(n+1)(2n+1)}{6}$$

$$\therefore \text{ Variance } = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{2(x+1)(2n+1)}{3} - (n+1)^2$$

$$= \frac{2(n+1)(2n+1) - 3(n+1)^2}{3} = \frac{(n+1)(4n+2-3n-3)}{3}$$

$$= \frac{(n+1)(n-1)}{3} = \frac{n^2-1}{3} \qquad \therefore \ S.D = \sqrt{\frac{n^2-1}{3}}$$

Comment 1 : Here the values of observations are 2, 4, 6, ... 2n i. e., 2(1, 2, 3, ..., n) i.e., each value is twice of first n natural numbers. So the S.D. of first n even

numbers will be twice of the S.D. of first n natural numbers. The S.D of first n natural numbers = $\sqrt{\frac{n^2-1}{12}}$.

So, S.D. of first n even numbers $= 2.\sqrt{\frac{n^2-1}{12}} = \sqrt{\frac{4(n^2-1)}{12}} = \sqrt{\frac{n^2-1}{3}}$

Comment 2 : We know that S.D of first n odd numbers is $\sqrt{\frac{n^2-1}{3}}$. i,e., S.D of 1, 3, 5, ..., (2n -1) is $\sqrt{\frac{n^2-1}{3}}$. Now, What will be the S.D. of first n even numbers i.e., S.D. of 2, 4, 6, ..., 2n = ?. We should note that the series of even numbers is obtained just by adding 1 to each number of the first series. Thus, if x is the series of the odd numbers and y is the series of the even numbers then y = x+1. We know that S.D. is independent of change in origin. So, S.D. of first n even numbers = S.D of first n odd numbers = $\sqrt{\frac{n^2-1}{3}}$. Similarly, $\bar{y} = \bar{x} + 1 = (n+1)$ i,e., mean of first n even numbers = mean of first n odd numbers +1 = n + 1

Some other Relative measures of dispersion

1) Coefficient of dispersion based on range $= \text{C.D.} = \frac{x\,\max - x\,\min}{x\,\max + x\,\min}$

2) Coefficient of dispersion based on Q.D. = $\text{C.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

3) Coefficient of dispersion based on M.D = $\text{C.D} = \frac{\text{M.D}}{\text{Average used in M.D}}$

2.5. Summary

If the available data are few in number or contain extreme values, avoid the standard deviation. If the data set is generally skewed, avoid the mean deviation as well. If the data have gaps around the quartiles, the quartile deviation should be avoided. If there are open-end classes, the quartile measure of dispersion should be preferred.

In an elementary treatment of statistical series in which a measure of variability is desired only for itself, any of the three measures viz., range, quartile deviation and mean deviation would be acceptable. Probably the mean deviation would be better. However, in usual pracitic, the measure of variability is employed in further statistical

analysis. For such a purpose the standard deviation, by far, is the most popularly used. Standard deviation is free from those defects from which other measures suffer. It lends itself to the analysis of variability in terms of normal curve of error. Practically all advanced statistical methods deal with variability and centre around the standard deviation. Hence, unless the circumstances warrant the use of any other measure, we should make use of standard deviation for measuring variability.

2.6. Questions

1. What do you mean by Dispersion?
2. Define Range
3. Define Mean Deviation
4. Define Quartile Deviation
5. Define standard Deviation
6. Define co-efficient of Variation
7. Explain the appropriateness of different measures of dispersion in different circumstances.
8. Explain the lorenz curve in measuring the inequality of income.
9. Find the co-efficient of variation from the following table :

Class boundaries	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120
Frequency	4	8	9	16	6	3	2	2

10. Prove that the sum of the absolute deviations from the median will be the least.
11. Calculate the standard deviation from the following data :

Age	5–7	8–10	11–13	14–16	17–19
No. of Students	7	12	19	10	2

12. For a group of 200 candidates the mean and standard deviation were found to be 40 and 15. Later on it was discovered that the score 43 was misread as 53. Find the correct mean and standard deviation. [Ans. $\bar{X} = 39.95$, $\sigma = 14.97$]
13. Calculate the combined average and combined standard deviation.

	A Series	B Series
No. of items	100	500
Mean	50	60
Variance	100	121

14. Select the correct Answer :

(i) If AM and co-efficient of variation of x are 6 and 50% respectively, then variance of x is

(a) 3 (b) 6 (c) 9 (d) None of these

(ii) If the relation between variables x and u, and first quartile of x are 3x+4u=21 and minus 1 respectively, then the first quartile of u is

(a) 0.75 (b) 6 (c) 7 (d) None of these.

(iii) Sum of the absolute deviations is minimum about

(a) mean (b) median (c) mode (d) none of these

(iv) Co-efficient of variation is equal to

(a) $\frac{s}{\bar{x}} \times 100\%$ (b) $\frac{\bar{x}}{s} \times 100\%$ (c) $\frac{\bar{x}}{s}$ (d) $\frac{s^2}{\bar{x}}$

(v) The range of observations –6, –9, –8, –1, –4, is

(a) 2 (b) –2 (c) –8 (d) 8

(vi) The S.D of x_1, x_2, x_3, x_4, x_5, x_6 and x_7 is S, then the S.D of $-x_1$, $-x_2$, ..., $-x_7$ is

(a) –s (b) s (c) o (d) 2s

(vii) SD is dependent on change of

(a) Origin only (b) Scale only (c) both (a) and (b) (d) None of these

(viii) If $\bar{x}$ and s stand for mean and SD of x then the SD of $\left(\frac{x - \bar{x}}{s}\right)$ will be—

(a) 1 (b) 0 (c) $1/s$ (d) none of these

2.7. References

1. Fundramentals of Statistics, vol I by Goon, Gupta and Dasgupta; The world Press.

Unit 3 ❐ Moments, Skewness and Kurtosis

Structure

3.1 Objectives

3.2 Introduction

3.3 Moments

3.3.1 Moments about arbitrary origin

3.3.2 Central moments expressed in terms of raw moments

3.3.3 Sheppard's Correction and charlier's check

3.3.4 Moments in Dimensionless form

3.3.5 Computation of Moments for grouped data

3.4 Measures of Skewness

3.4.1 Absolute measures of Skewness

3.4.2 Relative measures of Skewness

3.4.3 Karl Pearson's Co-efficient of Skewness

3.4.4 Bowley's co-efficient of skewness

3.4.5 Kelly's co-efficient of Skewness

3.4.6 Measure of Skewness based on third moment

3.5 Kurtosis

3.6 Measures of Kurtosis

3.7 Some Theorems on Moments, Skewness and Kurtosis

3.8 Summary

3.9 Questions

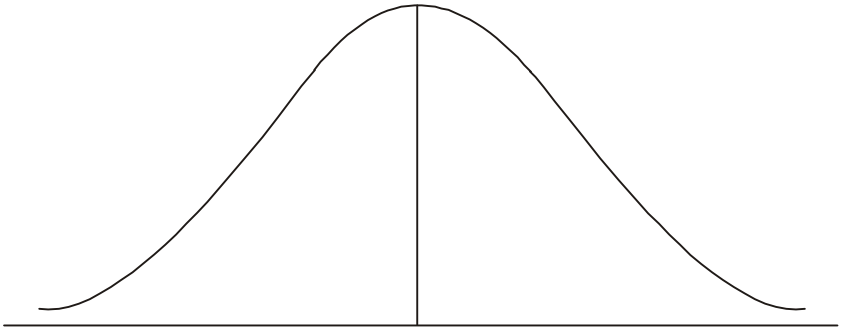
3.10 References

3.1 Objectives

The main objective of Studying Skewness and Kurtosis is to Know the shape of the frequency distribution more accurately. Measures of central tendency and the measures of variability cannot help us to draw the frequency distribution curve correctly. They may help us to find out the position of the curve. But the proper shape will be elusive till we know the skewness and Kurtosis of the frequency distribution. While Skewness helps us to know whether the curve is symmetrical or not, Kurtosis determines the peakedness of the curve. Two distributions may have the same mean and standard deviation but they may differ widely in shape if they have different Skewness and Kurtosis.

3.2 Introduction

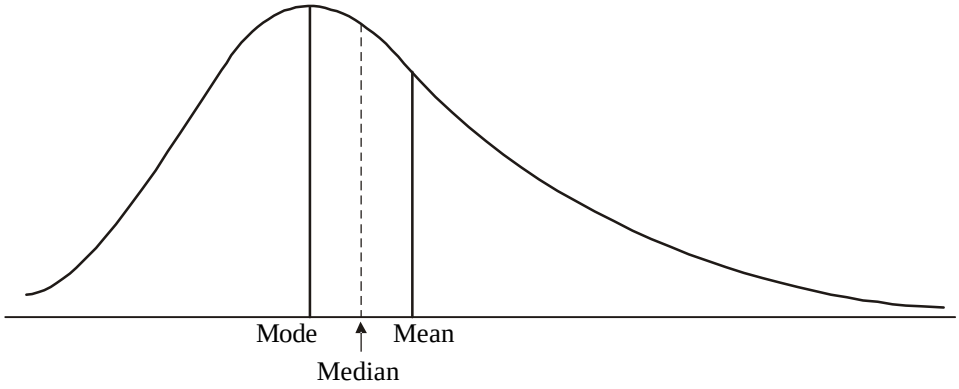
The term Skewness refers to lack of symmetry. When a distribution is not symmetrical it is called a skewed distribution.



Mean = Median = Mode
Symmetrical (normal) Distribution

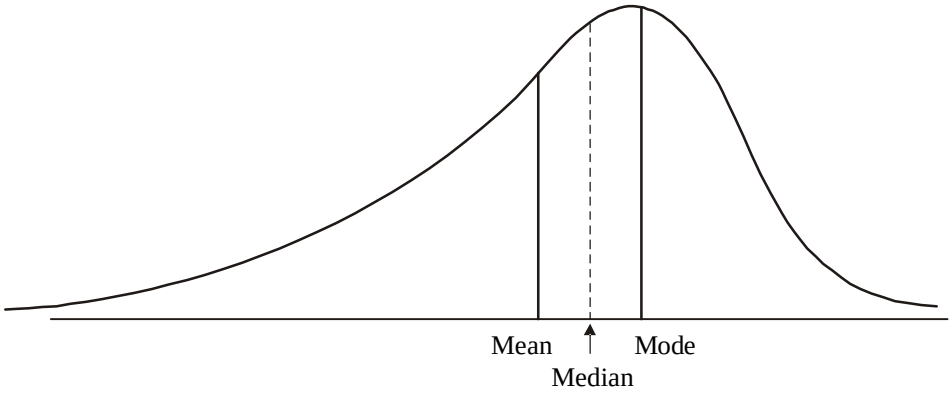
It is clear from the diagram that in a symmetrical distribution, the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.

A distribution which is not symmetrical will have mean $\neq$ median $\neq$ mode. Such a distribution is also called ''Asymmetrical Distribution''. Asymmetrical distribution could either be positively skewed or negatively skewed.



Positively Skewed Distribution

In a positively skewed distribution, the value of the mean is maximum and that mode is the least. The median lies in between the two. The long tail is on the right side and the hump is on the left.



Negatively Skewed Distribution

In a negatively skewed distribution, the value of mode is maximum and the value of mean is the least. The median lies in between the two. The long tail is spread over the left hand side of the distribution while the hump is on the right hand side.

Tests of Skewness :

In order to ascertain whether a distribution is skewed or not, the following tests may be applied.

Skewness is present if :

(i) The values of mean, median and mode do not coincide.

(ii) When the data are plotted on a graph they do not give the normal bell-shaped form i.e., when cut along the vertical line through the centre, the two halves do not match.

(iii) The sum of the positive deviations from the median is not equal to the sum of the negative deviations.

(iv) Quartiles are not equidistant from the median.

(v) Frequencies are not equally distributed at points of equal deviation from the mode.

3.3 Moments

Let the symbol 'd' be used to represent the deviation of any item in a distribution from the arithmatic mean of that distribution, ie., $d = \left(X - \bar{X}\right)$.

The arithmatic mean of these deviations with various powers is called the moment of the distribution. If we take the mean of the deviations with power one, we get the first moment about the mean, ie., $\frac{\Sigma d}{N}$ or, $\frac{\Sigma X - \bar{X}}{N}$ and it is denoted by μ_1.

Similrly, the mean of the squares of the deviations gives us the second moment about the mean, ie., $\frac{\Sigma d^2}{N}$ or, $\frac{\Sigma\left(X-\bar{X}\right)^2}{N}$ and it is denoted by μ_2.

The mean of the cubes of the deviations gives us the third moment about the mean, ie., $\frac{\Sigma d^3}{N}$ or, $\frac{\Sigma\left(X-\bar{X}\right)^3}{N}$ and it is denoted by μ_3 and so on.

The moments about mean is called the central moment. Since the sum of deviation of items from arithmatic mean is always zero,

$$\mu_1 = \frac{\Sigma\left(X-\bar{X}\right)}{N} = 0 \text{ and } \mu_2 = \frac{\Sigma\left(X-\bar{X}\right)^2}{N} = \sigma^2$$

In a symmetrical distribution, all odd moments ie. μ_1, μ_3 etc. would always be zero. In a symmetrical distribution, the deviation below the mean and the deviation above the mean will be exactly the same and will be of opposite sign, and therefore they will cancel each other when summed up. But with deviations with even power will have positive sign and will not cancel out. Thus odd moments will always be equal to zero for symmetrical distribution. But this will not hold true for asymmetrical distributions.

3.3.1 Moments about arbitrary origin

While we select an arbitrary origin A, then formula of moments for simple series

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

μ'_1 = First moment about arbitrary origin $= \frac{\Sigma X - A}{n} = \bar{\vec{X}} - A$

μ'_2 = Second moment about arbitrary origin $= \frac{\Sigma\left(X-A\right)^2}{n}$

μ'_3 = Third moment about arbitrary origin $= \frac{\Sigma\left(X-A\right)^3}{n}$

μ'_4 = Fourth moment about arbitrary origin $= \frac{\Sigma\left(X-A\right)^4}{n}$

In general, the r^{th} moment for grouped or frequency data,

$$\mu'_r = \frac{1}{N}\sum f_i \left(X_i - A\right)^r \quad \text{where } N = \sum f_i$$

Putting r = 0, we get $\mu'_0 = \frac{1}{N}\sum f_i = 1$

If A = 0, we call it raw moments

So, r-th order raw moment of x, $\mu'_r = \frac{1}{N}\sum x_i^r$. For frequency data, $\mu'_r = \frac{1}{N}\sum f_i x_i^r$

3.3.2 Central moments expressed in terms of raw moments

For the sake of simplicity in calculations, moments are first calculated about an arbitrary origin. If we want to obtain moments about mean, called central moments, we can do so with the help of the following relationship :

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

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

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

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\left(\mu'_1\right)^2\mu'_2 - 3\left(\mu_1\right)^4$$

Derivation : The r-th central moment about the mean is

$$\mu_r = \frac{1}{N}\sum f_i \left(X_i - \bar{X}\right)^r$$

$$= \frac{1}{N}\sum f_i \left[\left(X_i - A\right) - \left(\bar{X} - A\right)\right]^r$$

$$= \frac{1}{N}\sum f_i \left[\left(X_i - A\right) - d\right]^r$$

where, $\bar{X} - A = \mu'_1 = d$

$$= \frac{1}{N}\sum f_i \Big[\left(X_i - A\right)^r - {}^rC_1\left(X_i - A\right)^{r-1} d + {}^rC_2\left(X_i - A\right)^{r-2} d_2$$

$$+ \ldots\ldots\ldots + \left(-1\right)^r {}^rC_r d^r\Big]$$

$$= \frac{1}{N}\sum f_i\left(X_i - A\right)^r - {}^rC_1 d\frac{\Sigma f_i\left(X_i - A\right)^{r-1}}{N} + {}^rC_2 d^2\frac{\Sigma f_i\left(X_i - A\right)^{r-2}}{N} + ...$$

$$= \mu'_r - {}^rC_1 d\mu'_{r-1} + {}^rC_2 d^2\mu'_{r-2} + ...$$

Putting $d = \mu'_1$

$$\mu_r = \mu'_r - {}^rC_1\mu'_{r-1}\times\mu'_1 + ...$$

Now putting r = 1, 2, 3, 4.....

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - 2\mu'_1\mu'_1 + \left(\mu'_1\right)^2 = \mu'_2 - \left(\mu'_1\right)^2$$

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

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

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

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

and so on.

Moments about any point as expressed in terms of moments about the mean

$$\mu'_r = \frac{1}{N}\sum f_i\left(X_i - A\right)^r$$

$$= \frac{1}{N}\sum f_i\left[\left(X_i - \bar{X}\right) - \left(\bar{X} - A\right)\right]^r$$

$$= \frac{1}{N}\sum f_i\left[\left(X_i - \bar{X}\right) - d\right]^r$$

$$= \frac{1}{N}\sum f_i\left[\left(X_i - \bar{X}\right)^r + {}^rc_1\left(X_i - \bar{X}\right)^{r-1}.d + {}^rc_2\left(X_i - \bar{X}\right)^{r-2}.d^2 ... + {}^rc_r d^r\right]$$

$$= \frac{1}{N}\sum f_i\left(X_i - \bar{X}\right)^r + rd\frac{\Sigma f_i\left(X_i - \bar{X}\right)^{r-1}}{N} + rd^2\frac{\Sigma f_i\left(X_i - \bar{X}\right)^{r-2}}{N} + ...$$

$$\mu'_r = \mu_r + {}^rC_1 d\mu_{r-1} + {}^rC_2 d^2\mu_{r-2} + ... + d^r$$

where, $d = \bar{X} - A = \mu'_1$, $\mu_0 = 1$ and $\mu_1 = 0$

Now putting r = 1, 2, 3, 4.....

$\mu_1' = \mu_1 + d = d$ where, $d = \bar{X} - A$, $\mu_1 = 0$

$\mu_2' = \mu_2 + 2\mu_1 d + \mu_0 d^2 = \mu_2 + d^2$

$\mu_3' = \mu_3 + 3\mu_2 d + d^3$

$\mu_4' = \mu_4 + 4\mu_3 d + 6\mu_2 d^2 + d^4$ and so on.

Central moments are dependent on the change of scale but independent of any change in origin

Effect of a change in origin :

Let the origin be changed to point A. If X is the new variate, then

$$x = A + X \quad .....(1)$$

$$\bar{x} = A + \bar{X} \quad .....(2)$$

From (1)–(2) we get,

$$x - \bar{x} = X - \bar{X} \quad .....(3)$$

r-th moment about the mean of the variate x

$$= \frac{1}{N}\sum f_i \left(x_i - \bar{x}\right)^r$$

$$= \frac{1}{N}\sum f_i \left(X_i - \bar{X}\right)^r \text{ [From (3) we know } \left(x - \bar{x}\right) = \left(X - \bar{X}\right)]$$

= r-th moment about the mean of the variate X.

$\therefore \ \mu_r(x) = \mu_r(X)$

Effect of change in origin and change of scale

Let U is the new variate such that $U = \frac{x - a}{h}$

$\therefore \quad x = a + hU \quad .....(1)'$

$\bar{x} = a + h\bar{U} \quad .....(2)'$

From (2)′–(1)′ we get

$$\left(x - \bar{x}\right) = h\left(U - \bar{U}\right)$$

or, $\left(x_i - \bar{x}\right) = h\left(U_i - \bar{U}\right)$

r-th moment about the mean of the variate x

$$= \frac{1}{N}\Sigma f_i \left(x_i - \bar{x}\right)^r$$

$$= \frac{1}{N}\Sigma f_i\left[h^r\left(U_i - \bar{U}\right)^r\right]$$

$$= h^r \cdot \frac{1}{N}\Sigma f_i\left(U_i - \bar{U}\right)^r$$

$= h^r$ [r-th moment about mean of the variate U]

$$\therefore \quad \mu_r(x) = h^r \mu_r(U)$$

Hence the calculation of moments is not affected by the change of origin but are changed by the change of scale.

3.3.3 Sheppard's Correction and Charlier's check

While calculating moments it is assumed that all the values of a variable in a class interval are concentrated at the centre of that interval (i.e., mid-point). However, in practice, it is not so—the assumption is an approximation to facilitate calculations and it introduces some error which is known as grouping error. This grouping error can be eliminated by sheppard's correction.

$\mu_1 = 0$

$$\mu_2 \text{ (corrected)} = \mu_2 \text{ (uncorrected)} - \frac{i^2}{12}$$

μ_3 (corrected) = Uncorrected

$$\mu_4 \text{ (corrected)} = \mu_4 \text{ (uncorrected)} - \frac{1}{2}i^2\mu_2 \text{ (uncorrected)} + \frac{7}{240} \times i^4$$

where, ''i'' is the width of the class interval.

The first moment (μ_1) and the third moment (μ_3) need no correction.

Conditions for applying sheppard's corrections :

The following conditions should be satisfied for the application of Sheppard's correction :

(i) The data should relate to a continuous variable.

(ii) The total frequency should be fairly large.

(iii) The number of classes shlould not be two large.

(iv) The frequencies should taper off to zero in both directions, i.e., the curve should approach the base line gradually and slowly at each end of the distribution.

Generally, sheppard's corrections are not applied unless the total frequency is more than 1000 and the number of classes is less than 20.

Charlier's Check :

The coding method given in previous chapters for computing the mean and standard deviation can also be used to provide a short method for computing moments. This method uses the fact that

$$X_j = A + i\,d_j \quad \text{or briefly } X = A + id$$

$$\text{or,} \qquad d = \frac{X - A}{i}$$

The raw moment can be found out using the formula

$$\mu'_r = i^r \frac{\Sigma f d^r}{N} = i^r d^r$$

This can be used to find the central moments from the formula to convert raw moments to central moments.

Charlier's check in computing moments by the coding method uses the following identities :

$$\Sigma f\,(d+1) = \Sigma fd + \Sigma f$$

$$\Sigma f\,(d+1)^2 = \Sigma fd^2 + 2\Sigma fd + \Sigma f$$

$$\Sigma f\,(d+1)^3 = \Sigma fd^3 + 3\Sigma fd^2 + 3\Sigma fd + \Sigma f$$

$$\Sigma f\,(d+1)^4 = \Sigma fd^4 + 4\Sigma fd^3 + 6\Sigma fd^2 + 4\Sigma fd + \Sigma f$$

3.3.4 Moments in Dimensionless Form

Dimensionless means the measure is a pure number without any unit.

To avoid particular units, we can define the dimensionless moments about the mean as

$$a_r = \frac{\mu_r}{S^r} = \frac{\mu_r}{\left(\sqrt{\mu_2}\right)^r} = \frac{\mu_r}{\sqrt{\mu_2^r}}$$

where, $S = \sqrt{\mu_2}$ is the standard deviation.

Since $\mu_1 = 0$ and $\mu_2 = S^2$, we have $a_1 = 0$ and $a_2 = 1$

Moment coefficient of Skewness (dimensionless) :

$$a_3 = \frac{\mu_3}{S^3} = \frac{\mu_3}{\left(\sqrt{\mu_2}\right)^3} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

For normal curve $a_3 = 0$

Moment co-efficient of kurtosis (dimensionless) :

$$a_4 = \frac{\mu_4}{S^4} = \frac{\mu_4}{\mu_2^2}$$

For normal distribution, $a_4 = 3$

a_4 is often denoted by β_2.

$\beta_2 = 3 \Rightarrow$ The curve is Normal or Mesokurtic

$\beta_2 > 3 \Rightarrow$ More Peaked or Leptokurtic

$\beta_2 < 3 \Rightarrow$ Flat topped or Platykurtic

3.3.5 Computation of moments for grouped data

We summarise below the important results on moments :

1. $\mu_0 = 1$ and $\mu_1 = 0$
2. Mean $= \bar{X} = A + \mu_1'$

   In particular, if we take A = 0 we get

   $\bar{X} = 0 + \mu_1'$ (about origin)

   Hence, the first moment about origin gives mean.
3. Variance $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2$

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

   $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$
4. The following formula give the moments about any arbitrary point ‘A’ in terms of moments abut mean.

   Mean $= A + \mu_1'$

   $\mu_2' = \mu_2 + \mu_1'^2$

   $\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3$

   $\mu_4' = \mu_4 + 4\mu_3\mu_1' - 6\mu_2\mu_1'^2 + \mu_1'^4$
5. Beta (β) and Gamma (γ) coefficients based on Moments.

   $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \; ; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \qquad \left(\because\ \mu_2 = \sigma^2\right)$$

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

Example : Find the first, second, third and fourth central moments of the set of numbers 2, 4, 6, 8.

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

Mean $(\bar{x}) = \frac{\Sigma x}{N} = \frac{20}{4} = 5 \quad (\because N = 4)$

The first four central moments are given by

$$\mu_1 = \frac{1}{N}\sum(x - \bar{x}) = 0$$

$$\mu_2 = \frac{1}{N}\sum(x - \bar{x})^2 = \frac{20}{4} = 5$$

$$\mu_3 = \frac{1}{N}\sum(x - \bar{x})^3 = 0$$

$$\mu_4 = \frac{1}{N}\sum(x - \bar{x})^4 = \frac{164}{4} = 41$$

Example : Calculate β_1 and β_2 (measures of skewness and kurtosis) for the following frequency distribution :

x :	2	3	4	5	6
f :	1	3	7	2	1

x	f	d = x – 4	fd	fd^2	fd^3	fd^4
2	1	–2	–2	4	–8	16
3	3	–1	–3	3	–3	3
4	7	0	0	0	0	0
5	2	1	2	2	2	2
6	1	2	2	4	8	16
	$\Sigma f = N = 14$		$\Sigma fd = -1$	$\Sigma fd^2 = 13$	$\Sigma fd^3 = -1$	$\Sigma fd^4 = 37$

The moments about the point A = 4 are given by

$$\mu_1' = \frac{\Sigma fd}{N} = \frac{-1}{14} = -0.0714$$

$$\mu_2' = \frac{\Sigma fd^2}{N} = \frac{13}{14} = 0.9286$$

$$\mu_3' = \frac{\Sigma fd^3}{N} = \frac{-1}{14} = -0.0714$$

$$\mu_4' = \frac{\Sigma fd^4}{N} = \frac{37}{14} = 2.6429$$

The central moments ie., the moments about mean are given by :

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 0.9286 - (-0.0714)^2 = 0.9286 - 0.0051 = 0.9235$$

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

$$= -0.0714 - 3 \times 0.9286 \times (-0.0714) + 2 \times (-0.0714)^3$$

$$= -0.0714 + 0.1989 - 0.0007$$

$$= 0.1268$$

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

$$= 2.6429 - 4 \times (-0.0714) \times (-0.0714) + 6 \times (0.9286) \times (-0.0714)^2 - 3 \times (-0.0714)^4$$

$$= 2.6429 - 0.0204 + 0.0284 - 0.0001 = 2.6508$$

$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.1268)^2}{(0.9235)^3} = \frac{0.0161}{0.7876} = 0.0204$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2.6508}{(0.9235)^2} = \frac{2.6508}{0.8529} = 3.1080$$

Since, $\beta_1 \simeq 0$ and $\beta_2 \simeq 3$, the given distribution is approximately a Normal distribution.

Example : Calculate the values of β_1 and β_2 from the following data.

Marks :	20–30	30–40	40–50	50–60	60–70	70–80	80–90
No. of students :	4	7	10	20	4	3	2

Also apply sheppard's corrections for moments

Hint : X : Mid-Value

f : No. of students

$N = \Sigma f = 50$; $d = \frac{X - A}{i} = \frac{X - 55}{10}$. we shall get :

$\Sigma fd = -20$, $\Sigma fd^2 = 108$, $\Sigma fd^3 = -92$, $\Sigma fd^4 = 660$

The moments about the point A = 55 are given by

$$\mu_1' = i \cdot \frac{\Sigma fd}{N} = 10 \times \frac{-20}{50} = -4$$

$$\mu_2' = i^2 \cdot \frac{\Sigma fd^2}{N} = 10^2 \times \frac{108}{50} = 216$$

$$\mu_3' = i^3 \cdot \frac{\Sigma fd^3}{N} = 10^3 \times \frac{-92}{50} = -1840$$

$$\mu_4' = i^4 \cdot \frac{\Sigma fd^4}{N} = 10^4 \times \frac{660}{50} = 132000$$

The central moments are :

$\mu_1 = 0$

$$\mu_2 = \mu_2' - \mu_1'^2 = 216 - (-4)^2 = 200$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$$

$$= -1840 - 3\times 216\times(-4) + 2\times(-4)^3 = 624$$

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

$$= 132000 - 4\times(-1840)\times(-4) + 6\times 216\times(-4)^2 - 3\times(-4)^4$$

$$= 122528$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(624)^2}{(200)^3} = 0.0487$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{122528}{(200)^2} = 3.0632$$

Since, $\beta_1 \simeq 0$ and $\beta_2 \simeq 3$, the given distribution is nearly Normal.

Sheppard's Correction for Moments

$$\mu_2(\text{corrected}) = \mu_2 - \frac{i^2}{12} = 191.67 \qquad (i = 10)$$

$$\mu_3 \text{ (corrected)} = \mu_3$$

$$\mu_4(\text{corrected}) = \mu_4 - \frac{1}{2}i^2\mu_2 + \frac{7}{240}.i^4$$

$$= 122528 - \frac{1}{2}\times 100\times 200 + \frac{7}{240}\times(10)^4$$

$$= 112819.67$$

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

Ans. Given, A = 5, $\mu'_1 = 2$ and $\mu'_2 = 20$

We know that Mean $= A + \mu'_1 = 5 + 2 = 7$

and variance $= \mu_2 = \mu'_2 - \mu'^2_1 = 20 - 4 = 16$

3.4 Measures of Skewness

Measures of Skewness tell us the direction and extent of asymmetry in a series, and permit us to compare two or more series with regard to these. They may either be absolute or relative.

3.4.1 Absolute measures of Skewness

Skewness can be measured in absolute terms by taking the difference between mean and mode.

$$\text{Absolute Sk} = \bar{X} - \text{Mode}$$

When Skewness is based on quartiles, absolute Skewness is given by the formula :

$$\text{Absolute Sk} = Q_3 + Q_1 - 2 \text{ Median}$$

If the value of mean is greater than mode, Skewness will be positive i.e., we shall get a plus sign in the above formula. Conversely, if the value of mode is greater than the mean, we shall get a minus sign meaning thereby that the distribution is negatively skewed.

The reason why the difference between mean and mode can be used to measure Skewness is that in a symmetrical distribution the values of mean, median and mode are alike, but the mean moves away from the mode when the observations are asymmetrical. Consequently, the distance between the mean and the mode could be used to measure Skewness–the greater is this distance, whether positive or negative, the more asymmetrical the distribution. However, such a measure is unsatisfactory on two grounds :

1. This absolute measure of Skewness is expressed in the unit of value of the distribution and therefore cannot be compared with another such measure of a series expressed in different units.
2. Distributions vary greatly and the difference between Mean and the Mode in absolute terms might be considerable in one series and small in another, although the frequency curves of the two distributions were similarly Skewed.

3.4.2 Relative Measures of Skewness

The relative measures of Skewness are called the co-efficient of Skewness. They have the following properties :

1. It should be a pure number. The value should be independent of the units of the series and also of the degree of variation in the series.
2. It should have a zero value, when the distribution is symmetrical.
3. It should have some meaningful scale of measure so that we could easily interpret the measured value.

There are four important measures of relative skewness, viz.,

(i) Karl Pearson's co-efficient of skewness

(ii) Bowley's co-efficient of skewness

(iii) Kelly's co-efficient of skewness

(iv) Measure of Skewness based on Moments.

3.4.3 Karl Pearson's co-efficient of Skewness

It is based on the difference between mean and mode. This difference is divided by standard deviation to give a relative measure of skewness. The formula is

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\bar{X} - M_0}{\sigma}$$

Where Sk_p is the Karl Pearson's coefficient of Skewness. There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and usually lies between ± 1.

When a distribution is symmetrical, the values of mean, median and mode coincide and therefore the co-efficient of Skewness will be zero. When a distribution is positively Skewed, the co-efficient of Skewness shall have plus sign and when it is negatively Skewed, the coefficient of Skewness shall have minus sign. The degree of skewness shall be obtained by the numerical value, say 0.8 or 0.2 etc. Thus this formula gives both the direction as well as the extent of Skewness.

This method of measuring Skewness cannot be used where mode is ill-defined. However, in moderetly Skewed distribution, the averages have the following relationship:

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}$$

If this value of mode is substituted in the above formula we arrive at another formula for measuring Skewness

$$Sk_p = \frac{\left[\bar{X} - \left(3\text{Med.} - 2\bar{X}\right)\right]}{\sigma}$$

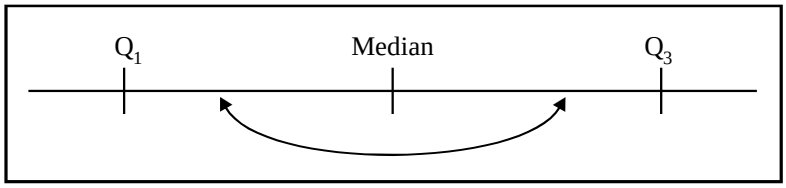
$$= \frac{\bar{X} - 3\text{Med.} + 2\bar{X}}{\sigma}$$

$$= \frac{3\left(\bar{X} - \text{Med.}\right)}{\sigma}$$

Theoretically, the value of this co-efficient varies between ± 3. However, in practice it is rare that the co-efficient of Skewness obtained by the above method exceeds ± 1.

3.4.4 Bowley's co-efficient of Skewness

Bowley's measure of Skewness is based on quartiles. In a symmetrical distribution first and third quartiles are equidistant from the median as can be seen from the following diagram.



In a symmetrical distribution, Q_3 and Q_1 both will maintain the same distrance from the median, ie.

$$Q_3 - \text{Med.} = \text{Med.} - Q_1$$

or, $$Q_3 + Q_1 - 2\ \text{Med.} = 0$$

If the distribution is positively skewed the top 25% of the values will tend to be farther away from the median than the botton 25% i.e., Q_3 will maintain greater distance from the median than Q_1. This position will be reversed when the distribution is negatively skewed.

So, Bowley's measure (co-efficient of skewness) is

$$Sk_B = \frac{(Q_3 - \text{Med.}) - (\text{Med.} - Q_1)}{(Q_3 - \text{Med.}) + (\text{Med.} - Q_1)}$$

or, $$Sk_B = \frac{Q_3 + Q_1 - 2\text{Med.}}{Q_3 - Q_1}$$

However, the results obtained from Karl Peasson's measure and Bowley's measure of skewness cannot be compared with one another. Especially, the numerical values are not related to one another, since Bowley's measure, because of its computational basis, is limited to values between –1 and +1, while pearson's measure has no such limits.

In rare occasions, with unusually shaped distributions, it is possible for them to emerge with opposite signs.

3.4.5 Kelly's co-efficient of Skewness

Bowley's measure of Skewness neglects the two extreme quarters of the data. It would be better for a measure to cover the entire data especially because in measuring

Skewness, we are often interested in the more extreme items. Bowley's measure can be extended by taking any two deciles equidistant from the median or any two percentiles equidistant from the median. Kelly has suggested the following formula for measuring Skewness based on the 10th and the 90th percentiles or the first and ninth deciles.

$$Sk_K = \frac{P_{10} + P_{90} - 2Med.}{P_{90} - P_{10}}$$

Also, $$Sk_K = \frac{D_1 + D_9 - 2Med.}{D_9 - D_1}$$

SK_K is the Kelly's co-efficient of Skewness

This measure of skewness has one theoretical attraction if skewness is to be based on percentiles. However, this method is not very popular in practice and generally Karl Pearson's meausre is used.

3.4.6 Measure of Skewness based on the Third Moment

A measure of skewness may be obtained by using second and third moment about the mean.

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ is used as a relative measure of skewness. In a symmetrical distribution β_1 will be zero. The greater the value of β_1 the more skewed will be the distribution. However, the co-efficient β_1 as a measure of skewness has a serious limitation. μ_3^2 is always positive. Also μ_2 being the variance is always positive.

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ is always positive.

Thus, β_1 as a measure of skewness is not able to tell us about the direction (positive or negative) of skewness. This drawback is removed in Karl Pearson's co-efficient

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

The sign of skewness would depend on the value of μ_3. If μ_3 is positive we will have positive skewness, and if μ_3 is negative, we will have negative skewness.

Example : Calculate Pearson's co-efficient of Skewness.

x	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5
f	28	42	54	108	129	61	45	33

Ans. Calculation of co-efficient of Skewness

x	f	$d = \frac{X - 27.5}{5}$	fd	fd^2
12.5	28	–3	–84	252
17.5	42	–2	–84	168
22.5	54	–1	–54	54
27.5	108	0	0	0
32.5	129	+1	+129	129
37.5	61	+2	+122	244
42.5	45	+3	+135	405
47.5	33	+4	+132	528
	N = 500		$\Sigma fd = 296$	$\Sigma fd^2 = 1780$

Co-efficient of $SK_P = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{\bar{X} - M_0}{\sigma}$

Mean $\bar{X} = A + \frac{\Sigma fd}{N} \times i$

Where, A = 27.5, Σfd = 296, N = 500, i = 5

$$\bar{X} = 27 \cdot 5 + \frac{296}{500} \times 5 = 30.46$$

Mode : Since the maximum frequency is 129, the corresponding value of X, ie, 32.5 is the modal value.

$$\text{S.D. } : \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Σfd_2 = 1780, N = 500, Σfd = 296, i = 5

$$\sigma = \sqrt{\frac{1780}{500} - \left(\frac{296}{500}\right)^2} \times 5$$

$$= \sqrt{3.56 - 0.35} \times 5 = 8.96$$

$\therefore$ Co-efficient of $SK_P = \frac{X - M_0}{\sigma} = \frac{30.46 - 32.5}{8.96} = \frac{-2.04}{8.96} = -0.228$

Example : Calculate Bowley's co-efficient of Skewness from the following data :

Variable	Frequency	Cumulative Frequency (cf)
0–10	12	12
10–20	16	28
20–30	26	54
30–40	38	92
40–50	22	114
50–60	15	129
60–70	7	136
70–80	4	140

Bowley's Co-efficient of Skewness is

$$SK_B = \frac{Q_3 + Q_1 - 2\text{Med.}}{Q_3 - Q_1}$$

Q_1 = size of $\frac{N}{4}$ th item $= \frac{140}{4} = 35$ th item. It lies in the class (20–30).

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

L = 20, $N/4$ = 35; CF = 28; f = 26; i = 10

$$Q_1 = 20 + \frac{35 - 28}{26} \times 10 = 20 + 2.69 = 22.69$$

Q_3 = Size of $3N/4$ th item $= \frac{3 \times 140}{4} = 105$ th item. It lies in the class (40–50)

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

L = 40; $3N/4 = 105$; c.f. = 92; f = 22; i = 10

$$Q_3 = 40 + \frac{105 - 92}{22} \times 10 = 40 + 5.91 = 45.91$$

Med. = Size of $\frac{N}{2}$ th item $= \frac{140}{2} = 70$ th item. It lies in the class (30–40)

$$\text{Med.} = L + \frac{N/2 - \text{c.f.}}{f} \times i$$

L = 30; $N/2 = 70$; c.f. = 54; f = 38; i = 10

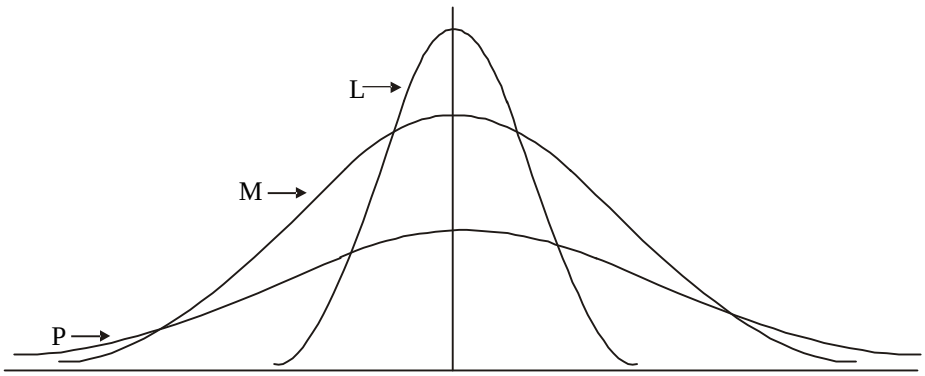
$$\text{Med.} = 30 + \frac{70 - 54}{38} \times 10 = 30 + 4.21 = 34.21$$

$\therefore$ Coefficient of Skewness $SK_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$

$$SK_B = \frac{45.91 + 22.69 - 2(34.21)}{45.91 - 22.69} = \frac{68.6 - 68.42}{23.22} = 0.008$$

3.5 Kurtosis

By Kurtosis of a frequency distribution we mean its degree of peakedness or steepness. Two distributions may be identical is respect of central tendency, dispersion and Skewness, but one may be more peaked than the other. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. In other words, measures of Kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called 'leptokurtic', In such a case items are more closely bunched around the mode. On the other hand, if a curve is more flat-topped than the normal curve, it is called 'platykurtic' The normal curve itself is known as 'mesokurtic'



L = Leptokurtic, M = Mesokurtic, P = Platykurtic

3.6 Measures of Kurtosis

As a measure of Kurtosis, Karl Pearson gave the co-efficient β_2 or its derivative γ_2.

β_2 is defined as :

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$ where μ_4 is the 4th central moment and μ_2 is the second central moment.

The greater the value of β_2, the more peaked is the distribution.

For a normal curve, the value of $\beta_2 = 3$ when the value of β_2 is greater than 3 the curve is more peaked than the normal curve and it is termed as ''Leptokurtic'' distribution. When the value of β_2 is less than 3 the curve is less peaked than the normal cuve and termed as ''platykurtic'' distribution. The normal curve with $\beta_2 = 3$ is called ''mesokurtic'' distribution.

γ_2 is also used to measure Kurtosis.

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

For a normal curve $\gamma_2 = 0$ (mesokurtic)

For a Leptokurtic distribution, $\gamma_2 > 0$

For a platykurtic distribution $\gamma_2 < 0$

3.7 Some theorems on Moments, Skewness and Kurtosis

Theorem 1 : Show that the measure of skewness given by $sk = \frac{3(\bar{x} - M_e)}{\sigma}$ must lie between –3 and +3.

Proof : We know that variance is non-negative

So, $\frac{1}{n}\Sigma ui^2 - \left(\frac{\Sigma ui}{n}\right)^2 \geq 0$ or, $\frac{1}{n}\Sigma ui^2 \geq \left(\frac{\Sigma ui}{n}\right)^2$

Putting $u_i = |x_i - \bar{x}|$ we get,

$$\frac{1}{n}\Sigma(x_i - \bar{x})^2 \geq \left[\frac{\Sigma|x_i - \bar{x}|}{n}\right]^2$$

Since MD about median is least,

$$\frac{1}{n}\Sigma\left(x_i-\bar{x}\right)^2 \geq \left[\frac{\Sigma\left|x_i-M_e\right|}{n}\right]^2 \quad .....(1)$$

Now, $\Sigma\left|x_i-M_e\right| = \left|x_1-M_e\right| + \left|x_2-M_e\right| + ... + \left|x_n-M_e\right|$

or, $\Sigma\left|x_i-M_e\right| \geq \left|x_1-M_e+x_2-M_e+.....+x_n-M_e\right|$

$$= \left|\Sigma x_i - nM_e\right|$$

$$\therefore \quad \frac{\Sigma\left|x_i-M_e\right|}{n} \geq \frac{1}{n}\left|\Sigma x_i - nM_e\right| = \left|\bar{x}-M_e\right|$$

$$\therefore \quad \frac{\Sigma\left|x_i-M_e\right|}{n} \geq \left|\bar{x}-M_e\right|$$

So, $$\left[\frac{\Sigma\left|x_i-M_e\right|}{n}\right]^2 \geq \left[\left|\bar{x}-M_e\right|\right]^2 \quad .....(2)$$

Comparing (1) and (2),

$$\frac{1}{n}\Sigma\left(x_i-\bar{x}\right)^2 \geq \left[\left|\bar{x}-M_e\right|\right]^2$$

or, $\sigma^2 \geq \left(\bar{x}-M_e\right)^2$

or, $$\left(\frac{\bar{x}-M_e}{\sigma}\right)^2 \leq 1$$

or, $$-1 \leq \frac{\bar{x}-M_e}{\sigma} \leq +1$$

or, $$-3 \leq \frac{3\left(\bar{x}-M_e\right)}{\sigma} \leq +3$$ **(Proved)**

Theorem 2 : Prove that $\beta_2 \geq 1$ (General proof)

We know that var (u) ≥ 0

or, $$\frac{1}{n}\Sigma u_i^2 - \left(\frac{\Sigma u_i}{n}\right)^2 \geq 0$$

or, $\frac{1}{n}\Sigma u_i^2 \geq \left(\frac{\Sigma u_i}{n}\right)^2$

Now, we put $u_i = \left(x_i - \bar{x}\right)^2$

So, $\frac{1}{n}\Sigma\left(x_i - \bar{x}\right)^4 \geq \left[\frac{1}{n}\Sigma\left(x_i - \bar{x}\right)^2\right]^2$

or, $m_4 \geq m_2^2$ or, $\frac{m_4}{m_2^2} \geq 1$

or, $\beta_2 \geq 1$ (Proved)

Theorem 3 : Prove that $\beta_2 = 1$ if the variable takes just two values with equal frequencies.

Proof : Let the variable x assume the value x_1 with frequency f and the value x_2 with the same frequency f. So, total number of observations = f + f = 2f = N and

$$\bar{x} = \frac{fx_1 + fx_2}{2f} = \frac{x_1 + x_2}{2}$$

Now, $m_4 = \frac{1}{N}\Sigma fi\left(x_i - \bar{x}\right)^4$

$$= \frac{1}{2f}\left[f\left(x_1 - \frac{x_1 + x_2}{2}\right)^4 + f\left(x_2 - \frac{x_1 + x_2}{2}\right)^4\right]$$

$$= \frac{1}{2f}\left[f\left(\frac{x_1 - x_2}{2}\right)^4 + f\left(\frac{x_1 - x_2}{2}\right)^4\right]$$

$$= \frac{1}{2f} \cdot 2f\left(\frac{x_1 - x_2}{2}\right)^4 = \left(\frac{x_1 - x_2}{2}\right)^4$$

Now, $m_2 = \frac{1}{N}\Sigma fi\left(x_i - \bar{x}\right)^2$

$$= \frac{1}{2f}\left[f\left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + f\left(x_2 - \frac{x_1 + x_2}{2}\right)^2\right]$$

$$= \frac{1}{2f}\left[f\left(\frac{x_1 - x_2}{2}\right)^2 + f\left(\frac{x_2 - x_1}{2}\right)^2 \right]$$

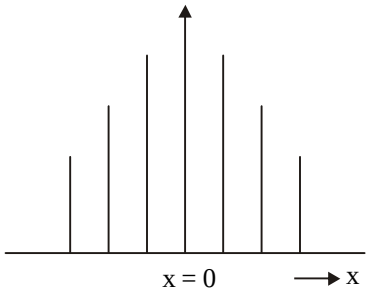
$$= \frac{1}{2f} \cdot 2f\left(\frac{x_1 - x_2}{2}\right)^2 = \left(\frac{x_1 - x_2}{2}\right)^2$$

$$\therefore \quad m_2^2 = \left(\frac{x_1 - x_2}{2}\right)^4 = m_4$$

$$\therefore \quad \beta_2 = \frac{m_4}{m_2^2} = \frac{m_4}{m_4} = 1$$ **[Proved]**

Theorem 4 : Show that all odd-order central moments are zero for symmetric distribution.

Proof : Let us take the origin at that point which corresponds to the point of symmetry of the frequency distribution.



Case (i) : When the number of observations is even.

Then the values of the observations are obviously $\pm x_i$. Let the corresponding frequencies be f_i.

then clearly, $\bar{x} = 0$

Now, $$m_{2r+1} = \frac{1}{N}\Sigma\left(x_i - \bar{x}\right)^{2r+1}.f_i$$

$$= \frac{1}{N}\left\{\Sigma {x_i}^{2r+1} \cdot f_i + \Sigma\left(-x_i\right)^{2r+1}.f_i\right\} \text{ as } \bar{x} = 0$$

Now, 2r+1 is odd. So, $\Sigma\left(-x_i\right)^{2r+1}.f_i = -\Sigma {x_i}^{2r+1}.f_i$

So, $m_{2r+1} = 0$ i.e., all odd-order central moments are zero for symmetric distribution.

Case (ii) : When the number of observations is odd.

We shall get one more term in addition to the values of the variable as in case (i).

That value is 0 with highest frequency f_0.

That is the value of x at the origin, the other observations being equally distributed, to the left and right hand sides of the origin.

Here again $\bar{x} = 0$

In this case, $m_{2r+1} = \frac{1}{N}\Sigma\left(x_i - \bar{x}\right)^{2r+1}.f_i$

$$= \frac{1}{N}\left\{\Sigma x_i^{2r+1}.f_i + \Sigma\left(-x_i\right)^{2r+1}.f_i + 0.f_0\right\} \quad \left(\text{as } \bar{x} = 0\right)$$

Now as 2r+1 is odd, $\Sigma\left(-x_i\right)^{2r+1}.f_i = -\Sigma x_i^{2r+1}.f_i$

So, $m_{2r+1} = 0$. Thus, in both cases, whether the number of observations is odd or even, all odd-order central moments are zero for symmetric distribution.

3.8 Summary

(i) Moments about mean

$$\mu_1 = \frac{\Sigma\left(X - \bar{X}\right)}{N} = 0 \qquad \mu_3 = \frac{\Sigma\left(X - \bar{X}\right)^3}{N}$$

$$\mu_2 = \frac{\Sigma\left(X - \bar{X}\right)}{N} \qquad \mu_4 = \frac{\Sigma\left(X - \bar{X}\right)^4}{N}$$

(ii) In a frequcny distribution

$$\mu_1 = \frac{\Sigma f\left(X - \bar{X}\right)}{N} \qquad \mu_2 = \frac{\Sigma f\left(X - \bar{X}\right)^2}{N}$$

(iii) Moments about arbitrary origin (A)

$$\mu_1' = \frac{\Sigma\left(X - A\right)}{N} \qquad \mu_3' = \frac{\Sigma\left(X - A\right)^3}{N}$$

$$\mu_2' = \frac{\Sigma\left(X - A\right)^2}{N} \qquad \mu_4' = \frac{\Sigma\left(X - A\right)^4}{N}$$

(iv) $\mu_1' = \frac{\Sigma f\left(X - A\right)}{N}$ or, $\mu_1' = \frac{\Sigma fd}{N} \times i$

$\mu_2' = \frac{\Sigma f\left(X - A\right)^2}{N}$ or, $\mu_2' = \frac{\Sigma fd^2}{N} \times i^2$ where $d = \frac{X - A}{i}$

i = class width

(v) Central moments expressed in terms of raw moments.

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

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

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

Skewness :

Karl Pearson's co-efficient of skewness

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$Sk_P = \frac{3\left(\text{Mean} - \text{Median}\right)}{\text{Standard Deviation}}$$

Bowley's Co-efficient of Skewness

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

Kelly's co-efficient of Skewness

$$Sk_K = \frac{P_{10} + P_{90} - \text{Med}}{P_{90} - P_{10}}$$

Measure of Skewness based on Moments

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ or, $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$

Kurtosis

Measure of Kurtosis based on moments

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

3.9 Questions

1. State whether the following statements are true or False.

   In case of false statement give the correct statement

   (i) Skewness studies the flatness or peakadness of the distribution.

   (ii) Kurtosis means lack of Symmetry.

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

   (iv) Skewness and Kurtosis help us in studying the shape of the frequency curve.

   (v) Bowley's co-efficient of Skewness lies between ± 3.

   (vi) Two distributions having the same values of mean, S.D., and Skewness must have the same kurtosis.

   (vii) A positively Skewed distributions curve is stretched more to the right than to the left.

   (viii) If $\beta_2 > 3$ the curve is called platykurtic

   (ix) If $\beta_2 = 3$ the curve is called normal.

   (x) β_1 is always non-negative.

   (xi) β_2 can be negative.,

   (xii) Variance = μ_2 (2nd moment about mean).

   (xiii) For a symmetrical distribution

   $\mu_1 = \mu_3 = \mu_5 = ... = 0$

   (xiv) For a symmetrical distribution

   Mean > Median > Mode

   (xv) $\beta_1 = \frac{\mu_4}{\mu_2^2}$

2. What is meant by moments of a fequency distribution?
3. Explain the use of moments in the measurement of Skewness and Kurtosis.
4. Define Skewness of a distribution.
5. What are the different measures of Skewness?
6. Write the formula expressing each of 2nd 3rd and 4th central moment in terms of raw moments.
7. What is sheppard's correction for moments.
8. Discuss the effect of change of origin and scale on third central moment.

9. Distinguish between positive Skewness and negative Skewness.
10. A distribution has standard deviation 3. What should be the value of 4th central moment so that the distribution is (i) mesokurtic, (ii) Platykurtic, (iii) Leptokurtic.

11. Calculate Karl Pearson's co-efficient of Skewness based on Mean and mode from the following.

Size	0–10	10–20	20–30	30–40	40–50	50–60	60–70
f	10	12	18	25	16	14	8

12. Find the first four moments about the mean for the following distribution :

Hight	60–62	63–65	66–68	69–71	72–74
f	5	18	42	27	8

13. For a distribution of 250 heights, calculations showed that the mean, S.D, Skewness and Kurtosis were 6, 5, 0 and 5 respectively. If was further found that two items 54 and 40 in the original data ware wronghly noted as 52 and 42 respectively. Calculate the correct values of mean, SD, skewness and kurtosis.

14. Find the Kurtosis for the following distribution

Class interval	0–10	10–20	20–30	30–40
Frequency	1	3	4	2

3.10 References

1. Das, N.G (1977) Statistical Methods, Part I & II, M. Das & Co.
2. Goon, Gupta, Dasgupta (1983) Fundamentals of Statistics, Vol. I, The World Press.

Unit 4 ❐ Correlation and Regression

Structure

4.1 Obejectives

4.2 Introduction

4.3 Correlation

4.3.1 Product Moment formula for the linear correlation co-efficient

4.3.2 Properties of correlation co-efficient

4.3.3 Computation of Correlation co-efficient from grouped data

4.3.4 Limitation of the correlation co-efficient

4.3.5 Rank Correlation

4.3.6 Interpretation of Rank Correlation co-efficient

4.4 Regression

4.4.1 Scatter Diagram and Regression Lines

4.4.2 Properties of Regression co-efficients

4.4.3 Standard Error of Estimate

4.4.4 Explained and Unexplained Variation

4.5 Summary

4.6 Questions

4.7 References

4.1 Objectives

If the data on two variables are recorded for a group of individuals, we have bivariate data. For example we may collect data indicating heights and weights of a group of students in a class or ages of husband and wife at the time of their marriage. In a bivariate distribution if the two variables vary in such a way that the changes in one are followed by changes in the other, then the variables are said to be correlated. The objective of the correlation analysis is to study the nature and extent of the association betwen the two variables. If the variables are found to be associated, we express one of the variables as a mathematical function of the other variable. The former is regarded as the dependent variable where as the latter is regarded as the independent variable. Now we may like to predict the value of the dependent variable corresponding to the given value of the independent variable. This objective can be achieved through regression analysis. Whether such estimation is robust or not can be studied at a still higher level of econometric analysis.

4.2 Introduction

Suppose, we have been given a data set contraining *n* pairs of values of the two variables *x* and *y*. If we plot each pair of given values in the first quadrant of the vector space, we get the diagrammatic representation of the bivariate data as a scatter diagram. From the scatter diagram one can easily identify the nature and intensity of the association of the two variables by inspecting the scatter diagram visually.

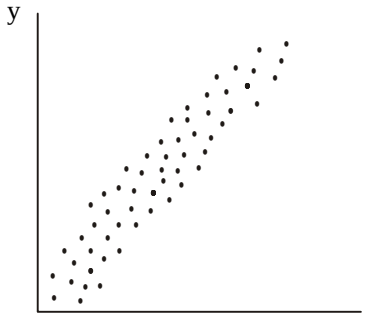


Fig (a)

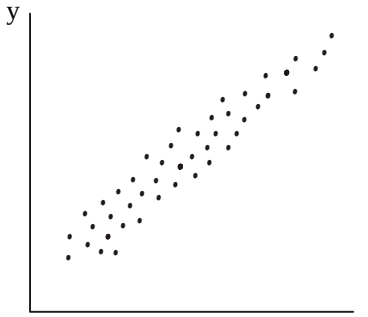


Fig (b)

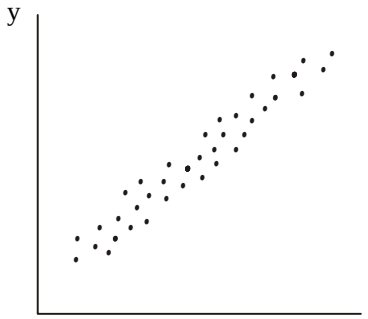


Fig (c)

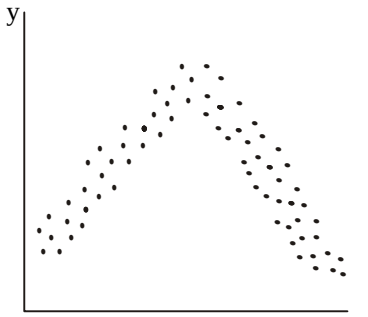


Fig (d)

Fig (a), Fig (b), Fig(c) and Fig (d) represent different types of data. Fig (a), Fig(b) and Fig(c) indicate that the association between x and y variables are linear whereas the data in Fig(d) shows a non-linear association of the variables.

Moreover, the intensity of the linear association in Fig(a), Fig(b) and Fig(c) are quite different. the intensity of assocation gradually increases as we move from Fig(a) to Fig(b) and to Fig(c).

For the linear association case, if we are allowed to draw a best fit line through the scatter diagram, it will be possible for us to estimate the value of the y-variable (dependent variable) given the value of the x-variable (independent variable) which is the basic working of the regression analysis.

4.3 Correlation

Correlation analysis helps us in determining the degree of relationship between two or more variables—it does not tell us anything about cause and effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables. However, the existence of causation always implies correlation.

Correlation can be classified in three different way :

(a) positive or negative correlation

(b) Simple, partial and multiple correlation

(c) linear and non-linear correlation

(a) Positive and negative correlation : Whether correlation is positive (direct) or negative (inverse) would depend on the direction of change of the variables. If both the variables are varying in the some direction ie, if one variable is increasing the other, on the average, is also incresing or, if one variable is decreasing, the other, on an average, is also decreasing, correlation is said to be positive. If on the other hand, the variables are varying in the opposite direction, i.e., of one variable is increasing, the other is decreasing or vice-versa, correlation is said to be negetive.

(b) Simple, Partial and multiple correlation : The distinction between simple, partial and multiple correlation is based on the number of variables studied. When only two variables are studied, it is a problem of simple correlation. When three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilisers used, it is a problem of multiple correlation. On the other hand, in partial correlation we recognise more than two variables, but consider only two variables to be influencing each other, the effect of other influencing variables being kept constant. For example, if we are interested in the association of the yield of rice per acre and the rainfall during periods when a certain average daily temperature existed.

(c) Linear and Non-Linear correlation : The distinction between linear and non-linear correlation is based on the constancy of the ratio of change between the variables. If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear.

Correlation will be non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. This has been discussed with diagrams in section 4.2.

4.3.1 Product moment formula for the linear correlation co-efficient

Suppose we are given a set of data containing pairs of values of x and y variables. Plotting the values of x and y we get a scatter diagram. (Fig. e). Now if we draw a new set of axes $x' = (x - \bar{x})$ and $y' = (y - \bar{y})$ through the point $(\bar{x}, \bar{y})$ which will

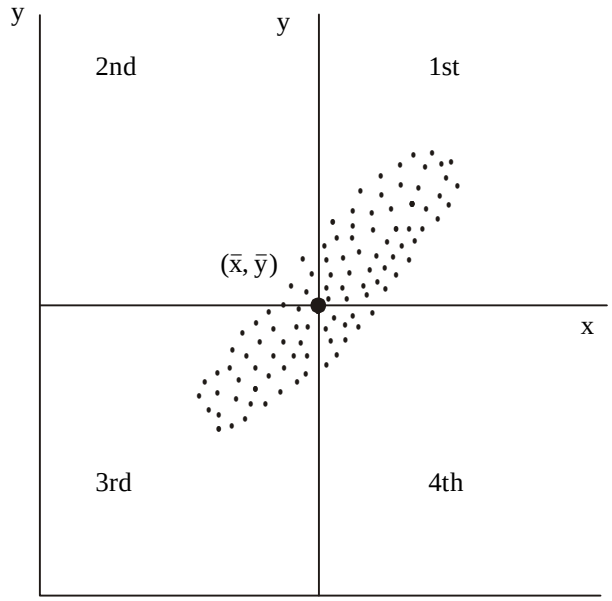


Fig (e)

be treated as the new origin. The (x', y') plane is divided into four quadrants. In the 1st and the 3rd quadrant, x' and y' will have the same sign, where as in the 2nd & 4th quadrant, x' and y' will have the opposite signs. If we calculate $\Sigma x'y'$, the sign of $\Sigma x'y'$ will be positive if most of the points lie in 1st and 3rd quadrant. On the other hand the sign of $\Sigma x'y'$ will be negative if most of the points lie in 2nd and 4th quadrant. In case of no correlation the points will be evenly distributed in the four quadrants and $\Sigma x'y'$ will be equal to zero.

So $\Sigma x'y' = \Sigma(x - \bar{x})(y - \bar{y})$ may be regarded as a measure of simple correlation. However, the value of $\Sigma x'y'$ depends on the number of pairs of values of the variables. So in order to neutralise this influence, we divide $\Sigma x'y'$ by n. Again $\Sigma x'y'$ is also influenced by units of measurement of the varibales and their variability. In order to neutralise this influence we divide $\Sigma x'y'$ by the standard deviations σ_x and σ_y. Thus we arrive of the product-moment formula for correlation as proposed by Karl Pearson :

The correlation co-efficient $= r_{xy} = \frac{1}{n}\frac{\sum(x_i - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}$

4.3.2 Properties of Correlation Co-efficient

1. Correlation co-efficient of any two variables is a pure number. It is independent of the units of measurement.
2 Correlation co-efficient r_{xy} is symmetric in x and y i.e, $r_{xy} = r_{yx}$.
3. **The value of the correlation coefficient is independent of the change of origin and scales of the variables.**

Proof : Suppose we are given n pairs of values (x_i, y_i) i = 1, 2,.....n of the variables x and y. we introduce two new variables u and v defined as follows :

$$u = \frac{x - a}{c} \text{ and } v = \frac{y - b}{d}$$

where, a, b, c and d are arbitrary constants, and $c \neq 0,\ d \neq 0$

So, corresponding to each pair (x_i, y_i), we have a pair of values (u_i, v_i) for the new variables where,

$$u_i = \frac{x_i - a}{c} \text{ and } v_i = \frac{y_i - b}{d}$$

or, $x_i = a + cu_i$ .....(1)

and $y_i = b + dv_i$ .....(2)

From (1) we get, $\bar{x} = a + c\bar{u}$ .....(3)

and (2) we get, $\bar{y} = b + d\bar{v}$ .....(4)

From (1)–(3) we get $(X - \bar{X}) = c(u_i - \bar{u})$

Similarly (2)–(4) we get $(y_i - \bar{y}) = d(v_i - \bar{v})$

$$\text{Var}(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{c^2}{n}\sum_i (u_i - \bar{u})^2 = c^2 \text{Var}(u)$$

$$\text{Var}(y) = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{d^2}{n}\sum_i (v_i - \bar{v})^2 = d^2 \text{Var}(v)$$

and $\text{Cov}(x, y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{cd}{n}\sum (u_i - \bar{u})(v_i - \bar{v})$

$$= cd\ cov\ (u, v)$$

Hence, $$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

$$= \frac{cd\ cov(u,v)}{\sqrt{c^2 var(u)}\sqrt{d^2 var(v)}}$$

$$= \frac{cd\ Cov(u,v)}{|c|\cdot|d|\sqrt{var(u)}\sqrt{var(v)}}$$

$$= \frac{cd}{|c|\cdot|d|} r_{uv}$$

$$= \begin{cases} +r_{uv}, & \text{when c and d are of the same sign.} \\ -r_{uv}, & \text{where c and d are of the opposite sign.} \end{cases}$$

So the corelation ocefficient is independent of the change of origin and scale.

4. The value of corrleation coefficient runs from –1 to + 1 ie, $-1 \le r \le +1$**.**

Proof. : Suppose we are given n pairs of values (x_i, y_i), i = 1, 2,.....n of the variables x and y.

$$\therefore \quad r = \frac{1}{n}\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = \frac{1}{n}\sum_i \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

$$= \frac{1}{n}\sum_i p_i q_i \quad \text{where, } p_i = \frac{x_i - \bar{x}}{\sigma_x} \text{ and } q_i = \frac{y_i - \bar{y}}{\sigma_y}$$

or, $$\sum_i p_i q_i = nr$$

$$\sum_i p_i^2 = \sum_i \left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 = \frac{1}{\sigma_x^2}\sum_i (x_i - \bar{x})^2 = \frac{n\sigma_x^2}{\sigma_x^2} = n$$

Similarly, $$\sum_i q_i^2 = n$$

Now, we know that

$$\sum_i (p_i + q_i)^2 \ge 0$$ Since squares of real quantities are non-negative

or, $\sum_i p_i^2 + \sum_i q_i^2 + 2\sum_i p_i q_i \geq 0$

or, $n + n + 2nr \geq 0$

or, $2n(1+r) \geq 0$

or, $1 + r \geq 0$

$\therefore \quad r \geq -1$ .....(i)

Again, $\sum_i (p_i - q_i)^2 \geq 0$

or, $\sum_i p_i^2 + \sum_i q_i^2 - 2\sum_i p_i q_i \geq 0$

or, $n + n - 2nr \geq 0$

or, $2n(1-r) \geq 0$

or, $1 - r \geq 0$

$\therefore \quad r \leq 1$ .....(ii)

Combining (i) and (ii) we get

$$-1 \leq r \leq +1$$

4.3.3 Computation of Correlation Co-efficint from grouped data

Suppose we have n pairs of values of x and y, then the Karl Pearson's measure of correlation coefficient will be as follows for bivariate data

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

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

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2} \quad \text{and} \quad \sigma_y = \sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}$$

$$\therefore \quad r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$= \frac{\Sigma xy - \bar{x}\Sigma y - \bar{y}\Sigma x + \Sigma \bar{x}\bar{y}}{\sqrt{\Sigma x^2 - 2\Sigma x\bar{x} + \Sigma(\bar{x})^2}\sqrt{\Sigma y^2 - 2\Sigma y\bar{y} + \Sigma(\bar{y})^2}}$$

$$= \frac{\Sigma xy - \bar{x}\Sigma y - \bar{y}\Sigma x + n\bar{x}\,\bar{y}}{\sqrt{\Sigma x^2 - 2\bar{x}\Sigma x + n(\bar{x})^2}\sqrt{\Sigma y^2 - 2\bar{y}\Sigma y + x(\bar{y})^2}}$$

Since $\Sigma x = n\bar{x}$ and $\Sigma y = n\bar{y}$

$$= \frac{\Sigma xy - n\bar{x}\bar{y} - n\bar{x}\,\bar{y} + n\bar{x}\,\bar{y}}{\sqrt{\Sigma x^2 - n(\bar{x})^2}\sqrt{\Sigma y^2 - n(\bar{y})^2}}$$

$$= \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\sqrt{\Sigma x^2 - n(\bar{x})^2}\sqrt{\Sigma y^2 - n(\bar{y})^2}}$$

Or, the formula can be designed as follows

$$r_{xy} = \frac{\Sigma xy - n \cdot \frac{\Sigma x \Sigma y}{n\,n}}{\sqrt{\Sigma x^2 - n\left(\frac{\Sigma x}{n}\right)^2}\sqrt{\Sigma y^2 - n \cdot \left(\frac{\Sigma y}{n}\right)^2}}$$

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

$$= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

For grouped data the same formula will be

$$r_{xy} = \frac{n\Sigma fxy - \Sigma fx\,\Sigma fy}{\sqrt{n\Sigma fx^2 - (\Sigma fx)^2}\sqrt{n\Sigma fy^2 - (\Sigma fy)^2}}$$

Since the correlation co-efficient is not affecfted by change in origin and change in scale, we transform the variables x an y to two new variables u and v.

$$u = \frac{x - A}{h} \text{ and } v = \frac{y - B}{k}$$

where h and k are the width of the x-classes and y-classes respectively and A and B are constants.

$$\therefore \quad r_{xy} = r_{uv} = \frac{n\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{n\Sigma fu^2 - (\Sigma fu)^2} \cdot \sqrt{n\Sigma fv^2 - (\Sigma fv)^2}}$$

Example : Calculate the product moment co-efficient of corrclation for the following bivariate distribution.

y \ x	5	10	15	20
11	2	4	5	4
17	5	3	6	2
23	3	1	2	3

Bivariate Correlation Table

y	x / u / v	5 / –2	10 / –1	15 / 0	20 / 1	f	fv	fv^2	fuv
11	–1	2 (4)	4 (4)	5 (0)	4 (–4)	15	–15	15	4
17	0	5 (0)	3 (0)	6 (0)	2 (0)	16	0	0	0
23	–1	3 (–6)	1 (–1)	2 (0)	3 (3)	9	9	9	–4
	f	10	8	13	9	$\Sigma f = 40$	$\Sigma fv = 6$	$\Sigma fv^2 = 24$	$\Sigma fuv = 0$
	fu	–20	–8	0	9	$\Sigma fu = -19$			
	fu^2	40	8	0	9	$\Sigma fu^2 = 57$			
	fuv	–2	3	0	–1	$\Sigma fuv = 0$			

Figures in brackets ( ) denote the product fuv for each cell. In the correlation table, x and y are mid-points of the respective classes and $u = \frac{x-15}{5}$ and $v = \frac{y-17}{6}$

$$r_{xy} = r_{uv} = \frac{n\Sigma uv - (\Sigma fu)(\Sigma fv)}{\sqrt{n\Sigma fu^2 - (\Sigma fu)^2} \cdot \sqrt{n\Sigma fv^2 - (\Sigma fv)^2}}$$

$$= \frac{40\times 0-(-6)\times(-19)}{\sqrt{40\times 24-(-6)^2}\cdot\sqrt{40\times 57-(-19)^2}} = \frac{-114}{\sqrt{924}\sqrt{1919}}$$

$$= \frac{-114}{30.4\times 43.81} = \frac{-114}{1331.82} = -0.0856 \simeq -0.09$$

Example : Let r be the correlation between x and y. What is the correlation between 3x + 1 and 2y – 3?

where

Suppose $u = 3x + 1$ or, $\bar{u} = 3\bar{x} + 1$

and $v = 2y - 3$ or, $\bar{v} = 2\bar{y} - 3$

$\therefore \quad u - \bar{u} = 3(x - \bar{x}) \qquad v - \bar{v} = 2(y - \bar{y})$

$$\therefore \quad \mathrm{Cov}(u,v) = \frac{1}{n}\sum(u-\bar{u})(v-\bar{v}) = \frac{1}{n}\sum\left[3(x-\bar{x})\cdot 2(y-\bar{y})\right]$$

$$= 6\times\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = 6\,\mathrm{cov}(x,\,y)$$

$$\sigma_u^2 = \frac{1}{n}\sum(u-\bar{u})^2 = \frac{1}{n}\sum\left[3(x-\bar{x})\right]^2 = 9\cdot\frac{1}{n}\sum(x-\bar{x})^2 = 9\sigma_x^2$$

$$\sigma_v^2 = \frac{1}{n}\sum(v-\bar{v})^2 = \frac{1}{n}\sum\left[2(y-\bar{y})\right]^2 = 4.\frac{1}{n}\sum(y-\bar{y})^2 = 4\sigma_y^2$$

We know, $r_{uv} = \dfrac{\mathrm{cov}(u,v)}{\sigma_u\sigma_v} = \dfrac{6\,\mathrm{cov}(x,y)}{\sqrt{9\sigma_x^2.4\sigma_y^2}}$

$$= \frac{6\,\mathrm{cov}(x,y)}{3\sigma_x\cdot 2\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} = r_{xy} = r$$

4.3.4 Limitation of the correlation co-efficient

If the relationship between x and y is non-linear, the correlation co-efficient fails to measure the intensity of association as they are not linearly related. So before the calculation of correlation co-efficient it is advisable to check whether the relationship is linear or not by drawing a scatter diagram.

Again, correlation coefficient may give misleading result if the data come from different sources. the variables may appear to be uncorrelated when the data from different sources are treated separately (see figure f).

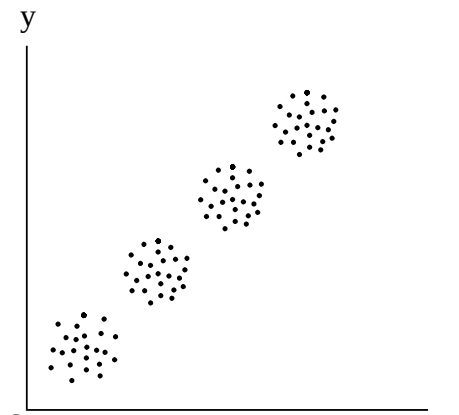


Fig.- f

More over, high correlation between two variables does not necessarily mean that the variables are causally related.

We may find high correctation only because both the variables depend on a third variable. This is the problem particularly with the time-series data. The size of the shoe and the IQ may be found to be highly correlated for a set of school children, but actually it is a non-sense correlation.

4.3.5 Rank Correlation

The product-moment correlation co-efficient (r) between the two variables is calculated by using the 'values' of the variables. But in many situations the measured values of the variables are not available. For example, the marks in two subjects say mathematics and Economics for a group of 10 students are not available, but their ranking in the two subjects are given.

In some other situations where attributes are involved, the measurement of values in numerical terms are not possible at all. For example, the 'intelligence' and 'efficiency in salesmanship' of a group of sales man cannot be measured quantitatively. Here the individuals are ranked according to their merit for both the attributes and the degree of associatiation of the attributes is calculated based on these rankings. This method of computation of correlation of the attaibutes is called 'Rank Correlation'.

Rank correlation co-efficient is denoted by 'R' and it is formulated according to Charles Spearman as

$$R = 1 - \frac{6\Sigma D^2}{N\left(N^2 - 1\right)}$$

where, D is the difference between the ranks of an individual and N is the number of individuals. R lies between –1 and +1.

Tied Ranks : In some cases it is sometimes necessary to assign equal ranks to two or more individuals. In such cases, it is customary to give each individual an average rank. Thus if two individuals are ranked 5th place, they will be given the average rank $\frac{5+6}{2} = 5.5$ each. If three individuals are ranked equally at 5th place, then they will be given the average rank $\frac{5+6+7}{3} = 6$ each.

When equal ranks are assigned to some entries, an adjustment in the above formula for calculating rank correlation co-efficient is made by adding $\frac{1}{12}\left(M^3 - M\right)$ to the value of ΣD^2, where M stands for the number of individuals whose ranks are common. If there are more than one such group of individuals with common rank, this value will be added as many times as the number of such groups. The changed formula for ties will be

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + ...\right]}{N^3 - N}$$

Example : Find out the Rank Correlation :

Maths	29	32	53	47	45	32	70	45	70	53
Physics	56	60	72	48	72	35	67	67	75	31

Ans.

x	y	x	y	D	D^2
		Rank	Rank		
29	56	10.0	7.0	+3	9.00
32	60	8.5	6.0	+2.5	6.25
53	72	3.5	2.5	+1.0	1.00
47	48	5.0	8.0	–3.0	9.00
45	72	6.5	2.5	+4.0	16.00
32	35	8.5	9.0	–0.5	0.25
70	67	1.5	4.5	–3.0	9.00
45	67	6.5	4.5	+2.0	4.00
70	75	1.5	1.0	+0.5	0.25
53	31	3.5	10.0	–6.5	42.25
					97.00

M_1 = 1.5, 3.5, 6.5, 8.5 In 6 cases we have ties

M_2 = 2.5, 4.5 So we use average ranks

$$R = 1 - \frac{6\left\{\begin{array}{c}\Sigma D^2 + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_1^3 - M_1\right) \\ + \frac{1}{12}\left(M_1^3 - M_1\right) + \frac{1}{12}\left(M_2^3 - M_2\right) + \frac{1}{12}\left(M_2^3 - M_2\right)\end{array}\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left\{\begin{array}{c}97 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) \\ + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\end{array}\right\}}{10^3 - 10}$$

$$= 1 - 6\frac{\{97 + .5 + .5 + .5 + .5 + .5 + .5\}}{990}$$

$$= 1 - 6\frac{(100)}{990} = 1 - \frac{20}{33} = 1 - 0.6 = 0.4$$

4.3.6 Interpretation of Rank Correlation co-efficient

(i) If R > 0, this means high rank in one characteristic corresponds to high rank in the other and low rank in one corresponds to low rank in the other. For example, if the two characteristics are intelligence in two subjects say Maths and Physics— $R > 0 \rightarrow$ A student good in Maths in also good in Physics.

(ii) $R < 0 \rightarrow$ A students good in Maths is poor in Physics.

(iii) R = 1, means perfect correlation in the two characteristics, i.e., every individual is getting exactly the same rank in the two characteristics. Ranks are of the type (1, 1), (2, 2).....(n, n)

(iv) R = –1 means perfect negative correlation. Ranks are of the type (1, n), (2, n–1), (3, n–3).....(n, 1)

(v) $R = 0 \rightarrow$ means no correlation between the two characteristics.

4.4 Regression

Regression of a variable ‘y’ on another variable ‘x’ indicates dependence of y on x,

on the average. In bivariate analysis, one of the major problems is prediction of the value of the dependent variable y when the value of the independent variable x is given. In the simplest case when y is linearly related with x, we can write

$$y = a + bx$$

So that $a + bx_1$ is the predicted value of y when $x = x_1$

Here, we confine our discussion to linear regression only.

4.4.1 Scatter Diagram and Regression lines

In a bivariate distribution, if there is any relation between the two variables, then the points of the scatter diagram concentrate round some curve (we consider only the linear one). This curve is calted the curve of regression. Let us suppose we consider the height and weight of adult males for some given population. If we plot the pair (height, weight), a diagram known as scatter diagram will result. for any given height, there is a range of observed weights and vice-versa. This variation will be partially due to measurement errors but primarily due to variations between individuals. Thus no unique relationship between actual hight and weight can be expected. But we can note that average observed weight for a given observed height increases as height increases. The locus of average observed weight for given observed height is called the regression curve of weight on height. There also exists a regression curve of height on weight similarly defined. Let us assume that these two curves are both straight lines (which in general they may not be). When we are concerned with the dependence of a random variable Y and quantity X, which is variable, but not a random variable, an equation that relates Y to X is usually called a regression equation.

Economic theory is mainly concerned with relations among economic variables. The relationship is mostly stochastic in nature. Simplest stochastic relation between two variables X and Y is a linear one.

$$Y_i = \alpha + \beta X_i + U_i$$

where, Y is the dependent variable; Xi is the independent or explanatory variable and Ui is the stochastic disturbance, α and β are parameters of this simple linear Regression model.

Though economic theory formulates exact functional relationship among the variables, handling with common data will reveal that all observations do not fall exactly on a straight line. The best we can expect is that the observed quantities will be closer to the line. This is the reason why our regression model requires introduction of stochastic disturbance term. This error term represents the effects of all those factors which are not suspected by the investigator. We assume that the disturbance term is distributed normally with mean 0, So $E(U_i) = 0$, $E(U_i^2) = \sigma^2$ and $E(U_i U_j) = 0$.

Regression lines :

The method of estimating the relationship between Y and X is termed as ''Least square estimation method''. Under this method, the line of best fit is said to be that which minimises the sum of squared residuals between the points of the scatter diagram and the points on the straight line.

Suppose we have n observations on Y and X.

$Y_1, Y_2, \ldots\ldots\ldots Y_n$

and $X_1, X_2, \ldots\ldots\ldots X_n$

The estimated line is denoted as $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, where $\hat{\alpha}, \hat{\beta}$ are the estimated parameters and $\hat{Y}$ is the estimated value of Y.

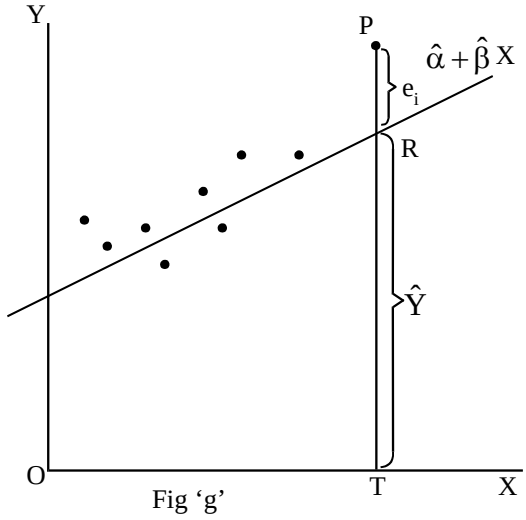


Fig 'g'

Taking any point P on the scatter diagram, we have OT = Xi, PT = Yi, RT = $\hat{Y}_i$

$\therefore$ Residuals or error is

$$e_i = Y_i - \hat{Y}_i = PR$$

These deviations of acutal values from estimated line will be positive or negative as the actual point lies above or below the estimated line. Squares of these residuals will be positive.

The principle of least squares is to choose such values of $\hat{\alpha}$ and $\hat{\beta}$ that will minimise the sum of squared deviations. The necessary condition for it is that the partial differentiation of Σe_i^2 with respect to $\hat{\alpha}$ and $\hat{\beta}$ should be equal to zero.

$$\Sigma e_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2$$

Differentiating w.r.t. $\hat{\alpha}$ and $\hat{\beta}$ we get,

$$\frac{\partial}{\partial\hat{\alpha}}\left[\Sigma e_i^2\right] = -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0$$

$$\frac{\partial}{\partial\hat{\beta}}\left[\Sigma e_i^2\right] = -2\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right).X_i = 0$$

On rearranging we get two equations

$$\Sigma Y_i = n\hat{\alpha} + \hat{\beta}\Sigma X_i \qquad .....(1)$$

$$\Sigma X_i Y_i = \hat{\alpha}\Sigma X_i + \hat{\beta}\Sigma X_i^2 \qquad .....(2)$$

Dividing both sides of equation (1) by n we get

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} \qquad .....(3)$$

Shifting the origin to $\left(\bar{X}, \bar{Y}\right)$ in equation (2) we get

$$\Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \hat{\alpha}\Sigma\left(X - \bar{X}\right) + \hat{\beta}\Sigma\left(X - \bar{X}\right)^2$$

But $\Sigma\left(X - \bar{X}\right) = 0$

$$\therefore \quad \hat{\beta} = \frac{\Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\Sigma\left(X - \bar{X}\right)^2} = r.\frac{\sigma_Y}{\sigma_X}$$

$$\left[\because\ r = \frac{1}{n}\frac{\Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sigma_X\sigma_Y}\right]$$

Subtracting(3) from the estimated line $Y = \hat{\alpha} + \hat{\beta}X$
we get the regression line Y on X,

$$\left(Y - \bar{Y}\right) = \hat{\beta}\left(X - \bar{X}\right) \quad \text{or,} \quad \left(Y - \bar{Y}\right) = \frac{r\sigma_Y}{\sigma_X}\left(X - \bar{X}\right)$$

$r\frac{\sigma_Y}{\sigma_X}$ is known as the regression co-efficient of Y on X.

From (3) we get the value of $\hat{\alpha}$

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} \text{ or, } \hat{\alpha} = \bar{Y} - r\frac{\sigma_Y}{\sigma_X}\bar{X}$$

Using the same line of argument we can derive the regression line X on Y

$$\left(X - \bar{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y - \bar{Y}\right)$$

where, $r\frac{\sigma_X}{\sigma_Y}$ is known as the regression coefficient of X or Y.

Remarks :

1. The two regression lines intersect at $\left(\bar{X}, \bar{Y}\right)$ since both the equations are satisfied by $X = \bar{X}$ and $Y = \bar{Y}$.
2. If r = 0, the two regression lines will be represented by the equations $Y = \bar{Y}$ and $X = \bar{X}$ which are parallel lines parallel to the two axes and perpendicular to each other at point $\left(\bar{X}, \bar{Y}\right)$.
3. If $r = \pm 1$, the two regression lines will coincide.

Finding the angle between the two lines of regression

Two regression lines are

$$Y = \bar{Y} + r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right) \qquad \text{.....(i)}$$

and $$X = \bar{X} + r\frac{\sigma_x}{\sigma_y}\left(Y - \bar{Y}\right) \qquad \text{.....(ii)}$$

The gradient of (i) is $r\frac{\sigma_y}{\sigma_x} = m_1$ (say)

and that of (ii) is $\frac{\sigma_y}{r\sigma_x} = m_2$ (say)

If θ be the acute angle between the two lines,

$$\theta = \tan^{-1}\left|\frac{m_2 - m_1}{1 + m_1 m_2}\right| = \tan^{-1}\left|\frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y}{r\sigma_x}\cdot\frac{r\sigma_y}{\sigma_x}}\right|$$

$$= \tan^{-1}\left|\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right|$$

The other angle between the two lines is $(\pi - \theta)$

when $r = \pm 1$, $\tan\theta = 0$ i.e., $\theta = 0$ and the two lines coincide

when $r = 0$, $\cot\theta = 0$ i.e., $\theta = \frac{\pi}{2}$ and the two lines are at right angles.

4.4.2 Properties of Regression co-efficients

Property 1 : The covariance, the correlation co-efficient and the two regression co-efficients all have the same sign.

Proof : $\text{cov}(x, y) = r\,\sigma_X\sigma_Y$

Regression co-efficient of Y on X = $b_{YX} = \frac{r\sigma_Y}{\sigma_X}$

Regression co-efficient of X on Y = $b_{XY} = \frac{r\sigma_X}{\sigma_Y}$

Since σ_X and σ_Y are both positve, $\sigma_X\sigma_Y$ and $\frac{\sigma_Y}{\sigma_X}, \frac{\sigma_X}{\sigma_Y}$ are all positive.

$\therefore$ Cov (X, Y), b_{YX}, b_{XY} have the same sign as r— the correlation co-efficient

Property 2 : Correlation co-efficient is the geometric mean between the two regression co-efficients.

Proof : The two regression co-efficients are

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

GM of the two regression co-efficients

$$= \sqrt{b_{YX}\times b_{XY}} = \sqrt{r\frac{\sigma_Y}{\sigma_X}\times r\frac{\sigma_X}{\sigma_Y}} = r$$

Property 3 : If one of the regression co-efficients is greater than unity numerically, the other is less than unity numerically.

Proof : $b_{YX}\cdot b_{XY} = r^2$

$$\left|b_{YX} \cdot b_{XY}\right| = r^2 \leq 1$$

or, $$\left|b_{YX}\right|\left|b_{XY}\right| = r^2 \leq 1$$

or, $$\left|b_{YX}\right| < \frac{1}{\left|b_{XY}\right|}$$

If $\left|b_{YX}\right| > I$, then $\frac{1}{\left|b_{XY}\right|} > 1 \;\therefore\; \left|b_{XY}\right| < 1$

If $\left|b_{XY}\right| > 1$ then $\frac{1}{\left|b_{XY}\right|} < 1 \;\therefore\; \left|b_{YX}\right| < 1$

Property 4 : Arithmatic mean of regression co-efficients is greater than the co-efficient of correlation.

Proof : We know that

$$AM > GM$$

$$\frac{b_{YX} + b_{XY}}{2} > \sqrt{b_{YX} \cdot b_{XY}}$$

$$\frac{b_{YX} + b_{XY}}{2} > r$$

Property 5 : Regression co-efficients are independent of the change of origin but not of change of scale.

Proof : Let, $u = \frac{x - a}{h}$, $v = \frac{Y - b}{k}$

where, a, b, h and k are arbitrary constants.

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}, \quad b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$\therefore$ $X = a + hu$ and $\bar{X} = a + h\bar{u}$

$\therefore$ $\left(X - \bar{X}\right) = h\left(u - \bar{u}\right)$

Similarly, $\left(Y - \bar{Y}\right) = k\left(v - \bar{v}\right)$

Hence, $$Var\left(X\right) = h^2 Var\left(u\right)$$

$$Var\left(Y\right) = k^2 Var\left(v\right)$$

and $$Cov\left(X, Y\right) = hk\, cov\left(uv\right)$$

Then, $b_{YX} = \frac{Cov(XY)}{Var(X)} = \frac{hk\,Cov(u,v)}{h^2 Var(u)} = \frac{k}{h}\frac{Cov(u,v)}{Var(u)} = \frac{k}{h}b_{vu}$

Similarly $b_{XY} = \frac{h}{k}b_{uv}$

which are independent of a and b but not of h and k.

4.4.3 Standard Error of Estimate

Standard deviation of e_i, the error term is called the Standard error of the estimate of Y from its linear regression on X.

$$Var(e) = \frac{1}{n}\Sigma e_i^2 \quad (\because \bar{e} = 0)$$

$$= \frac{1}{n}\Sigma\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{n}\sum_i\left[\left(Y_i - \bar{Y}\right) - r\frac{\sigma_y}{\sigma_x}\left(X_i - \bar{X}\right)\right]^2$$

$$= \frac{1}{n}\Sigma\left(Y_i - \bar{Y}\right)^2 - 2r\frac{\sigma_y}{\sigma_x}\cdot\frac{1}{n}\Sigma\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) + \frac{r^2\sigma_Y^2}{\sigma_X^2}\cdot\frac{1}{n}\Sigma\left(X_i - \bar{X}\right)^2$$

$$= \sigma_Y^2 - 2r\frac{\sigma_y}{\sigma_x}\cdot r\sigma_X\sigma_Y + r^2\frac{\sigma_Y^2}{\sigma_X^2}\cdot\sigma_X^2$$

$$= \sigma_Y^2 - 2r^2\sigma_Y^2 + r^2\sigma_Y^2$$

$$= \sigma_Y^2 - r^2\sigma_Y^2$$

$$= \sigma_Y^2\left(1 - r^2\right)$$

$\therefore$ Standard error $= \sqrt{\sigma_Y^2\left(1 - r^2\right)} = \sigma_Y\sqrt{\left(1 - r^2\right)}$

Since Var (e) ≥ 0 we have,

$$\sigma_Y^2\left(1 - r^2\right) \geq 0$$

or, $\left(1 - r^2\right) \geq 0$

or, $r^2 \leq 1$

$\therefore$ $-1 \leq r \leq +1$

It has been proved before,

From Var(e) $= \sigma_Y^2\left(1-r^2\right)$ we find that if r = 0 then Var(e) $= \sigma_Y^2$, so that errors of estimation are as much variable as the observed values, and hence the regression line is of no help as a prediction formula.

It is also noted that as the numerical value of r increases, Var(e) decreases and when $r = \pm 1$, Var(e) = 0 which implies that, for each i,

$$e_i = 0 \text{ or } Y_i = \hat{Y}_i$$

So that all the points in the scatter diagram lie on the regression line and thus, the regression line becomes a perfect prediction formula. From these observations, it is clear that the numerical value of r can be taken as a measure of the efficacy of the regression equation as a prediction formula.

4.4.4 Explained and Unexplained Variation

The total sum of squares of the deviations of the observed values of Y from their mean can be split up into two components viz., the sum of squares explained by linear regression of Y on X and the sum of squares unexplained by the regression of Y on X.

Total sum of squares

$$= \sum_{i=1}^{n}\left(Y_i = \bar{Y}\right)^2$$

$$= \sum_i\left[\left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}\right)\right]^2$$

$$= \sum_i\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_i\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_i\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right)$$

$$= \sum_i\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_i\left(Y_i - \hat{Y}_i\right)^2 \quad \text{.....(i)}$$

$$\left[\because \sum_i\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) = \sum_i \beta\left(X_i - \bar{X}\right)e_i = \beta\sum_i X_i e_i - \beta\bar{X}\Sigma e_i = 0\right.,$$

from Normal equations]

The first term in equation (i) on the RHS is the component which is explained by the linear regression of Y on X and the second term is the component which is unexplained, (see Fig. h)

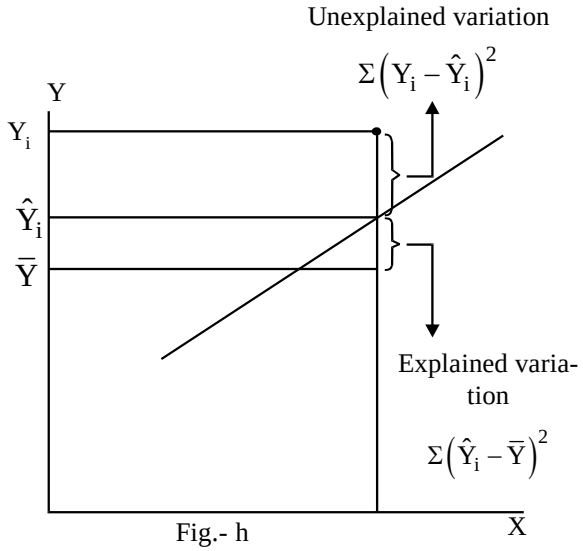


Fig.- h

R^2 : The co-efficient of determination

R^2 is defined as the explained variation as a proportion of the total variation, which is explained by the regression model. An $R^2 = 1$ indicates that all the variations have been explained, ie., each predicted value for the dependent variable must be exactly equal to the corresponding observed value.

At the other extreme, R^2 will be zero. In this case, the regression equation has been totally unable to explain variation in the dependent variable. In actual regression study, R^2 will seldom be equal to either 0 or 1. Usually in empirical works, R^2 is higher for time-series studies than cross-section studies, as exogenous factors are held constant in cross-section studies. A low R^2 indicates the inadequacy of the model, which generally arises for the omission of important variables from the model.

The co-efficient of determination is not highly reliable as it can be made artificially high if too small a sample is used to estimate the model's co-efficient. So a substantial number of data observations are needed to fit a regression model adequately, so that there is substantial number of degrees of freedom (d_f). Degree of freedom is defined as the number of data observations beyond the mininum necessary to calculate a given regression co-efficient or statistic. Since R^2 always approaches zero, statisticians have developed a method for correcting R^2 to account for the number of degrees of freedom. The corrected co-efficient of determination $\bar{R}^2$ is given by

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left[\frac{n-1}{n-(K+1)}\right]; \qquad 0 \le \bar{R}^2 \le 1$$

Where, n = Sample size

k = the number of independent variables in the regression equation

The reliability of a given regression model will be high when both $\bar{R}^2$ and degrees of freedom are substantial.

Example : Obtain the equations of the two lines of regression for the data given below :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

X	Y	$dx = X - \bar{X}$ $= x - 5$	$dy = Y - \bar{Y}$ $= Y - 12$	dx^2	dy^2	dxdy
1	9	–4	–3	16	9	12
2	8	–3	–4	9	16	12
3	10	–2	–2	4	4	4
4	12	–1	0	1	0	0
5	11	0	–1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
$\Sigma x = 45$	$\Sigma y = 108$	$\Sigma dx = 0$	$\Sigma dy = 0$	$\Sigma dx^2 = 60$	$\Sigma dy^2 = 60$	$\Sigma dxdy = 57$

$$\bar{X} = \frac{\Sigma x}{n} = \frac{45}{9} = 5 \; ; \qquad \bar{Y} = \frac{\Sigma y}{n} = \frac{108}{9} = 12$$

$$dx = X - \bar{X} = (x - 5); \qquad dy = Y - \bar{Y} = (Y - 12)$$

Regression co-efficients

$$b_{YX} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2} = \frac{\Sigma dxdy}{\Sigma dx^2} = \frac{57}{60} = 0.95$$

$$b_{XY} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (Y - \bar{Y})^2} = \frac{\Sigma dxdy}{\Sigma dy^2} = \frac{57}{60} = 0.95$$

Regression equation of Y on X :

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \quad \text{or, } Y - 12 = 0.95\ (X - 5)$$

or, $Y = 0.95X + 7.25$

Regression equation of X on Y

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

or, $\left(X-5\right)=0.95\left(Y-12\right)$

or, $X = 0.95Y - 6.40$

Example : Estimate the loss in production in a week when the number of workers on strike is 1800, from the following data :

Mean no. of workers on strike = 800
Mean loss of daily production in '000 Rs = 35
Standard deviation of No. of workers on strike = 100
S.D. of loss of daily production in '000 Rs. = 2
Co-efficient of correlation between No. of workers on strike and daily production loss = 0.8

Ans. Let the no. of workers on strike be denoted by variable X and the loss of daily production by the variable Y. In the usual notation we are given

$\bar{X}=800$, $\bar{Y}=35$, $\sigma_X = 100$, $\sigma_Y = 2$, $r_{XY} = 0.8$

To estimate the loss in daily production (Y) when the number of workers on strike (X) is 1800, we need the regression equation of Y on X which is given by

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

or, $$Y=r\frac{\sigma_y}{\sigma_x}\cdot\left(X-\bar{X}\right)+\bar{Y}$$

or, $$Y=\frac{0.8\times 2}{100}\left(X-800\right)+35=0.016\left(X-800\right)+35$$

The loss of daily production when the number of workers on strike is 1800, is obtained on taking X = 1800 in the above equation,

$$Y_{X=1800}=0.016\left(1800-8000\right)+35=16+35 = \text{Rs. } 51 \text{ ('000)}$$

Hence the loss in production in a week (6 working days)

= Rs 51 × 5('000) = Rs. 306 ('000) = 306 × 1000 = Rs. 306000/-

Example : Given the standard deviation σ_X and σ_Y for two correlated variables X and Y in a large sample.

(a) What is the standard error in estimating Y, X if r = 0?

(b) By how much is the error get reduced if r = 0.5?

(c) What is the standard error in estimating Y from X if r = 1 ?

Ans. Standard error of estimate of Y for given X

$$SE_{YX} = \sigma_Y \sqrt{1-r^2}$$

(a) when r = 0, $SE_{YX} = \sigma_Y \sqrt{1-0} = \sigma_Y$

(b) When r = 0.5, $SE_{YX} = \sigma_Y \sqrt{1-(0.5)^2} = \sigma_Y \sqrt{0.75} = 0.87\sigma_y$

Hence if the value of r is increased from 0 to 0.5, the reduction in the standard error of the estimate is :

$$\sigma_y - 0.87\sigma_y = (1-0.87)\sigma_y = 0.13\sigma_y$$

(c) when r = 1, $SE_{YX} = \sigma_y \sqrt{1-r^2} = \sigma_y.0 = 0$

4.5 Summary

Measures of Correlation to find out the degree of association between the variables.

Karl Pearson's Correlation co-efficient

(when deviations are taken from the actual means)

1. $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{N\,\sigma_x\sigma_y}$ or, $\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot (y-\bar{y})^2}}$

2. (when deviations are taken from assumed mean)

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x \cdot \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\ \sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

where, d_x = (X – A) and d_y = (Y – A)

3. In a bivariate frequency distribution where f is the frequency

$$r = \frac{N\Sigma f d_x d_y - \Sigma f d_x \cdot \Sigma f d_y}{\sqrt{N\Sigma f d_x^2 - (\Sigma f d_x)^2} \cdot \sqrt{N\Sigma f d_y^2 - (\Sigma f d_y)^2}}$$

4. When we deal with actual values of x and y

$$r = \frac{\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{N\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{N\Sigma y^2 - (\Sigma y)^2}}$$

5. Spearman's Rank correlation co-efficient

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

where, D refers to the difference of rank between paired items in two series.

In case ranks are repeated

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}\left(m^2 - m\right) + \frac{1}{12}\left(m^3 - m\right) + ...\right]}{N^3 - N}$$

Regression :

6. Regression equation of x on y :

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\Sigma\left(y - \bar{y}\right)^2}$$

(if deviations are taken from actual means of x and y)

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \cdot \Sigma d_y}{N}}{\Sigma d_y^2 - \frac{\left(\Sigma d_y\right)^2}{N}}$$

(if deviations are taken from assumed means of x and y)

7. Regression equation of Y on X :

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(X - \bar{X}\right)$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\Sigma\left(x - \bar{x}\right)^2}$$

(if deviations are taken from the actual means)

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \cdot \Sigma d_y}{N}}{\Sigma d_x^2 - \frac{\left(\Sigma d_x\right)^2}{N}}$$

(if deviations are taken from the assumed means)

8. Regression co-efficients :

$r\frac{\sigma_x}{\sigma_y}$ or b_{xy} is the regression co-efficient of x on y.

$r\frac{\sigma_y}{\sigma_x}$ or b_{yx} is the regression co-efficient of y on x.

Correlation co-efficient $= r = \sqrt{b_{xy} \cdot b_{yx}}$

9. Standard errors of estimate :

$$Se_{xy} = \frac{\sqrt{\Sigma\left(X - \hat{X}\right)^2}}{N}$$

or, $Se_{xy} = \sigma_x\sqrt{1-r^2}$

Standard error of estimate :

$$Se_{yx} = \frac{\sqrt{\Sigma\left(y - \hat{y}\right)^2}}{N}$$

or, $Se_{yx} = \sigma_y\sqrt{1-r^2}$

4.6 Questions

1. Choose the correct answer :

(i) The correlation co-efficient lies between
(a) 1 and 2 (b) –1 and +1 (c) 0 and 1 (d) None

(ii) If the two regression lines coincide, then r =
(a) 1 (b) –1 (c) ±1 (d) 0

(iii) Two regression lines are 2x + 3y – 4 = 0 and x + 2y + 6 = 0
The correlation co-efficient between x and y is
(a) $-\frac{3}{4}$ (b) $\frac{3}{4}$ (c) $\frac{\sqrt{3}}{2}$ (d) $\pm\frac{\sqrt{3}}{2}$

(iv) If the correlation coefficient between x and y is 0.4, then the correlation co-efficient between 3x and –2y is

(a) 0.4 (b) –0.4 (c) –1 (d) None

(v) If the two regression lines are mutually perpendicular, then the correlation coefficient equals

(a) 0 (b) ±1 (c) 1 (d) –1

(vi) If b_{yx} = –0.8, r = –0.96, S.D. of y = 12 then S.D. of x is equal to

(a) 12 (b) 20 (c) 10 (d) None

(vii) The minimum value of co-efficient of determination is

(a) –1 (b) 0 (c) 1 (d) None

(viii) If x + y = 50 then the correlation coefficient between x and y is

(a) –1 (b) 0 (c) 1 (d) 0.5

(ix) The lines of regression concerning to the variables x and y are given by y = 32 – x and x = 13 – 0.25y. the values of the means are

(a) 6.7, 25.3 (b) 4.2, 9.7 (c) 7.9, 24.8 (d) None

(x) b_{xy} = –0.2, b_{yx} = –1.8 then r_{xy} is equal to

(a) 0.36 (b) 0.6 (c) ±0.6 (d) –0.6

2. Fill in the blanks :

(i) Regression co-efficients are affected by change of _____ only.

(ii) Pearson's co-efficient of correlation is a measure of _____ association between two variables.

(iii) If the sum of squares of differences of ranks given by two judges of 10 students is 33, then spearman's rank correlation co-efficient is equal to_____.

(iv) If one regression co-efficient is greater than unity, then the other must be _____ than unity.

(v) If perfect agreement exists between two series of ranks then spearman's rank correlation co-efficient equals to _____.

3. Define product-moment correlation co-efficient.
4. What is a scatter diagram?
5. Prove that the correlation co-efficient lies between –1 and +1.
6. What are the limitations of the correlation co-efficient as a measure of association between two variables?

7. Prove that the correlation co-efficient is independent of change of origin and numerically it is also independent of the change of scale.

8. Are the following data consistent?
Cov(x, y) = 14, Var(x) = 5 and Var(y) = 20

9. Two Judges in a beanty competition rank the 12 entries as :

x	1	2	3	4	5	6	7	8	9	10	11	12
y	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the two judges?

10. From the following data calculate the co-efficient of rank correlation between x and y.

x	32	55	49	60	43	37	43	49	10	20
y	40	30	70	20	30	50	72	60	45	25

11. The following data gives the marks obtained by 10 students in Accountancy and statistics.

Student	1	2	3	4	5	6	7	8	9	10
Marks is Accountancy	45	70	65	30	90	40	50	75	85	60
Marks in Statistics	35	90	70	40	95	40	60	80	80	50

12. Obtain the equations of the two lines of regression for the data below :

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

13. Assuming that we conduct an experiment with eight fields planted with corn, four fields having no nitrogen fertiliser and four fields having 80 kg of nitrogen fertiliser. The resulting corn yields are shown in the table in bushels per hectare :

Field	:	1	2	3	4	5	6	7	8
Nitrogen (kgs)	:	0	0	0	0	80	80	80	80
Corn yield/Hecture	:	120	360	60	180	1280	1120	1120	760

(a) Compute a linear regression equation by least square method. Explain the meaning of regression equation in terms of fertiliser and corn yield.

(b) Predict corn yield for a field treated with 60 kg of fertiliser

14. The following table gives the age and blood pressure of 10 women :

Age (X) :	56	42	36	47	49	42	60	72	63	55
Blood Pressure (Y) :	147	125	118	128	145	140	155	160	149	150

(i) Find the correlation co-efficient between x and y.

(ii) Determine the least square regression equation of Y on X.

(iii) Estimate the blood pressure of a Woman whose age is 45.

4.7 References

1. Kenney & Keeping (1953) Mathematics of statistics, Van Nostrand Co.
2. Bowen and starr (1982) Basic statistics for Business and Economics, Macgraw Hill.

Unit 5 ❐ Index Numbers and their Applications

Structure

5.1 Objectives

5.2 Introduction

5.3 Important Factors regarding construction of Index Number

5.4 Construction of Price and Quantity Index Numbers

5.4.1 Various Formulae

5.4.2 Construction of Quantity Index Numbers

5.4.3 Value Indices

5.4.4 Simple Average of Price Relatives

5.4.5 Weighted Average of Price Relatives

5.4.6 Construction of General Index from Group Indices

5.4.7 Errors in Index numbers

5.5 Test of Consistency for Index Number Formulac

5.5.1 Fulfilment of Tests of Index Numbers

5.6 Chain Indices

5.6.1 Chain Base method and Fixed Base Method

5.7 Base Shifting, Splicing and Deflating of Index Numbers

5.8 Cost of Living Index Number

5.8.1 Steps in the Construction of Cost of Living Index (CLI) or, Consumer Price Index (CPI)

5.8.2 Uses of cost of Living Index Number

5.9 Uses of Index Number

5.10 Bias in Index Number

5.11 Some Worked out Example

5.12 Summary

5.13 Questions

5.14 References

5.1 Objectives

Newspapers flash in their head lines the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular time period compared to a previous one. In each and every case we use 'Index Number' to indicate the rise or fall in the economic variables. The main objective of this exercise with index numbers is to feel the pulse of the economy. These indicators in the form of index numbers are called the 'Barometers of economic activity'. If one wants to get an idea as to what is happening to an economy, one should look into the important indices like the index number of industrial production, agricultural production, business activity, GDP, Imports and Exports.

5.2 Introduction

Historically, the first index was constructed in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in prices, index numbers have become today one of the most widely used statistical devices and there is hardly any field left where they are not used.

An index number is a summary measure of change is the magnitude of a certain variable. For instance, consider the price of food crops. We may be interested in knowing whether the price of food crops has changed this year as compared to last year. We are here comparing two price situations over time. Secondly, food crops do not comprise a homogeneous commodity but we have several crops with several varieties. By constructing a price index number we try to summarise the differential price movements of this heterogeneous collection into a single number. Instead of the price of food crops if we are interested in the production of food crops, then the comparison is between two quantity situations. The index number concerned is the quantity index. We may enquire about the prices of food crops in Bihar and Punjab in a particular year. This will be spatial comparison. Theoretically the basic problem of index number construction remain the same wheather we are making inter temporal or spatial comparisons.

When people say the prices have increased or that the cost of living has gone up, it does not necessarily mean that the prices of all commodities have gone up or that prices have gone up equally for all the commodities, prices of some commodities may actually have fallen.

We read in news papers that Dearness Allowance of all Govt. employees have been increased when consumer price index number goes up by certain points. Thus index number helps the Govt. in its policy making exercise.

5.3 Important Factors regarding construction of Index Number

The following steps are considered for the construction of Index Number :

1. Purpose for which the index number is needed

The objective of the desired index number must be stated clearly otherwise the construction of the index number cannot be appropriately carried on. For example, the whole sale prices of commodities have to be collected for the construction of the wholesale price index number, while retail prices will be collected for the construction of cost of living index number.

2. Selection of base year

Suppose the index number measures the relative changes in the price level in a time period as compared to the price level on a previous date. The previous date is known as the 'Base year' where as the former time period is konwn as the 'current year'. The selection of the base period or base year is very impotant. Base year must be a normal year. It should not be affected by natural calamities like floods, earthquakes or economic depression, boom, war, strikes etc. The base year should not be a distant past relative to the current year. If the base year is in the remote past, it should be shifted to a year in the recent past.

3. Selection of items to be included

Since all the items cannot be taken into consideration which requires considerable time and money, we have to make judgement sampling of the items instead of 'random sampling'. The representative item should possess the following characteristics :

(a) It should be representative of the taste, habits, customs and necessities of the people to whom the index number relates.

(b) It should be stable in quality and preferably should be graded or a standardised one.

(c) It should be as far as possible large in number. Since larger the number the greater is the chance of accuracy.

(d) There should be varieties of commodities to make them more representative

(e) There should be proper classification of items included in the construction of index number

4. Collection of data

Suppose, we are preparing the price index number. In a specified period, the price of the commodities concerned do not remain the same for its various brands/grades and also in all the markets.

In practice, we collect retail or wholesale prices (depending on the purpose of the index number) of the commodities from a few renowned and representative markets for some of

their major brands/grades. Since the reliability of the price index number depends much on the precision of data on price quotations, such data should be collected from reliable sources.

5. Selection of Average

The technique of index number is used to study the changes in general price level and in such cases more than one commodity have to be taken into account. Here arise the problem of selecting an average. Theoretically, any average can be used. But practically a choice has to be made between Arithmatic mean, median and geometric mean. The geometric mean is supposed to be the better average so far as the construction of index number is concerned. In index number we deal with ratios and relative changes and GM gives equal weights to equal ratios of change. (GM of ratios = Ratios of GM)

6. Selection of weights

Weights should be assigned in a proper and logical way. But it is difficult to define rigidly the rational weights. Weights may be perfectly rational for one investigation but may be quite unsuitable for another. The purpose of indices, the nature of data decides as to what shall constitute the rational weights. Apart from the question of rational weights, another question arises as to whether there should be fixed weight or a fluctuating one. The simple answer is that it should be fluctuating since changing weights are more accurate as they give reliability to the indices. Commonly adopted systems of weighting are :

(i) Quantity weights in which the various commodities are attached importance according to the amount of their quantity used, purchased or consumed.

(ii) The value weights in which the importance to the various items is assigned according to the expenditure involved on them.

7. Interpretation of the Index number

The interpretation of an index number depends on the purpose for which it has been constructed. Suppose, the wholesale price index number for a country in 2018 with 2010 as base period is found to be 307.5. This indicates that the general price level of the country has increased 3.075 times or by 207.5%.

5.4 Construction of Price and Quantity Index numbers

Price Index numbers can be constructed by using two methods (a) aggregative method and (b) average of price relatives method.

(a) (i) Simple Aggregative method

Here simple aggregate of actual price for the current year is to be compared with that for the base year

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100 \qquad ...(1)$$

p_0 is the price in the base year '0' and p_1 is the price in the current year '1'.

(a) (ii) Weighted aggregative method

The commodities that are included in the construction of the price index number are not of equal importance. For example wheat should be given more importance than tobacco in constructing the whole sale price index number for India. Hence, we must assign appropriate weights (or relative importance) to the prices of commodities and subsequently construct the index number.

Generally, the quantities consumed, produced or sold in the base period, the current period or some other reference period are used as weights.

If w is the weight attached to a commodity, then the price index number is given by

$$p_{01} = \frac{\sum wp_1}{\sum wp_0} \times 100 \qquad ...(2)$$

5.4.1 Various Formulac

By using different systems of weighting we get a number of formulae. Some of the important formulae are given below :

Laspeyre's Price Index (or Base year method)

Taking base year quantities as weights i.e., $w = q_0$ in (2), we get Laspeyre's Price index given by

$$p_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \qquad ... (3)$$

Paasche's Price Index

If we take current year quantities as weights i.e., $w = q_1$ in (2) we get Paasche's price Index

$$p_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 \qquad ... (4)$$

Dorbish-Bowley Price Index

This index is the arithmatic mean of Laspeyre's and Passche's price index numbers.

$$P_{01}^{DB} = \frac{1}{2}\left[P_{01}^{La} + P_{01}^{pa}\right]$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100 \qquad \ldots (5)$$

Fisher's Price Index

Irving Fisher's index is the geometric mean of Laspeyre's and paasche's price index numbers and is formulated as

$$P_{01}^{F} = \left[P_{01}^{La} \times P_{01}^{pa}\right]^{\frac{1}{2}} = \left[\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}\right]^{\frac{1}{2}} \times 100 \quad \ldots(6)$$

Marshall-Edgeworth price Index

Taking w = $\frac{q_0 + q_1}{2}$ in (2), we get the Marshell-Edgeworth formula for Price Index :

$$P_{01}^{ME} = \frac{\sum p_1 (q_0 + q_1)/2}{\sum p_0 (q_0 + q_1)/2} \times 100$$

$$= \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100 \qquad \ldots (7)$$

$$= \left[\frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1}\right] \times 100 \qquad \ldots (7a)$$

Walsch Price Index

Taking w = $\sqrt{q_0 q_1}$, we get Walsch price index from equation (2)

$$P_{01}^{w} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100 \qquad \ldots (8)$$

Kelly's Price Index or Fixed Weight Index

$$P_{01}^{k} = \frac{\sum p_1 q}{\sum p_0 q} \times 100 \qquad \ldots (9)$$

where the weights are the quantities (q) which may refer to some period (not necessarily base year or the current year) and are kept constant for all periods.

REMARKS

1. Laspeyre's Index vs Paasche's Index

Laspeyre's price index is expected to have an 'upward bias' as it over-estimates the true value, where as Pasche's price index has a 'downward bias' and is expected to under-estimate the true value.

2. Marshall-Edgeworth and Fisher's Index Numbers

(i) These formulae are a sort of compromise between Laspeyre's index (which has an upward bias) and Paasche's price index (which has a downward bias) and provide a better estimate of the true price index.

(ii) Fisher's index is termed as 'ideal index' since it satisfies the Time Reversal and the Factor Reversal tests for the consistency of Index numbers.

(iii) Both the Fisher's ideal index and Marshall-Edgeworth index lie between Laspeyre's and Paasche's indices

5.4.2 Construction of Quantity Index Number

The formulae for quantity indices are obtained from the Price Index formulae (3) to (9) on interchanging prices (p) and quantities (q).

Thus, for example : Laspeyre's quantity index

$$Q_{01}^{La} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100 = \frac{\sum p_0 q_1}{\sum p_0 q_0} \times 100 \qquad \ldots (10)$$

Paasche's quantity index $Q_{01}^{pa} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{\sum p_1 q_1}{\sum p_1 q_0} \times 100 \qquad \ldots (11)$

Fisher's quantity index

$$Q_{01}^{F} = \left[Q_{01}^{La} \times Q_{01}^{Pa} \right]^{1/2} = \left[\frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0} \right]^{\frac{1}{2}} \times 100 \qquad \ldots (12)$$

Marshall-Edgeworth quantity Index

$$Q_{01}^{ME} = \frac{\sum q_1(p_0 + p_1)}{\sum q_0(p_0 + p_1)} \times 100 = \frac{\sum q_1 p_0 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_0 p_1} \times 100 \quad \text{... (13)}$$

ans so on.

5.4.3 Value Indices

Value index numbers are obtained on expressing the total value (or expenditure) in any given year as a percentage of the same in the base year. Symbolically, the formula is given by

$$V_{01} = \frac{\text{Total value in the current year}}{\text{Total value in the base year}} \times 100$$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100 \quad \text{... (14)}$$

5.4.4 Simple Average of Price Relatives

$$P = \text{Price Relative for a commodity} = \frac{p_1}{p_0} \times 100 \quad \text{... (15)}$$

Price relatives are the simplest form of the index numbers for each commodity. The price index for the composite group is obtained on averaging these price relatives by using arithmatic mean (A.M.) or geometric mean (G.M) or Harmonic Mean (HM).

Price index using simple arithmatic mean of the relatives is given by :

$$p_{01}\text{A.M.}) = \frac{1}{n}\sum\left(\frac{p_1}{p_0} \times 100\right) = \frac{1}{n}.\sum p \quad \text{... (16)}$$

where, n is the number of commodities in the group.

Using simple geometric mean of the price relatives, the price index is given by :

$$p_{01}(\text{G.M.}) = \left[\Pi\left(\frac{p_1}{p_0} \times 100\right)\right]^{\frac{1}{n}} = [\Pi p]^{\frac{1}{n}} \quad \text{... (17)}$$

where, Π denotes the product of the price relatives for the n commodities

$$p_{01}(\text{G.M.}) = \text{Anti}\log\left[\frac{1}{n}\sum \log p\right] \quad \text{... (17a)}$$

Using harmonic mean of the price relatives, the price index is given by :

$$P_{01}(HM) = \frac{n}{\sum \frac{p_0}{p_1} \times 100} = \frac{n}{\sum P} \qquad \ldots (18)$$

5.4.5 Weighted Average of Price Relatives

As in the case of aggregative index method, we introduce weights which is the value of commodities sold, produced or consumed in the base year.

So the weighted price relatives averaged with the arithmatic mean is given by

$$P_{01}(\text{weighted AM}) = \frac{\sum w\left(\frac{p_1}{p_0} \times 100\right)}{\sum w}$$

$$= \frac{\sum wp}{\sum w} \qquad \ldots (19)$$

where w is the weight attached to the pirce relative p.

Similarly, weighted GM is given by,

$$P_{01}(\text{weighted G.M.}) = \left[\Pi\left(\frac{p_1}{p_0} \times 100\right)^w\right] \frac{1}{\sum w} \qquad \ldots(20)$$

$$\text{or, } P_{01}(\text{weighted GM}) = \text{Anti log}\left[\frac{\sum w \log p}{\sum w}\right]$$

Likewise, Weighted Harmonic Mean is given by

$$P_{01}(\text{weighted HM}) = \frac{\sum w}{\sum\left(\frac{p_0}{p_1} \times 100\right) w} = \frac{\sum w}{\sum pw} \qquad \ldots (21)$$

Remarks : If we use the base year values, p_0q_0 for 'w' in (19), then we find Laspeyre's formula. If we substitute p_1q_1, the current year values for 'w' in equation (21), we get the Paasche's formula.

Similarly, if we take $w = p_1q_0$ in (21) and $w = p_0q_1$ in (19), we get Laspeyre's and Paasche's formula respectively.

So, it should be noted that Laspeyre's and Paasche's index numbers may be obtained as the weighted overages of price relatives.

5.4.6 Construction of General Index from Group Indices

The General Index for the composite group of commodities is obtained on taking the weighted average (usually A.M.) of group indices as given below.

$$\text{General Index} = \frac{\sum IW}{\sum W} \qquad \ldots (22)$$

where, I represents the group index and w is the group weight.

5.4.7 Errors in Index Numbers

The index numbers are subject to some errors. The errors are generally classified as (i) formula error. (ii) sampling error and (iii) homogeneity error.

The formula error arises due to the choice of a particular formula in the construction of an index number. There is no universally accepted formula which can measure the price change exactly. Hence each formula is subject to some inherent errors.

The sampling error arises due to the selection of commodities out of complete list. All commodities of the base and current periods cannot be included in the list. So there will be a sampling error. Naturally, as the number of commodities included increases, sampling error decreases.

Homogeneity errors arises due to the fact that index numbes are calculated from data on binary commodities. But it should be based on all the commodities marketed in the base period and current period. With the passage of time, many old commodities disappear and new commodities appear in the market. So, as the gap between the base period and the current period increases, the homogeneity error also increases.

5.5 Test of Consistency for Index Number Formulae

Irving Fisher has considered certain tests in order to judge the efficiency of an index number.There are 3 such tests– time reversal test, factor reversal test and circular test.

Time Reversal Test

According to this test, a good formula of index number should be time-consistent. The test requires that we should get the same picture of the change in the price level if the base and current periods are interchanged. Symbolically, $I_{on} \times I_{no} = 1$. ... (23)

An index number formula which obeys this relation is said to satisfy the time reversal test. Form example, if the price of a commodity changes form Rs. 4 per unit in 2000 to Rs 5/- in 2009, the price in 2009 is 125% of the price in 2000, and the price in 2000 is 80% of the price in 2009. The product of the two price ratios is $1.25 \times .80 = 1$.

Time reversal test is satisfied by simple aggregative formula. Marshall-Edgeworth formula,

Fisher's ideal index formula and simple G.M. of relatives formula.

Factor Reversal Test

An index number formula is said to satisfy the factor reversal test if the product of price and quantity indices gives the true value ratio. Symbolically,

$$P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0} \qquad \ldots (24)$$

It thus states that the product of price ratio and quantity ratio equals the value ratio. For example, suppose price rises from Rs. 4 to Rs. 8 quantity rises from 10 to 30 units. So, the product of price and quantity ratios = $\frac{8}{4} \times \frac{30}{10} = 6$. This is equal to the value ratio = $\frac{p_n q_n}{p_0 q_0} = \frac{8 \times 30}{4 \times 10} = 6.$

Fisher's ideal index is the only formula which satisfies this test.

Circular test

This is an extension of time reversal test. An index number formula is said to satisfy the circular test if the time-reversal test is satisfied through a number of inter mediate years. Symbolically.

$I_{01} \times I_{12} \times I_{23} \times \ldots \times I_{(n-1),n} \times I_{no} = 1 \qquad \ldots (25)$

This means that the relation is satisfied in a circular fashion through 0 to 1, 1 to 2, ..., (n – 1) to n, and finally from n back to zero. Simple aggregative formula and simple GM of relatives formula satisfy this test. Weighted aggregative formula and weighted GM of relatives formula satisfy this test if constant weights are used for all time periods.

5.5.1 Fulfilment of Tests of Index Numbers

Theorem 1 : Neither Laspeyres' formula nor Paasche's formula obeys time reversal or factor reversal tests.

Proof : Time reversal test may be symbolically expressed as, $I_{on} \times I_{no} = 1$. Now, using Laspeyres' formula, $I_{on}^{L} = \frac{\sum p_n q_0}{\sum p_0 q_0}$ (omitting 100). Interchanging the suffixes 0 and n, $I_{no}^{L} = \frac{\sum p_0 q_n}{\sum p_n q_n}$.

$\therefore\ I_{on}^{L} \times I_{no}^{L} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_n}{\sum p_n q_n} \neq 1.$

Thus, Laspeyres' formula does not obey the time reversal test.

Again, using Paache's formula, $I_{on}^{p} = \frac{\sum p_n q_n}{\sum p_0 q_n}$ (omitting 100).

Interchanging the suffixes o and n, $I_{no}^{p} = \frac{\sum p_0 q_0}{\sum p_n q_0}$

Now, $I_{on}^{p} \times I_{no}^{p} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0} \neq 1$

Thus, Paasche's formula does not obey the time reversal test.

Now, one factor-reversal test requires, $P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$. From Laspeyres' price inex, $P_{on}^{L} = \frac{\sum p_n q_0}{\sum p_0 q_0}$. Interchanging p and q, We get Laspeyres' quantity index, $Q_{on}^{L} = \frac{\sum q_n p_0}{\sum q_0 p_0}$.

$\therefore\ P_{on}^{L} \times Q_{on}^{L} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum q_n p_0}{\sum q_0 p_0} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$

Thus, Laspeyres' formula does not satisfy factor reversal test.

Again, Paasche price index = $P_{on}^{P} = \frac{\sum p_n q_n}{\sum p_0 q_n}$.

Interchanging p and q, we get Paasche Quantity index = $Q_{on}^{P} = \frac{\sum q_n p_n}{\sum q_0 p_n}$.

$\therefore\ P_{on}^{p} \times Q_{on}^{p} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum q_n p_n}{\sum q_0 p_n} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$

Thus, Paasche formula does not obey the factor-reversal test.

Theorem 2 : Show that Fisher's Index Satisfies (a) Time reversal test and (b) Factor-reversal test.

Proof : (a) Time reversal test requires. $I_{on} \times I_{no} = 1$

Now, from Fisher's price index formula, $I_{on}^{F} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$ (omitting 100).

Interchanging 0 and n, $I_{no}^{F} = \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}}$.

Now, $I_{on}^{F} \times I_{no}^{F} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}}$

= 1. so, Fisher's formula obeys time-reversal test.

(b) Factor Reversal test requires, $P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$ From Fisher's formula,

$P_{on}^{F} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$.

Interchanging p and q, $Q_{on}^{F} = \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum q_n p_n}{\sum q_0 p_n}}$

So, $P_{on}^{F} \times Q_{on}^{F} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum q_n p_n}{\sum q_0 p_n}} = \frac{\sum p_n q_n}{\sum p_0 q_0}$

Thus, Fisher's formula obeys factor reversal test also. Hence it is termed as Fisher's ideal index.

Question : Examine whether Edgeworth-Marshall formula obeys(a) Time reversal test, (b) Factor reversal test.

Ans. (a) Our time reversal test requires, $I_{on} \times I_{no} = 1$

From Edgeworth Marshall price index, $I_{on}^{EM} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)}$.

Interchanging o and n, $I_{no}^{EM} = \frac{\sum p_0 (q_n + q_0)}{\sum p_n (q_n + q_0)}$.

So, $I_{on}^{EM} \times I_{no}^{EM} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times \frac{\sum p_0 (q_n + q_0)}{\sum p_n (q_n + q_0)} = 1$

Thus, Edgeworth-Marshall formula obeys the time reversal test.

(b) Factor-reversal test requires, $P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$.

From Edgeworth-Marshall price index, $P_{on}^{EM} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)}$.

Interchanging p and q, we get Edgeworth-Marshall quantity index, $Q_{on}^{EM} = \frac{\sum q_n (p_0 + p_n)}{\sum q_0 (p_0 + p_n)}$.

So, $P_{on}^{EM} \times Q_{on}^{EM} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times \frac{\sum q_n (p_0 + p_n)}{\sum q_0 (p_0 + p_n)} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$.

So, Edgeworth-Marshall formula does not obey the factor reversal test.

5.6 Chain Indices

The chain base method consists in computing a series of index number by a suitable method for each year with the preceding year as the base year.

The steps in the construction of the chain base index number may be summarised as follows :

1. For each commodity express the price in any year as a percentage of its price in the preceding year. This gives the Link Relatives" (L.R.). Thus,

$$\text{L.R. for period i} = \frac{p_i}{p_{i-1}} \times 100, \qquad (i = 1, 2 \ldots r) \qquad \ldots(26)$$

2. Chain Base Indices (CBI) are obtained from the Link Relatives (LR) by the formula :

$$\text{CBI for any year} = \frac{\text{current year L.R.} \times \text{Preceding bear C.B.I}}{100} \qquad \ldots(27)$$

The CBI for the first period being the same as FBI (Fixed Base Index) for the first period.

Conversion of Chain Base Index Number to Fixed Base Index Number

Fixed Base Index (FBI) numbers can be obtained from the Chain Base Index (CBI) numbers by using the following formula :

$$\text{Current year FBI} = \frac{\text{Current year CBI} \times \text{Previous year FBI}}{100} \qquad \ldots (28)$$

The FBI for the first period being the same as the CBI for the first period.

5.6.1 Chain Base vs Fixed Base Method

The fixed base index numbers become more and more inaccurate as the distance between the base period and the current period increases. As the chain base index numbers are based on a number of link-indices, they are more accurate than fixed base method specially in long-term comparison.

Further, a chain index fully utilises the information regarding prices and quantities of all the intermediate periods between the base period and the current period. But a fixed base index is based on data of the base period and current period only.

Some argue that as a chain index is obtained by multiplying a number of link indices, it may involve a cumulative error. However, there is no convincing proof of this statement.

Lastly, fixed base indices are easier to calculate and are more easily understood than chain base method.

5.7 Base Shifting, Splicing and Deflating of Index Numbers

1. Base Shifting : Base shifting means changing of the given base year of a series of index numbers and recasting them into a new series based on some recent new base period.

Taking the index number of the new base year as 100, the series of index numbers, recast with a new base is obtained by the formula :

$$\text{Recast index number of any year} = \frac{\text{Old Index number of the year}}{\text{Index number with new base year}} \times 100 \ldots (29)$$

$$= \left(\frac{100}{\text{Index Number with new base year}}\right) \times (\text{old index number of the year}) \ldots (30)$$

In other words, the new series of index numbers is obtained on multiplying the old index numbers with a common factor : $\frac{100}{\text{Index No. with new base year}}$

Example : The following are the index number of wholesale prices of a certain commodity based on (1972 = 100).

Year :	1972	1973	1974	1975	1976
Index no :	100	108	120	150	210

Shift the base to 1974 and obtain the new series.

Ans. Shifting of base from 1972 to 1974

Year (1)	Index no (2)	(3) Index no. (Base 1974 = 100)
1972	100	$\frac{100}{120}\times 100 = 83.33$
1973	108	$\frac{100}{120}\times 108 = 90.00$
1974	120	$\frac{100}{120}\times 120 = 100$
1975	150	$\frac{100}{120}\times 150 = 125$
1976	210	$\frac{100}{120}\times 210 = 175$

Multiplying factor = $\frac{100}{120} = 0.8333$

Working formula : (3) = $\frac{100}{120}\times(2) = 0.8333\times(2)$

So the new series with (1974 = 100) as the base year is :

1972	1973	1974	1975	1976
83.33	90.00	100	125	175

2. Splicing : The technique of splicing consists in combining two or more overlapping series of index numbers to obtain a single continuous series. This continuity of the series of index number is required to facilitate comparisons.

Let us suppose that we have a series of index numbers with some base period, say, 'a' and it is discontinued in the period 'b' and with period 'b' as base, a second series of index numbers (with the same items) is constructed by the same method. The two series are put together or spliced together to get a continuous series.

The spliced index number may be obtained as follows :

$$\text{Spliced index no.} = \frac{\text{Index no. of current year} \times \text{old index no. of new base year}}{100}$$

$$= \frac{\text{Old index no. of new base year}}{100} \times \text{Index no. of current year} \quad \ldots (31)$$

Example : Two series of index numbers, series A with 1974 as base and series B with 1980 as base are given below : (1) splice the index B to index A (2) splice the index A to index B

Splicing				
Year				
	Series A Index (1974 = 100)	Series B Index (1980 = 100)	Series B spliced to Series A (1994 = 100)	Series A Spliced to Series B (1980 = 100)
1974	100		100	$\frac{100}{400}\times 100 = 25$
1975	120		120	$\frac{100}{400}\times 120 = 30$
1976	150		150	$\frac{100}{400}\times 150 = 37.5$
1977	200		200	$\frac{100}{400}\times 200 = 50$
1978	300		300	$\frac{100}{400}\times 300 = 75$
1979	350		350	$\frac{100}{400}\times 350 = 87.5$
1980	400	100	400	100
1981		115	$\frac{400}{100}\times 115 = 460$	115
1982		90	$\frac{400}{100}\times 90 = 360$	90
1983		95	$\frac{400}{100}\times 95 = 380$	95

Splicing				
Year				
	Series A Index (1974 = 100)	Series B Index (1980 = 100)	Series B spliced to Series A (1994 = 100)	Series A Spliced to Series B (1980 = 100)
1984		102	$\frac{400}{100}\times 102 = 408$	102
1985		110	$\frac{400}{100}\times 110 = 440$	110
1986		98	$\frac{400}{100}\times 98 = 392$	98

Note : In 4th column we have done forward splicing. In the last column we have done backward splicing

3. Deflating of the price index numbers : Deflating means adjusting, correcting or reducing a value which is inflated. Hence by deflating of the price index numbers we mean adjusting them after making allowance for the effect of changing price levels.

The purchasing power is given by the reciprocal of the index number and the real income (or wages) are obtained by the formula :

$$\text{Real wages} = \frac{\text{Money or Nominal Wages}}{\text{Price Index}}\times 100 \qquad \ldots (32)$$

The real income is also known as deflated income.

Example : Given the following data :

Year		Consumar Price Index
	weekly take home pay (wages)	
1968	109.50	112.8
1969	112.20	118.2
1970	116.40	127.4
1971	125.08	138.2
1972	135.40	143.5
1973	138.10	149.8

(i) What was the real average weekly wage for each year?

(ii) In which year did the employes have the greatest buying power?

(iii) What percentage increase in the weekly wages for the year 1973 is required (if any) to provide the same buying power that the employees enjoyed in the year in which they had the highest real wages?

Calculations of Real Wags

Year	Weekly take home wages (Rs)	Consumer Price Index	Real wags (Rs)
(1)	(2)	(3)	(4) = $\frac{(2)}{(3)} \times 100$
1968	109.50	112.8	97.07
1969	112.20	118.2	94.92
1970	116.40	127.4	91.37
1971	125.08	138.2	90.51
1972	135.40	143.5	94.36
1973	138.10	149.8	92.19

(i) Real weekly wages (Rs) are given in the last column (4)

(ii) Since Real wages are the highest in 1968, the employees had the highest buying power in 1968.

(iii) In order that employees had the same buying power in 1973 as they enjoyed in the year 1968, there should be an increase of Rs. 97.07 – Rs. 92.19 = Rs. 4.88 in their weekly wages. Hence the required percentage increase in their weekly wages is

$$\frac{4.88}{92.19} \times 100 = 5.29$$

5.8 Cost of Living Index Number

Cost of living index numbers also termed as 'Consumer Price Index Numbers' or 'Retail Price Index Numbers' are designed to measure the effects of changes in the prices of a basket of goods and services on the purchasing power of a particular section or class of society during any given (current) period with respect to some fixed (base) period. They reflect upon

the average increase in the cost of the commodities consumed by a class of people so that they can maintain the same standard of living in the current year as in the base year.

Such indices are helpful in wage negotiations and dearness allowance adjustments etc. The Govt. can make use of such indices in framing wage policy, price policy, rent control, taxation and general economic policies. Changes in the purchasing power of money and real income can be measured and markets of particulars kinds of goods and services can be analysed with the help of these indices.

Construction of cost of living index numbers

Cost of Living index numbers are constracted by the following methods :

(i) Aggregate Expenditrue method or Weighted Aggregate Method

In this method, the quantities consumed in the base year are used as weights. Thus,

$$\text{Cost of living index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \qquad \ldots (33)$$

which is nothing but Laspeyre's price index

(ii) Family Budget Method or Method of Weighted Relatives :

In this method the cost of living index is obtained on taking the weighted average of price relatives, the weights being the values of the quantities consumed in the base year. Thus,

$$I = \text{Price Relative} = \frac{p_1}{p_0} \times 100 \text{ and } w = p_0 q_0$$

$$\text{Then, cost of living Index} = \frac{\sum WI}{\sum W} \qquad \ldots (34)$$

Substituting the values of W and I we get,

$$\text{Cost of Living Index} = \frac{\sum p_0 q_0 \left(\frac{p_1}{p_0} \times 100 \right)}{\sum p_0 q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Which is the same Laspeyre's Index as we found in equation(33). So the cost of Living index numbers obtained from these two methods are the same.

Remark : Purchasing power of a rupee in the current period as compared to the base period is given by parchasing power of Rupee $= \frac{100}{\text{Cost of Living Index Number}}$ ... (35)

5.8.1 Steps in the Construction of Cost of Living Index (CLI) or, Consumer Price Index (CPI)

Cost of living index numbes are special-purpose index numbers measuring the relative change in the cost for maintaining similar standard of living of a group of people in two different situations. It measures the relative change in the amount of money required for equal consumption under two different situations. It represents the average changes in prices over time paid by the ultimate consumer for specified goods and services. Hence it is also called "consumer price index number". Generally the consumption pattern varies with the class of people and the geographical area. Hence, cost of living index must always relate to a specified class of people and specified geographical area.

The steps in the construction of a cost of living index are as follows :

(1) We are to decide on the class of people for whom the index number is intended.

(2) The next step is to conduct a "family budget enquiry" in the base period relating to the class of people concerned by the method of random sampling.

(3) The items of expenditure are classified in certain main groups–food, clothing, fuel and light, housing and miscellaneous. These groups are further sub-divided into smaller groups and sub-groups so that the items are individually mentioned.

(4) Retail prices of the items at regular time intervals should be collected from important local markets.

(5) For each item, there will be a number of price quotations covering different qualities and markets. We are to take the simple average of them as the price relative for the particular year.

(6) A separate index number is then computed for each group by the method of weighted average of price relatives.

The weight given is the percentage of expenditure on an item in relation to the total expenditure in the group, as obtained from the family budget data.

(7) The weighted, average of the group index numbers gives the final cost of living index number. Here the weight of a group index is the percentage of total expenditure on that group, as obtained from the family budget data.

(8) The cost of living index numbers are generally constructed for each week. The average of the weekly index numbers is taken as the index number for a month. The average of monthly index numbers gives the cost of living index for the whole year.

5.8.2 Uses of cost of Living Index Number

(i) Cost of living index (CLI) numbers are primarily used for the calculation of deamess allowance (DA) so that the same standard of living as in the base year can be maintained.

(ii) The reciprocal of the cost of living index may be used to measure that purchasing power of money.

(iii) Cost of living index numbers are also used to find the real wages by the method of deflation.

5.9 Uses of Index Number

An index number is used to measure the average change in a set of related variables over two different time periods or two different places. The most commonly used index numbers relate to prices of seleted group of commdities, volume of production in different sectors of an industry, quantum of exports and imports of different commodities and business activities. Price index numbers are used for various purposes. A wholesale price index number is used to measure the change in the general price level of a country, to reveal the flactuations in the purchasing power of money, and to study the general economic and business conditions of the country. A cost of living index number primarily serves as a measure of change in the retail prices of a specified set of goods and services representing the consumption level of the given group of people. Such indices also help in wage negotiations, for adjustment of dearness allowance, for determining real income, and in framing policies relating to wage, price, rent control, taxation, etc. Index number of stock prices are used by economists, bankers, speculators for different purposes.

Index number of industrial production presents the position in productivity in the current period relative to the base period. Similarly, index number of business activity reveals the progress in business conditions.

5.10 Bias in Index Number

In a bivariate data relating to the variables x and y, the weighted coefficient of correlation is,

$$r_{xy} = \frac{\text{cov}(x,y)}{s_x s_y} \text{ where } \text{cov}(x, y) = \frac{\sum fxy}{\sum f} - \frac{\sum fx}{\sum f} \cdot \frac{\sum fy}{\sum f}$$

Let x = price relative = $\frac{p_n}{p_0}$, y = quantity relative = $\frac{q_n}{q_0}$ and f = value in the base period = p_0q_0.

$$\text{So, } \operatorname{cov}(x, y) = \frac{\sum p_0 q_0 \cdot \frac{p_n}{p_0} \cdot \frac{q_n}{q_0}}{\sum p_0 q_0} - \frac{\sum p_0 q_0 \cdot \frac{p_n}{p_0}}{\sum p_0 q_0} \cdot \frac{\sum p_0 q_0 \cdot \frac{q_n}{q_0}}{\sum p_0 q_0}$$

$$= \frac{\sum p_n q_n}{\sum p_0 q_0} - \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0}$$

$$= \frac{\sum p_n q_n}{\sum p_0 q_n} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0} - \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0} = P_p L_q - L_p . L_q = L_q (P_p - L_p)$$

Now, from the law of demand we know that there is an inverse relation between price and quantity demanded, ceteris paribus. So, the correlation between price and quantity relatives is negative, $r_{xy} < 0$. It implies, $\operatorname{cov}(x, y) < 0$, or, $L_q(P_p - L_p) < 0$.

As $L_q > 0$, $P_p - L_p < 0$, or $P_p < L_p$.

Thus, Laspeyre's price index has an upward bias than the Paasche's price index. In other words, P_p has a downward bias than L_p.

5.11 Some Worked out Examples

Example : What is the difference between Laspeyre's and Paasche's systems of weights in compiling a price index? Calculate both Laspeyre's and Paasche's aggregative price indices for the year 1960 from the following data.

Commodities	Quantities		Prices per unit	
	1959	1960	1959	1960
A	3	5	2.0	2.5
B	4	6	2.5	3.0
C	2	3	3.0	2.5
D	1	2	1.0	0.75

Ans. In Laspeyre's price index, base year quantities are taken as weights and in Paasche's price index current year quantities are taken as weights.

Calculation of Laspeyre's and Paasche's Price Indices

Commodities	1959 (q_0)	1960 (q_1)	1959 (p_0)	1960 (p_1)	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	3	5	2.0	2.5	6.0	10	7.5	12.5
B	4	6	2.5	3.0	10.0	15	12.0	18.0
C	2	3	3.0	2.5	6.0	9	5.0	7.5
D	1	2	1.0	0.75	1.0	2	0.75	1.5

Total = $\sum p_0q_0 = 23, \sum p_0q_1 = 36, \sum p_1q_0 = 25.25, \sum p_1q_1 = 39.5$

Laspeyre's Price Index for 1960 (1959 = 100)

$$p_{01}^{La} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{25.25}{23.00} \times 100 = 1.0978 \times 100 = 109.78$$

Paasche's Price Index for 1960 (1959 = 100)

$$p_{01}^{pa} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{39.50}{36.00} \times 100 = 1.0972 \times 100 = 109.72$$

Example : (a) Using paasche's formula compute price and quantity index no for 1977

Commodity	Quantity		Value	
	1966	1970	1966	1970
A	100	150	500	900
B	80	100	320	500
C	60	72	150	360
D	30	33	360	297

(b) For the above problem also compute price index by

(i) Marshall – Edgeworth formula

(ii) Fisher's Formula

(iii) Dorbish – Bowley Formula

(iv) Walsch Formula

Computation of Indices by different formulae

Commodities	q_0	q_1	p_0q_0	p_1q_1	p_0	p_1	p_1q_0	p_0q_1
A	100	150	500	900	5.0	6	600	750
B	80	100	320	500	4.0	5	400	400
C	60	72	150	360	2.5	5	300	180
D	30	33	360	297	12.0	9	270	396
Total			1330	2057			1570	1726

(a) $P_{01}^{pa} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{2057}{1726} \times 100 = 1.192 \times 100 = 119.2$

$q_{01}^{pa} = \frac{\sum p_1q_1}{\sum p_1q_0} \times 100 = \frac{2057}{1570} \times 100 = 1.31019 \times 100 = 131.019$

(b) (i) $P_{01}^{ME} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100 = \left(\frac{2057+1570}{1726+1330}\right) \times 100$

$= \frac{3627}{3056} \times 100 = 1.1868 \times 100 = 118.68$

(ii) $P_{01}^{F} = \left[\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}\right]^{\frac{1}{2}} \times 100 = \sqrt{\left(\frac{1570}{1330} \times \frac{2057}{1726}\right)} \times 100$

$= \sqrt{(1.1804 \times 1.192)} \times 100 = \sqrt{118.04 \times 119.2)} = 118.62$

(iii) $P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1q_0}{\sum p_0q_0} + \frac{\sum p_1q_1}{\sum p_0q_1}\right] \times 100 = \frac{1}{2}\left[\frac{1570}{1330} + \frac{2057}{1726}\right] \times 100$

$= \frac{1}{2}(1.18045 + 1.192) \times 100 = \frac{237.245}{2} = 118.6225$

Computation of Walsch Price Index

Commodity	q_1q_0	$(q_1q_0)^{\frac{1}{2}}$	$p_1(q_1q_0)^{\frac{1}{2}}$	$p_0(q_1q_0)^{\frac{1}{2}}$
A	15000	122.47	734.82	612.35
B	8000	89.44	447.20	357.76
C	4320	65.73	328.65	164.33
D	990	31.46	283.14	377.52
Total	28310		1793.81	1511.96

$$P_{01}^{W} = \frac{\sum p_1.(q_0q_1)^{\frac{1}{2}}}{\sum p_0.(q_0q_1)^{\frac{1}{2}}} \times 100 = \frac{1793.81}{1511.96} \times 100 = 118.64$$

Example : Find out Fisher's Index from the following table and prove that this index satisfies time reversal test.

	1975 (Base year)		1975 (Base year)	
Commodity	Price	Value	Price	Value
A	4	16	6	12
B	6	24	4	32
C	8	40	10	30
D	10	50	15	45

We know value = Price × Quantity

Calculations for Fisher's Price Index

Commodity	Price	Value	Price	Value	q_0	q_1	p_0q_1	p_1q_0
	(p_0)	(p_0q_0)	(p_1)	(p_1q_1)				
(1)	(2)	(3)	(4)	(5)	(3)/(2)	(5)/(4)		
A	4	16	6	12	4	2	8	24
B	6	24	4	32	4	8	48	16
C	8	40	10	30	5	3	24	50
D	10	50	15	45	5	3	30	75
Total		$\sum p_0q_0$ = 130		$\sum p_1q_1$ = 119			$\sum p_0q_1$ = 110	$\sum p_1q_0$ = 165

Fisher's Price Index (without the factor 100)

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{\frac{165 \times 119}{130 \times 110}} = \sqrt{\frac{19635}{14300}}$$

$$P_{10}^{F} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{\frac{110 \times 130}{119 \times 165}} = \sqrt{\frac{14300}{19635}}$$

$$P_{01}^{F} \times P_{10}^{F} = \sqrt{\frac{19635}{14300} \times \frac{14300}{19635}} = 1$$

Hence Fisher's Price Index Satifies
Time Reversal test,

Example : Compute Fisher's Index number from the following data

	(Base year)		(Current year)	
Commodity	Price	Expenditure	Price	Expenditure
A	5	25	10	60
B	1	10	2	24
C	4	16	8	40
D	2	40	5	75

Apply Factor Reversal Test to the above Index
We know Expenditure = Price × Quantity

Calculation of Fisher's Indices

	(Base year)		(Current year)	
Commodity (1)	Price (2)	Expenditure (3)	Price (4)	Expenditure (5)
A	5	25	10	60
B	1	10	2	24
C	4	16	8	40
D	2	40	5	75
		$\sum p_0 q_0 = 91$		$\sum p_1 q_1 = 199$

	q_0	q_1	p_0q_1	p_1q_0
	= (3)/(2) (6)	= (5)/(4) (7)	(8)	(9)
	5	6	30	5
	10	12	12	20
	4	5	20	32
	20	15	30	100
			$\sum p_0q_1 = 92$	$\sum p_1q_0 = 202$

Fisher's Price Index

$$p_{01}^F = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} = \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100$$

$$= \sqrt{\frac{40198}{8372}} \times 100 = \sqrt{4.8015} \times 100 = 2.1912 \times 100 = 219.12$$

Fisher's Quantity Index

$$Q_{01}^F = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum p_1q_1}{\sum p_1q_0}} = \sqrt{\frac{92}{91} \times \frac{199}{202}}$$

Hence, the factor reversal test demands that $p_{01}^F \times Q_{01}^F = V_{01}$

$$p_{01}^F \times Q_{01}^F = \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{91} \times \frac{199}{202}} = \frac{199}{91} = \frac{\sum p_1q_1}{\sum p_0q_0} = V_{01} \text{ (proved)}$$

$\therefore$ Fisher's Ideal Index No. satisfies both Time and Factor Reversal Test.

Example : Computation of consumer price index number for 1996 (1995 = 100) by Aggregative expenditure method

Commodity	Quantity	p_0 (1995)	p_1 (1996)	p_1q_0	p_0q_0
A	6	5.75	6.00	36.00	34.50
B	6	5.00	8.00	48.00	30.00
C	1	6.00	9.00	9.00	6.00
D	6	8.00	10.00	60.00	48.00
E	4	2.00	1.50	6.00	8.00
F	1	20.00	15.00	15.00	20.00
				$\sum p_1q_0 = 174$	$\sum p_0q_0 = 146.5$

Consumer Price Index = $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{174}{146.5} \times 100 = 118.77$

CPI by the Family Budget Method

Commodity	Quantity	p_0 (1995)	p_1 (1996)	$\left(\frac{p_1}{p_0} \times 100\right) = p$	$p_0q_0 (= V)$	pV
A	6	5.75	6.0	104.34	34.5	3600
B	6	5.00	8.0	160.00	30.0	4800
C	1	6.00	9.0	150.00	6.0	900
D	6	8.00	10.0	125.00	48.0	6000
E	4	2.00	1.5	75.00	8.0	600
F	1	20.00	15.0	75.00	20.0	1500
				$\sum pV = 17400$		$\sum V = 146.5$

CPI = $\frac{\sum pV}{\sum V} = \frac{17400}{146.5} = 118.77$

$\therefore$ Both the methods produce the same result.

5.12 Summary

The following are the important Formulae used for the construction of Index Numbers.

Simple Aggregative Index $P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$

Simple Average of Price Relative Index

$$P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{N}$$

Weighted Aggregative Indices

Laspeyre's Index $P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

Paasche's Index $P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$

Dorbish-Bowley's Index $P_{01}^{DB} = \frac{L + P}{2}$

Fisher's Ideal Index $P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$

Marchall-Edgeworth's Index $P_{01}^{ME} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} \times 100$

Kelly's Index $P_{01}^{k} = \frac{\sum p_1 q}{\sum p_0 q}$

Walsch's Price Index $P_{01}^{w} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$

Weighted Average of Price Relative Method $P_{01} = \frac{\sum WP}{\sum W}$

Where, 'W' stands for the value of commodity consumed.
Consumer Price Index

(1) $\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

(2) $\frac{\sum WI}{\sum W}$, Where I = $\frac{p_1}{p_0} \times 100$ and W = $p_0 q_0$

Time Reversal, Factor Reversal and Circular Tests

Time Reversal Test in satisfied when $p_{01} \times p_{10} = 1$
[i.e., when base period(0) and current period(i) are reversed]

Factor Reversal Test is satisfied when $P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$ and Circular Test is satisfied when $p_{01} \times p_{12} \times p_{20} = 1$

5.13 Questions

A. Choose the correct answer

(i) Fisher's ideal index is

(a) AM of Laspeyre's and Paasche's index

(b) Median of Laspeyre's and Paasche's index

(c) GM of Laspeyre's and Paasche's index

(d) None of the above

(ii) Factor Reversal test is satisfied by

(a) Laspeyre's index (b) Paasche's index

(c) both (a) and (b) (d) Fisher's index

(iii) Fixed-base index and chain-base index are equal when the formula satisfies

(a) circular test (b) Time reversal test

(c) Factor reversal test (d) none of the above

(iv) Laspeyre's index is based on

(a) current year quantities (b) base year quantities

(c) average of current and base year quantities

(d) none of the above

(v) If the consumer price index for middle class people in Delhi in 2010 with 2000 as base period is 225, then the retail prices have increased on the average

(a) 225% (b) 25% (c) 125% (d) 22.5%

(vi) Constructing one continuous series from two different index number series with a common base is

(a) deflating (b) base shifting

(c) splicing (d) none of the above

(vii) Time revesal test is satisfied by

(a) Laspeyre's index (b) Simple GM of price relatives

(c) Marshall-Edgeworth index (d) both (b) and (c)

(viii) If the prices of all items change in the same ratio, then

(a) Laspeyre's index = Paasche's index

(b) Laspeyre's index < Paasche's index

(c) Laspeyre's index > Paasche's index

(d) None of the above

(ix) The best average in construction of index number is

(a) AM (b) GM (c) HM (d) Median

(x) Weighted HM of price relatives using current year value as weights is

(a) Paasche's index (b) Fisher's index

(c) Laspeyre's index (d) Bowley's index

B. True or False

(i) Marshall - Edgeworth price index number lies between Laspeyre's and Paasche's price index numbers.

(ii) Factor reversal test is satisfied only by Fisher's index

(iii) Laspeyre's price index is based on current year quantities

(iv) Weighted AM of price relatives with base period values as weights give Laspeyre's formula

(v) Chain base index numbers are easier to calculate than fixed base index numbers.

C. Fill in the blanks :

(i) Link indices are successively multiplied to obtain _____ indices.

(ii) Fisher's index is called _____ index

(iii) Purchasing power of money is _____ proportional to the price index

(iv) Cost of living index is also known as _____ price index

(v) Bowley's index is the _____ mean of Laspeyre's and Paasche's Index

1. What is an Index number? What are its uses?
2. Discuss the problems in the construction of index number
3. Show that Laspeyre's and Paasche's price index numbers can be obtained by the method of averaging price-relatives as well as by aggregative method.
4. What are the tests proposed by Fisher for checking the goodness of an Index Number?

   Do the Laspeyre's and Paasche's price index numbers satisfy these tests? Mention an index number that satisfies them.
5. Show that Marshall -Edgeworth price index number is weighted Arithmatic Mean of price relatives.
6. What is meant by cost of living index number? How it is constructed?
7. What is meant by

   (a) Base shifting (b) Splicing and (c) Deflating of index numbers?
8. From the following data calculate

   (i) Laspeyre's (ii) Paasche's

   (iii) Marshall-Edgeworth's and

   (iv) Fisher's Price index numbers for 1986

	(1976 = 100)		(1986)	
Commodity	Quantity	Money value	Price	Money value
A	10	40	5.50	66
B	6	18	4.40	22
C	5	75	18.20	91
D	8	48	7.60	76

5.14 References

1. Allen, R.G.D. (1949) Statistics for Economics, Hutchinson and co.
2. Croxton, Cowden and Klein (1967) Applied General Statistics, Prentice Hall
3. Yule and Kendall (1968) An Introduction to the Theory of statistics, London C. Griffin and Co. Ltd.

Unit 6 ❐ Introduction to the Theory of Probability and Distribution

Structure

6.1 Objectives

6.2 Introduction

6.3 Terminology

6.4 Concept of Probability

6.4.1 Set theoretical Notations and Terminology

6.4.2 Conditional Probability

6.4.3 Theorem of Compound Probability

6.4.4 Bayes' Theorem

6.4.5 Statistical or Empirical definition of probability

6.4.6 Axiomatic definition of probability

6.5 Random Variable

6.5.1 Discrete and continuous random variable

6.5.2 Probability distribution of a random variable

6.5.3 Mathematical Expectation

6.5.4 Joint and Margianl probability distribution

6.6 Probability Distribution–Discrete and continuous

6.6.1 Binomial distribution

6.6.2 Poisson distribution

6.6.3 Normal distribution

6.6.4 Relationship between Binomial and Normal distribution

6.6.5 Moments of Normal Distribution

6.6.6 Mode of Normal Distribution

6.6.7 Median of Normal Distribution

6.6.8 Points of Inflexion of Normal curve

6.6.9 Properties of Normal Distribution

6.6.10 Fitting a Normal Distribution to an observed Distribution

6.6.11 Importance of Normal Distribution

6.6.12 Some worked out examples on Probability Distribution

6.7 Moment Generating Function

6.7.1 MGF of Binomial Distribution

6.7.2 MGF of Poisson Distribution

6.7.3 MGF of Normal Distribution

6.8 Summary

6.9 Questions

6.10 References

6.1 Objectives

Theory of Probability is the foundation of research Methodology. It is one of the major building blocks of statistical inference which is at the heart of research methodology. Statistical inference refers to the process of selecting and using a sample statistic to draw inferences about a population parameter. It is concerned with using probability concept to deal with uncertainly in decision making. Statistical inference treats two different classes of problems viz., 'Hypothesis testing' and 'Estimation'. Hypothesis testing is to test some hypothesis about the parent population from which the sample is drawn.

Let us assume that the purchase manager of a machine tool making company has to decide whether to buy castings from a new supplier or not. The new supplier claims that his castings have higher hardness than those of his competitors. If the claim is true, then it would be worth while to switch over from the existing suppliers to the new one. However, if the claim is not true, the purchase manager should continue to buy from the existing suppliers. In such a probabilistic dilemma, testing of hypothesis provides such a tool to the decision maker which helps him arrive at convincingly near-accurate solution. This tesing of hypothesis technique is totally based on the theory of probability.

6.2 Introduction

If an experiment is performed repeatedly under essentially homogeneous and similar conditions, the result (outcome) may be classified as follows :

(i) Unique or certain

(ii) not definite but may be one of the possibilities depending on the experiment.

The phenomenon under (i) is deterministic or predictable phenomenon. Most of the outcomes of the experiments under physical and chemical sciences are of this nature. However, phenomena under (ii) are unpredictable or probabilistic phenomena. These are observed in economics, business and social sciences. Even in our day to day life we find such unpredictable phenomena.

(a) In case of a new born baby, the sex can not be predicted with certainty.

(b) A sales manager can not say with certainty if he will achieve the sales target.

(c) If the electric bulb has lasted for 3 months, nothing can be said about its future life

(d) While tossing a uniform coin we are not sure if we shall get head or tail.

In all these cases there is an element of uncertainty or chance. A numerical measure of uncertainty is provided by a very important branch of statistics called the 'theory of probability'.

The theory of probability has its origin in the games of chance related to gambling, for instance, throwing of dice or coin, drawing of cards from a pack of cards and so on. The main contributors to the development of theory of probability and its applications are : French mathmatician Blaise Pascal and Pierre de Format, Swiss mathematician James Bernoulli, De Moivre (French), Thomas Bays (British), Pierre-Simon de Laplace (French), R.A. Fisher (British), Von Mises (Austrian) and Russian mathematicians chebychev, A. Markov and A.N. Kolmogorov.

Today, the subject has been developed to a great extent and there is not even a single discipline in social, physical or natural sciences where probability theory is not used. It is extensively used in the quantitative analysis of business and economic problems and forms the basis of the 'Decision Theory'.

6.3 Terminology

There are three approaches to probability

(a) Classical approach

(b) Empirical approach

(c) Axiomatic approach

In this section we shall define various terms that are used in the definition of probability under different approaches.

Random Experiment : An experiment is called random experiment if it is conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the various possible outcomes.

Trial and Event : Performing a random experiment is called a trial and outcome or combination of outcomes are termed as events.

Example : If a coin is tossed repeatedly, the result is not unique. We may get either head or tail. Thus tossing of a coin is a random experiment or trial and getting of a head or tail is an event.

Exhaustive cases : The total number of possible outcomes of a random experiment is called the exhaustive cases for the experiment.

Thus, while tossing a single coin, we get head (H) or tail (T). Hence exhaustive number of cases is 2^1 viz., (H, T).

If two coins are tossed, the various possibilities are HH, HT, TH, TT where, HT means head on the first coin and tail on second coin and TH means tail on the first coin and head on the second. Thus in the case of tossing of two coins, exhaustive number of cases is $2^2 = 4$.

Similarly, in a toss of three coins, the possible number of outcomes is

$(H, T) \times (H, T) \times (H, T)$

$= (HH, HT, TH, TT) \times (H, T)$

$= (HHH, HTH, THH, TTH, HHT, HTT, THT, TTT)$

Therefore, in the case of tossing of 3 coints the exhaustive number of cases is $2^3 = 8$.

In general, is a throw of n coins, the exhaustive number of cases is 2^n.

In a throw of a die, the exhaustive number of outcomes is 6. We can get any one of the six faces marked 1, 2, 3, 4, 5 or 6. If two dice are rolled the possible outcomes are

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

i.e., 36 ordered pairs where pair (i, j) means number i on the first die and j on the second die, i and j both taking the values from 1 to 6. Hence, in the case of a throw of two dice exhaustive number of cases is $36 = 6^2$. Thus for a throw of 3 dice, exhaustive number of cases will be $216 = 6^3$, and for n dice they will be $6^{n.}$

If r cards are drawn from a pack of n cards, the exhaustive number of cases is $n_{c_r} = \binom{n}{r}$,

since r cards can be drawn out of n cards in $\binom{n}{r}$ ways.

Favourable Cases or Events :

The outcome of a random experiment is called an Event. For example (i) in a toss of two coins, the number of cases favourable to the event 'exactly one head' is 2, viz., HT, TH and for getting 'two heads' is one viz., HH.

(ii) In drawing a card from a pack of cards, the cases favourable to 'getting a diamond' are 13 and to 'getting an ace of spade' is only one.

Mutually Exclusive Events or Cases :

Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiment. For example, in toss of a coin the events, 'head' and 'tail' are mutually exclusive because if head comes, we can't get tail and if tail comes we can't get head.

Similarly, in the throw of a die, the six faces numbered, 1, 2, 3, 4, 5 and 6 are mutually exclusive. Thus events are said to be mutually exclusive if no two or more of them can happen simultaneously.

Equally Likely cases : The outcomes are said to be equally likely or equally probable if none of them is expected to occur is preference to other. Thus in tossing of a coin (die), all the outcomes viz., H, T (the faces 1, 2, 3, 4, 5, 6) are equally likely if the coin (die) is unbiased.

Independent Events : Events are said to be independent of each other if happening of any one of them is not afected by and does not affect the happening of any one of others. For example : (i) In, tossing of a die repeatedly the event of getting '5' in 1st throw is independent of getting '5' in second, third or subsequent throws.

(ii) In drawing cards from a pack of cards, the result of the second draw will depend upon the card drawn in the first draw. However, if the card drawn in the first draw is replaced before drawing the second card, then the result of second draw will be independent of the 1st draw.

Permutation and Combination : The word permutation in simple language means 'arrangement' and the word combination means 'group' or 'selection'. Let us consider three letters A, B and C.

The permutation of these three letters taken two at a time will be AB, BC, CA, BA, CB, AC i.e., 6 in all whereas the combination of three letters taken two at a time will be AB, BC, CA i.e., 3 in all. It should be noted that in combinations, the order of the elements (letters in this case) is immaterial i.e., AB and BA form the same combination but these are different arrangements.

Permutation : A permutation of n different objects taken r at a time, denoted by n_{p_r}, is an ordered arrangement of only r objects out of the n objects. Some important results are given below on permutation in the form of theorems.

Theorem 1 : The number of different permutations of n different objects taken r at a time without repetition is

$$n_{p_r} = n(n-1)(n-2) \ldots (n-r+1)$$

$$= \frac{n!}{(n-r)!} \qquad \ldots (1)$$

where, $n! = n(n-1)(n-2) \ldots 3.2.1$ ... (2)

eg., $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

By convention we take $0! = 1$.

In particular, the total number of permutations of n distinct objects, taken all at a time is given by : (Take r = n in (1))

$$n_{p_n} = n(n-1)(n-2) \ldots 3.2.1 = n! \qquad \ldots (3)$$

Theorem 2 : The number of different permutations of n different (distinct) objects, taken r at a time with repetition is

$$n_{p_r} = n^r \qquad \ldots (4)$$

In particular, $np_n = n^n$

Theorem 3 : The number of permutations of n different objects all at a time round a circle is $(n-1)!$

Theorem 4 : (Permutation of objects not all distinct)

The number of permutations of n objects taken all at a time, when n_1 objects are alike of one kind, n_2 objects are alike of second kind, ... n_k objects are alike of kth kind is given by

$$\frac{n!}{n_1!n_2!\ldots n_k!} \qquad \ldots (5)$$

For example, total number of arrangements in the letters of the word ALLAHABAD taken all at a time is given by :

$$\frac{9!}{4!2!} = \frac{9\times8\times7\times6\times5\times4\times3\times2\times1}{4\times3\times2\times1\times2\times1} = \frac{9\times8\times7\times6\times5}{2} = 7560$$

because is this word, there are 9 letters out of which 4 are of one kind i.e., A, 2 are of 2nd kind, i.e., L and rest are all different occurring once and 1! = 1.

Theorem 5 : (Fundamental Rule of Counting)

If one operation can be performed in p different ways and another operation can be

performed in q different ways, then the two operation when associated together can be performed in p × q ways.

The result can be generalised in more than two operations. For example, if there are 5 routes of journey from place A to place B, then the total number of ways of making a return journey (i.e., going from A to B and then coming back from B to A are 5 × 5 = 25 ways, since one can go from A to B in 5 ways and come back from B to A in 5 ways and any one of the ways of going can be associated with any one of the ways of coming.

Combination : A combination of n different objects taken r at a time, denoted by n_{c_r} or $\binom{n}{r}$, is a selection of only r objects out of the n objects, without any regard to the order of arrangement.

Theorem 6 : The number of different combinations of n different objects taken r at a time, without repetition, is $n_{c_r} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$; $r \le n$ ... (6)

6.4 Concept of Probability

Mathematical or classical or 'a priori' probability

Definition : If a random experiment results in N exhaustive mutually exclusive and equally likely outcomes (cases) out of which m are favourable to the happening of an event A, then the probability of accurrence of A, usually denoted by P(A) is given by :

$$P(A) = \frac{\text{Favourable number of cases to A}}{\text{Exhaustive number of cases}} = \frac{m}{n} \quad \text{... (7)}$$

This definition was given by James Bernoulli.

Remarks :

1. Obviously, the number of cases favourable to the complementary event $\overline{A}$ i.e., non-happening of event A are (N – m).

$$\therefore\ P(\overline{A}) = \frac{\text{Favourable No. of cases to } \overline{A}}{\text{Exhaustive number of cases}} = \frac{N-m}{N} = 1-\frac{m}{n}$$

or, $P(\overline{A}) = 1 - P(A)$ or, $P(A) = 1 - P(\overline{A})$ ... (8)

or, $P(A) + P(\overline{A}) = 1$ ... (9)

2. Since m and N are non-negative integers, $P(A) \geq 0$. Further, since $m \leq N$, we have $P(A) \leq 1$

$\therefore$ For any event A, $0 \leq P(A) \leq 1$ ... (10)

3. The above definition enables us to obtain probability by logical reasoning prior to making any actual trials and hence it is also known as 'a priori' or theoretical or mathematical probability.

4. **Limitations :** The classical probability fails in the following situations :

(i) If N, the exhaustive number of outcomes of the random experiment is infinite or unknown

(ii) If the various outcomes of the random experiment are not equally likely. For example, if a person jumps from the top of Qutab Minar, then the probability of his survival will not be 50%, since in this case the two mutually exclusive and exhanstive outcomes, viz., survival and death are not equally likely.

Problem 1 : What are the probabilities of obtaining (i) an odd number and (ii) a multiple of 3 in the throw of a fair die?

Ans. There are 6 elementary events in the sample space corresponding to the number 1, 2, 3, 4, 5 and 6 on the upper most face of the die. Since the die is fair, these 6 elementary events are equally likely.

(i) Let E denote the event that an odd number of points is obtained, occurance of 1, 3 and 5 are favourable to E

$$\therefore P(E) = \frac{3}{6} = \frac{1}{2}$$

(ii) Of the 6 elementary events, 3 and 6 are favourable of obtaining 'a multiple of 3'

Hence the required probability is $\frac{2}{6} = \frac{1}{3}$

Problem 2 : If the letters of the word MOTHER are aranged at random, then find the probability that the vowels will be next to each other.

Ans. The letters can be mutually arranged in 6! ways. So the total number of elementary events is 6! which are equally likely, since the letters are arranged at random.

When the two vowels, viz., O and E are placed next to each other, then we may arrange this single entity (containing O and E) and the remaining letters among themselves in 5! different ways. In each of these 5! arrangements, the vowels can be mutually arranged in 2! ways. Thus the number of elementary events favourable to the given event is $5! \times 2!$.

So the required probability is $\frac{5! \times 2!}{6!} = \frac{2}{6} = \frac{1}{3}$

Problem 3 : If a fair coin is tossed thrice, find the probability that there are (a) at most one tail, (b) at least one head.

Ans. Hence, the sample space is [HHH, HHT, HTH, HTT, THH, THT, TTH, TTT] which has $2^3 = 8$ elementary events. Since the coin is fair, these elementary events are equally likely.

(a) Let E denote the event of getting at most one tail. Then there are 4 elementary events, viz., HHH, HHT, HTH, THH which are favourable to E.

So, P(E) = $\frac{4}{8} = \frac{1}{2}$

(b) Of the 8 elementary events, 7 are favourable to the event of getting at least one head.

Hence the required probability is $\frac{7}{8}$.

Alternative solution : Required probability = 1 – P(no head) = $1 - \frac{1}{8} = \frac{7}{8}$

Problem 4 : A club consisting of 15 married couples chooses a president and then a secretary by random selection, what is the probability that (i) both are men (ii) one is a man and the other is a women (iii) the president is a man and secretary is a woman?

Ans. Total number of elementary events is equal to the number of ways in which two positions can be occupied by 30 persons = ${}^{30}P_2 = 30 \times 29$

These elementary events are equally likely as the choice is made by random selection.

(i) The number of elementary events favourable to the event that both are men = ${}^{15}P_2 = 15 \times 14$

So the required probability = $\frac{15 \times 14}{30 \times 29} = \frac{7}{29}$

(ii) The number of favourable cases = ${}^{15}C_1 \times {}^{15}C_1 \times 2! = 15 \times 15 \times 2$, since the chosen man and the chosen woman can occupy the positions in 2! = 2 ways. Thus the required probability = $\frac{15 \times 15 \times 2}{30 \times 29} = \frac{15}{29}$

(iii) Since the president's post can be filled in ${}^{15}P_1 = 15$ ways and the secretary's post in

$^{15}P_1 = 15$ ways, the number of favourable elementary events in this case is 15×15.

So the required probability = $\frac{15 \times 15}{30 \times 29} = \frac{15}{58}$

Problem 5 : A box contains 7 white and 5 black balls. Three balls are drawn at random. Find the probability that they are all of the same colour when (i) the balls are drawn at a time, (ii) The balls are drawn one by one without replacement and (iii) the balls are drawn one by one with replacement.

Ans. (i) Total number of emelentary events is $^{12}C_3 = 220$, which are equally likely. Among them $^7C_3 + {}^5C_3 = 35 + 10 = 45$ are favourable to the event that all are of the same colour, since all may be white or all black. Then the required probability is

$$1 - P(\text{all are of the same colour}) = 1 - \frac{45}{220} = 1 - \frac{9}{44} = \frac{35}{44}$$

(ii) Total number of elementary events = $^{12}P_3 = 1320$, and among them $^7P_3 + {}^5P_3 = 210 + 60 = 270$ are favourable to the event that all are of the same colour. So the required probability is

$1 - \frac{270}{1320} = 1 - \frac{9}{44} = \frac{35}{44}$, same as in (i).

(iii) Here, total number of elementary events = $12^3 = 1728$ and among them $7^3 + 5^3 = 343 + 125 = 468$ are favourable to the event that all are of the same colour. So the required

probability = $1 - \frac{468}{1728} = 1 - \frac{13}{48} = \frac{35}{48}$

Problem 6 : 5 different letters are put at random inside 5 addressed envelopes. Find the probability of putting exactly 2 letters in the correct envelopes.

Ans. There are 5 letters ($L_1, L_2, L_3 \ldots L_5$) and 5 corresponding envelops ($E_1, E_2, E_3 \ldots E_5$).

So the total number of elementary events is $5! = 120$, which are equally likely as the letters are put at random.

The two letters are to be put correctly can by chosen from 5 letters in $^5C_2 = 10$ ways.

In each case, the chosen letters can be put in correct envelops in 1 way and other 3 letters wrongly in the remaining 3 envelopes in 2 ways. This is because, supposing that L_1 and L_2 are placed correctly in E_1 and E_2, other 3 letters can be placed wrongly as

E_3	E_4	E_5
L_4	L_5	L_3
L_5	L_3	L_4

So, the total number of elementary events that are favourable to the given event is $10 \times 2 = 20$.

Thus, the required probability = $\frac{20}{120} = \frac{1}{6}$

6.4.1 Set Theoretical Notations and Terminology

Union : $A \cup B$, read as 'A union B', with denote the occurrence of either A or B or both A and B;

Similarly, $\bigcup_{i=!}^{k} A_i = A_1 \cup A_2 \cup ... \cup A_k$ will denote the occurrence of at least one of the events $A_1, A_2, ..., A_k$

Intersection : $A \cap B$, read as 'A intersection B', will denote the occurrence of both A and B. Similarly, $\bigcap_{i=1}^{k} A_i = A_1 \cap A_2 \cap . \cap A_k$ will denote the simultaneous occurrence of the events $A_1, A_2, ... A_k$.

Complement : $\bar{A}$ read as 'A complement', will denote the non-occurrence of event **A**

Difference : A – B, read as 'A minus B' will denote the occurrence of event A together with the non-occurrence of event B. It follows that $A - B = A \cap \bar{B}$.

Example : In the throwing of a die, let A denote the appearance of at most 4 points and B the appearance of an odd number of points. Then in set theoretic notation, the sample space is

$S = \{1, 2, 3, 4, 5, 6\}$ and $A = \{1, 2, 3, 4\}$ and $B = \{1, 3, 5\}$

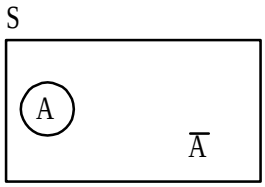
$\therefore A \cup B = \{1, 2, 3, 4, 5\}$

$A \cap B = \{1, 3\}$

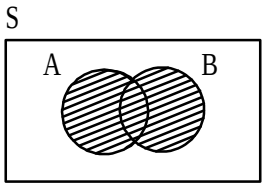
$\bar{A} = \{5, 6\}$ and $A - B = \{2, 4\}$

Under the operations of union, intersection and complement events satisfy various Laws, some of which are noted below.

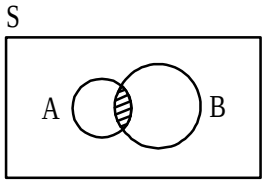
Venn Diagram



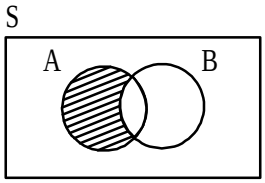
A and $\overline{A}$



$A \cup B$



$A \cap B$



$(A - B)$ or $(A \cap \overline{B})$

1. Commutative Laws :

$A \cup B = B \cup A$

$A \cap B = B \cap A$

2. Associative Laws :

$(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

3. Distributive Laws :

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

4. De Morgan's Laws :

$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$

$\overline{(A \cap B)} = \overline{A} \cup \overline{B})$

Theorem 7 : If $A_1, A_2 \ldots A_K$ are mutually exclusive events, then

$P(A_1 \cup A_2 \cup \ldots \cup A_K) = P(A_1) + P(A_2) + \ldots + P(A_K)$ This is known as **theorem of total probability**

Proof : Let the total number of elementary events of the random experiment be n, where n is finite and the elementary events are equally likely. Also let $n(A_i)$ of them be favourable to the event A_i, i = 1(1)K. Since the events are mutually exclusive, the elementary event that are favourable to any of the events are entirely different from those favourable to others. So the number of elementary events that are favourable to A_1 or A_2 or ... or A_K is $n(A_1) + n(A_2) + \ldots + n(A_K)$

Hence, $P(A_1 \cup A_2 \cup \ldots \cup A_K) = \dfrac{n(A_1) + n(A_2) + \ldots n(A_k)}{n}$

$$= \frac{n(A_1)}{n} + \frac{n(A_2)}{n} + \ldots + \frac{n(A_K)}{n}$$

$= P(A_1) + P(A_2) + \ldots + P(A_K)$ ... (11)

Corollary :

(i) When k = 2, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ ... (12)

(ii) Let the events $A_1, A_2 \ldots, A_k$ be exhanstive and mutually exclusive.

As the events are exhaustive $A_1 \cup A_2 \cup \ldots, \cup A_k$ is a sure event. Hence,

$P(A_1, \cup A_2 \cup \ldots \cup A_k) = 1$

Again, as the events are mutually exclusive,

$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$

Then combining the two results, we get

$P(A_1) + P(A_2) + \ldots + P(A_k) = 1$ ... (13)

In particular, the events A and $\bar{A}$ are exhaustive and mutually exclusive. Hence,

$P(A) + P(\bar{A}) = 1$ which implies that $P(A) = 1 - P(\bar{A})$ ... (14)

(iii) If $A_1, A_2, \ldots A_k$ be exhaustive forms of A and be mutually exclusive, then

$P(A) = P(A_1) + P(A_2) + \ldots + P(A_k)$

(iv) If the occurrence of event A implies the occurrence of event B, then B can occur in one of the mutually exclusive forms A and (B – A). Hence,

$P(B) = P(A) + P(B - A)$ or, $P(B - A) = P(B) - P(A)$ ... (15)

Problem 7 : An integer is chosen at random from 50 integers 1, 2, ..., 50. What is the probability that the selected integer is divisible by 7 or 10?

Ans. Let A_1 denote the selection of an integer divisible by 7 and A_2 denote the selection of an integer divisible by 10. Since none of the integer from 1 to 50 is a multiple of 7 as well as that of 10, A_1 and A_2 are mutually exclusive events. So the required probability is

$P(A_1 \cup A_2) = P(A_1) + P(A_2)$

Here, the total number of elementary events, all equally likely, is n = 50. The number of cases favourable to A_1 and A_2 are $n(A_1) = 7$ and $n(A_2) = 5$.

$$\therefore P(A_1) = \frac{n(A_1)}{n} = \frac{7}{50} \text{ and } P(A_2) = \frac{n(A_2)}{n} = \frac{5}{50}$$

$$\text{Hence, } P(A_1 \cup A_2) = P(A_1) + P(A_2) = \frac{7}{50} + \frac{5}{50} = \frac{12}{50} = \frac{6}{25}$$

Theorem 8 : Addition Theorem of Probability

Whatever be the events $A_1, A_2, \ldots, A_n$,

$P(A_1 \cup A_2 \cup \ldots \cup A_n)$

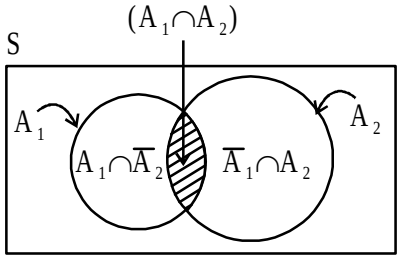
$$= \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j) + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} P(A_i \cap A_j \cap A_k) - \ldots$$

$$+(-1)^{n-1} P(A_1 \cup A_2 \cap \ldots \cap A_n)$$

Proof : Let us consider two events A_1 and A_2. The probability of occurrence of at least one of the two events A_1 and A_2 is givn by $P(A_1 \cup A_2)$. Let us suppose that a random experiment results in a sample space S with N sample points which are exhaustive.

Then by definition $P(A_1 \cup A_2) = \dfrac{n(A_1 \cup A_2)}{n(S)} = \dfrac{n(A_1 \cup A_2)}{N}$

where, $n(A_1 \cup A_2)$ is the number of occurrences favourable to the event $(A_1 \cup A_2)$,



From the venn diagram above, we get,

$$P(A_1 \cup A_2) = \frac{[n(A_1) - n(A_1 \cap A_2)] + n(A_1 \cap A_2) + [n(A_2) - n(A_1 \cap A_2)}{N}$$

$$= \frac{n(A_1) + n(A_2) - n(A_1 \cap A_2)}{N}$$

$$= \frac{n(A_1)}{N} + \frac{(n(A_2)}{N} - \frac{n(A_1 \cap A_2)}{N}$$

$$= P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad \ldots (16)$$

The theorem thus holds for n = 2. For 3 events A_1, A_2 and A_3

$P(A_1 \cup A_2 \cup A_3) = P[(A_1 \cup A_2) \cup A_3]$

$= P(A_1 \cup A_2) + P(A_3) - P[(A_1 \cup A_2) \cap A_3]$, by (16)

$= P(A_1) + P(A_2 - P(A_1 \cap A_2) + P(A_3) - P[A_1 \cup A_2) \cap A_3] \qquad \ldots (17)$

Now, $P[(A_1 \cap A_2) \cap A_3] = P[(A_1 \cap A_3) \cup (A_2 \cap A_3)$

$= P(A_1 \cap A_3) + P(A_2 \cap A_3) - P[(A_1 \cap A_3) \cap (A_2 \cap A_3)$

$= P(A_1 \cap A_3) + P(A_2 \cap A_3] - P(A_1 \cap A_2 \cap A_3)$ ... (18)

So, combining equations (17) and (18) we get,

$$P(A_1 \cup A_2 \cup A_3) = \sum_{i=1}^{3} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{3} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3)$$

Thus the theorem holds for n = 3 also.

Let us assume that the theorem is true for n = m. Then,

$$P(A_1 \cup A_2 \cup ... \cup A_m) = \sum_{i=1}^{m} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{m} P(A_i \cap A_j) + ...$$

$... + (-1)^{m-1} P(A_1 \cap A_2 \cap ... \cap A_m)$ ... (19)

Now, $P(A_1 \cup A_2 \cup ... \cup A_{m+1}) = P[(A_1 \cup A_2 \cup ... \cup A_m) \cup A_{m+1}]$

$= P(A_1 \cup A_2 \cup ... \cup A_m) + P(A_{m+1}) - P[(A_1 \cup A_2 \cup ... \cup A_m) \cap A_{m+1}]$

But, $P[(A_1 \cup A_2 \cup ... \cup A_m) \cap A_{m+1}]$ ... (20)

$= P[(A_1 \cap A_{m+1}) \cup (A_2 \cap A_{m+1}) \cup ... \cup (A_m \cap A_{m+1})]$

$$= \sum_{i=1}^{m} P(A_i \cap A_{m+1}) - \sum_{\substack{i,j=1 \\ i<j}}^{m} P(A_i \cap A_j \cap A_{m+1}) + ...$$

$+ (-1)^{m-1} P(A_1 \cap A_2 \cap ... \cap A_m \cap A_{m+1})$ ... (21)

Substituting the results of (19) and (21) in (20) we get, after rearrangement of terms,

$P(A_1 \cup A_2 \cup ... \cup A_{m+1})$

$$= \sum_{i=1}^{m+1} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{m+1} P(A_i \cap A_j) + ... + (-1)^m P(A_1 \cap A_2 \cap ... \cap A_{m+1})$$

Thus the theorem holds for n = m + 1 if it holds for n = m. But we have seen it to be true for n = 2 and n = 3. Being true for n = 3, it is true for n = 4, and so on for all positive integral values of n.

Remarks :

1. For any two events A_1 and A_2 we have,

$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$,

Since $P(A_1 \cap A_2) \geq 0$

Using this result we may write $P(A_1 \cup A_2 \cup A_3) = P[(A_1 \cup A_2) \cup A_3] \leq P(A_1 \cup A_2)$

$+ P(A_3) \leq P(A_1) + P(A_2) + P(A_3)$

Proceeding in this way, it may be shown that,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \qquad \ldots (22)$$

This result is known as **Boole's Inequality**.

2. Probability of an event cannot exceed 1

$P(A_1 \cup A_2) \leq 1$

or, $P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq 1$

$\therefore\ P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$

Using this result, we have

$P(A_1 \cap A_2 \cap A_3) = P[(A_1 \cap A_2) \cap A_3] \geq P(A_1 \cap A_2) + P(A_3) - 1$

$\geq P(A_1) + P(A_2) - 1 + P(A_3) - 1$

$= P(A_1) + P(A_2) + P(A_3) - 2$

Proceeding in similar way, it may be shown that

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1) \qquad \ldots (23)$$

This result is known as **Bonferroni's Inequality**.

Problem 8 : Two newspaper N_1 and N_2 are published in Kolkata. It is found from a survey that 28% read N_1, 21% read N_2 and 6% read both the newspapers. Find the probability that a person randomly selected (i) does not read any of the two newspapers (ii) read only N_2.

Let A and B denote that the selected person reads N_1 and N_2 respectively. Since the person is randomly selected, $P(A) = 0.28$, $P(B) = 0.21$, $P(A \cap B) = 0.06$

(i) The probability that the person chosen does not read any of the newspapers is

$P(\bar{A} \cap \bar{B}) = (\overline{A \cap B}) = 1 - P(A \cup B)$

$= 1 - P(A) - P(B) + P(A \cap B)$

$= 1 - 0.28 - 0.21 + 0.06$

$= 0.57$

(ii) The probability that the selected person reads only N_2 is

$P(\bar{A} \cap B) = P(B) - P(A \cap B) = 0.21 - 0.06 = 0.15$

6.4.2 Conditional Probability

If A and B are two dependent events in sample space, then the conditional probability of A given B ie. probability of occuring of A on the assumption that B has already been occurred is defined as

$$P(A/B) = \frac{n(A \cap B)}{n(B)} \text{ provided } n(B) \neq 0 \qquad \ldots (24)$$

Similarly, the conditional probability of B given A is defined as

$$P(B/A) = \frac{n(B \cap A)}{n(A)}, \text{ provided } n(A) \neq 0 \qquad \ldots (25)$$

6.4.3 Theorem of compound probability

Theorem 9 : Theorem of compound probability or multiplication Law of probability

The probability of simultaneous happening of two events A and B is given by

$$\left.\begin{array}{l} P(A \cap B) = P(A), P(B/A); P(A) \neq 0 \\ \text{or } P(B \cap A) = P(B), P(A/B); P(B) \neq 0 \end{array}\right\} \qquad \ldots (26)$$

Where, P(B/A) is the conditional probability of happening of B under the condition that A has already been occurred and P(A/B) is the conditional probability of happening of A under the condition that B has already been occurred.

Multiplicative Law for Independent Events

If A and B are independent so that the probability of occurrence or non-occurrence of A is not affected by the occurrence or non-occurrence of B, we have,

$P(A/B) = P(A)$ and $P(B/A) = P(B)$

Hence, substituting in (26) we get

$$P(A \cap B) = P(A).P(B) \qquad \ldots (27)$$

So, the probability of simultaneous happening of two independent events is equal to the product of their individual probabilities,

Generalisation : For n events $A_1, A_2, \ldots, A_n$ we have,

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2/A_1)\, P(A_3/A_1 \cap A_2)$$

$$x \ldots x P(A_n/A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \qquad \ldots (28)$$

In particular, if $A_1, A_2, \ldots A_n$ are independent events then,

$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2) \ldots P(A_n)$ i.e., the probability of the simultaneous happening of independent events is equal to the product of their individual probabilities.

Some Important Results :

Theorem 10 : $P(\bar{A}) = 1 - P(A)$ ... (30)

Theorem 11 : (i) $P(\bar{A} \cap B) = P(B) - P(A \cap B)$ ... (31)

(ii) $P(A \cap \bar{B}) = P(A) - P(A \cap B)$ ... (32)

Theorem 12 : If $A \subset B$ then $P(A) \leq P(B)$ ... (33)

Remarks : Since $A \cap B \subset A$ and $A \cap B \subset B$ we get,

$P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$ ... (34)

Theorem 13 : If events A and B are independent then the complementary events $\bar{A}$ and $\bar{B}$ are also independent.

Remarks : In fact we also have the following results. If A and B are independent then :

(a) A and $\bar{B}$ are independent and (b) $\bar{A}$ and B are independent.

Theorem 14 : $P(A_1 \cup A_2 \cup ... \cup A_n) = 1 - (A_1 \cup A_2 \cup ... \cup A_n)^C$

$= 1 - (\bar{A}_1 \cap \bar{A}_2 \cap ... \cap \bar{A}_n)$

[By De-Morgan's Law of complementation, i.e., the complement of the union of sets is equal to the intersection of their complements.]

If $A_1, A_2, ..., A_n$ are independent, their complements $\bar{A}_1, \bar{A}_2, ..., \bar{A}_n$ are also independent.

By the compound probability theorem we get :

$P(A_1 \cup A_2 \cup ... \cup A_n) = 1 - P(\bar{A}_1)\, P(\bar{A}_2) ... P(\bar{A}_n)$ ... (36)

Remarks : The result in equation (35) is very important and can be stated in words as follows :

P[Happening of at least one of the events $A_1, A_2, ..., A_n$]

= 1 – P[None of the events $A_1, A_2, ..., A_n$ happens] ... (37)

or, equivalently,

P[None of the given events happens]

= 1 – P[At least one of them happens] ... (37a)

Pairwise and Mutual Independence :

Let A_1, A_2, A_3 are 3 events asociated with sample space S. They are said to be pairwise independent if

$P(A_i \cap A_j) = P(A_i)\, P(A_j);\ \forall\ i \neq j = 1, 2, 3$ ... (38)

A_1, A_2, A_3 are said to be mutually independent, if the law of independence holds for every subset of A_1, A_2, A_3.

Thus, A_1, A_2, A_3 are mutually independent if

$P(A_i \cap A_j) = P(A_i)\, P(A_j)$; $i \neq j = 1, 2, 3$

and $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ ...(39)

6.4.4 Bayes' Theorem (Rule for the Inverse Probability) :

If an event A can only occur in conjunction with one of the n mutually exclusive and exhaustive events $E_1, E_2, ..., E_n$ and if A actually happens, then the probability that it was preceded by the particular event $E_i (i = 1, 2, ..., n)$ is given by

$$P(E_i/A) = \frac{P(A \cap E_i)}{\sum_{i=1}^{n} P(E_i).P(A/E_i)} = \frac{P(E_i).P(A/E_i)}{\sum_{i=1}^{n} P(E_i).P(A/E_i)} \qquad ... (40)$$

Proof : Since the event A can occur in combination with any of the mutually exclusive and exhaustive events $E_1, E_2, ..., E_n$ we have :

$A = (A \cap E_1) \cup (A \cap E_2) \cup ... \cup (A \cap E_n)$

where, $A \cap E_1, A \cap E_2, ..., A \cap E_n$ are all disjoint (mutually exclusive) events. Hence by addition theorem of probability we have

$P(A) = P(A \cap E_1) + P(A \cap E_2) + ... + P(A \cap E_n)$

$= P(E_1)P(A/E_1) + P(E_2)P(A/E_2) + ... + P(E_n)P(A/E_n)$

$$= \sum_{i=1}^{n} P(E_i)P(A/E_i) \qquad ... (41)$$

For any particular event E_i, the conditional probability $P(E_i/A)$ is given by

$P(E_i \cap A) = P(A)P(E_i/A)$

$$\text{or, } P(E_i/A) = \frac{P(E_i \cap A)}{P(A)} = \frac{P(E_i)P(A/E_i)}{\sum_{i=1}^{n} P(E_i)P(A/E_i)} \qquad \text{(using equation 41)}$$

Which is the Bayes' rule for obtaining the conditional probabilities.

Remarks : The probabilities $P(E_1), P(E_2) ... P(E_n)$ which are already given or known before conducting an experiment are termed as a priori or prior probabilities. The conditional probabilities $P(E_1/A), P(E_2/A), ..., P(E_n/A)$ which are computed after conducting the experiment viz., occurrence of A are termed as posterior or inverse probabilities,

6.4.5 Statistical or Empirical definition of Probability

Definition (Von Mises) : If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the total number of trials, as the number of tirals becomes indefinitely large, is called probability of happening of the event, it being assumed that the limit is finite and unique.

Suppose that an event A occurs m times in N repetitions of a random experiment, then

$$P(A) = \lim_{N\to\infty} \frac{m}{N}$$

6.4.6 Axiomatic Definition of Probability (A.N. Kolmogorov)

Given a sample space of a random experiment, the probability of the occurrence of any event A is defined as a set function P(A) satisfying the following axioms :

Axiom 1. P(A) is defined, is real and non-negative i.e., $P(A) \geq 0$

Axiom 2. $P(S) = 1$

Axiom 3. If $A_1, A_2, ..., A_n$ is any finite or infinite sequence of disjoint events of S, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ or, } P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The above axioms are known as axioms of positiveness, certainty and additivity respectively.

Probability – Mathematical notion

Let us suppose that S is the sample space of a random experiment with sample points N i.e., n(S) = N. Let the number of occurrences (sample points) favourable to the event A be denoted by n(A). Then the frequency interpretation of the probability gives :

$$P(A) = \frac{n(A)}{n(S)} = \frac{n(A)}{N}$$

In particular, we have, $P(\phi) = \frac{n(\phi)}{n(S)} = 0$

$$P(S) = \frac{n(S)}{n(S)} = 1$$

because $n(\phi) = 0$, as the null set does not contain any sample point

Problem 9 : A box contains 4 defective and 6 good electronic calculators. Two calculators are drawn out one by one without replacement :

(i) What is the probability that the two calculators so drawn are good?

(ii) One of the two calculator so drawn is tested and found be good. What is the probability that the other one is also good?

Ans. Let us define the following events :

E_1 : First calculator is good;

E_2 : Second calculator is good

Then we have $P(E_i) = \frac{6}{6+4} = \frac{6}{10} = \frac{3}{5}$

(ii) Required Probability = $P(E_2/E_1) = \frac{5}{5+4} = \frac{5}{9}$ because, if the 1st tested calculator is good then in the box, we are left with 5 good and 4 defective calculators :

(i) The probability that both the drawn calculators are good is given by :

$$P(E_1 \cap E_2) = P(E_1).P(E_2/E_1) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3} \text{ (using part ii)}$$

Problem 10 : A box contains 6 red, 4 white and 5 blue balls. From this box 3 balls are drawn in succession. Find the probability that they are drawn in the order red, white and blue if each ball is (i) replaced, (ii) not replaced.

Ans. Let us define he following events :

A : Drawing a red ball in 1st draw

B : Drawing a white ball in 2nd draw

C : Drawing a blue ball in 3rd draw

(i) Draws with replacement : If the three balls are drawn from the box in succession with replacement, then the three events A, B and C are independent. In this case the required probability is given by

$$P(A \cap B \cap C) = P(A).P(B).P(C) = \frac{6}{15} \times \frac{4}{15} \times \frac{5}{15} = \frac{8}{225}$$

because, in this case (draws with replacement), the total number of balls in the box remains

15 for each draw.

(ii) Draws without replacement : If the three balls drawn without replacement are red, white and blue in the 1st, 2nd and 3rd draw respectively, then the constitution of the balls in the box for the 3 draws will be :

1st draw	2nd draw	3rd draw
6R, 4W, 5B	5R, 4W, 5B	5R, 3W, 5B

Hence by compound probability theorem, the required probability is given by

$$P(A \cap B \cap C) = P(A).P(B/A).P(C/A \cap B) = \frac{6}{15} \times \frac{4}{14} \times \frac{5}{13} = \frac{4}{91}$$

Problem 11 : Assume that a factory has two machines. Past records show that machine 1 produces 30% of the items of output and machine 2 produces 70% of the items. Further, 5% of the items produced by machine 1 were defective and only 1% produced by machine 2 were defective. If a defective item is drawn at random, what is the probability that it was produced by machine 1 or machine 2?

Ans. Let E_1 and E_2 denote the events that the item selected at random is produced by machines 1 and 2 respectively and let E denote the event that it is defective.. Then we are given :

$P(E_1) = 0.30$; $(P(E/E_1) = 0.05$

$P(E_2) = 0.70$; $P(E/E_2) = 0.01$

$\therefore P(E_1) \times P(E/E_1) = 0.30 \times 0.05 = 0.015$

and $P(E_2) \times P(E/E_2) = 0.70 \times 0.01 = 0.007$

$P(E) = P(E_1) \times P(E/E_1) + P(E_2) \times P(E/E_2)$

$= 0.015 + 0.007 = 0.022$

The probability that the defective item, drawn at random is produced by machine 1 is given by Bayes' rule as :

$$P(E_1/E) = \frac{P(E_1) \times P(E/E_1)}{\sum P(E_i)P(E/E_i)} = \frac{P(E_1)P(E/E_1)}{P(E)} = \frac{0.015}{0.022} = 0.6818$$

Similarly we get, $P(E_2/E) = \frac{P(E_2).P(E/E_2)}{P(E)} = \frac{0.007}{0.022} = 0.3182$

6.5 Random Variable

By a random variable we mean a real number x associated with the outcomes of a

random experiment. It can take any one of the various possible values each with a definite probability. For example, in a throw of a die if x denotes the number obtained then x is a random variable which can take any one of the values, 1, 2, 3, 4, 5 or 6, each with equal probability 1/6. Similarly, in toss of a coin if x denotes the number of heads, then x is a random variable which can take any one of the two values : 0 (i.e., no head or tail) and 1 (i.e., head), each with equal probability 1/2.

In terms of symbols if a variable x can assume discrete set of values $x_1, x_2, ..., x_n$ with respective probalities $p_1, p_2, ..., p_n$ where, $p_1 + p_2 + ... + p_n = 1$, we say that a discrete probability distribution for x has been defined.

The probability distribution of a pair of dice tossed is given below

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Where x denotes the sum of the points obtained. For example the probability of getting sum 4 is $\frac{3}{36}$ (1 + 3 = 4; 2 + 2 = 4; 3 + 1 = 4). Thus in 1200 tosses of the dice we would expect 100 tosses to give the sum 4.

6.5.1 Discrete and Continuous Random Variable

If the random variable x assumes only a finite or countably infinite set of values it is known as discrete random variable. For example marks obtained by students in a test, the number of students in a college, no. of accidents taking place in a busy road etc. are all discrete random variables.

On the other hand, if the random variable x can assume infinite and uncountable set of values it is said to be a continuous random variable, e.g., the age, height or weight of students in a class are all continuous random variables. So Random variables are essentially a real valued function on the sample space taking values on the real line $R(-\infty, +\infty)$.

6.5.2 Probability Distribution of a Random Variable

Let us consider a discrete random variable X which can take the possible values $x_1, x_2, ..., x_n$. with each value of the variable X, we associate a number

$p_i = P(X = x_i), i = 1, 2, ..., n$

which is known as the probability of x, and satisfies the following conditions

(i) $p_i = P(X = x_i) \geq 0$, i = 1, 2, ..., n

i.e., p_i's are all non-negative and

(ii) $\sum p_i = p_1 + p_2 + ... + p_n = 1$

i.e., the total probability is one.

$p_i = P(X = x_i)$ or p(x) is called the probability mass function (p.m.f) of the random variable X and the set of all possible ordered pairs {x, p(x)} is called the probability distribution of the random variable X.

A random variable X is said to be continuous if it can take all possible values between certain limits. In case of a continuous random variable, we do not talk of probability at a particular point (which is always zero) but we always talk of probability in an interval. If P(x)dx is the probability that the random variable X takes the value in a small interval of magnitude dx, eg., (x, x + dx) or $\left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)$, then P(x) is called the probability density function (p.d.f.) of the random variable X.

Moments

If X is a **discrete random variable** with probability function P(x), then

μ'_r = r-th moment about an arbitrary point 'A'.

$= (x - A)^r P(x)$ ... (42)

μ_r = r-th moment about mean $(\bar{x})$

$= \sum (x - \bar{x})^r p(x)$ ... (43)

In particular,

Mean = $\bar{x}$ = First moment about origin

$= \sum xp(x)$ ... (44)

By taking A = 0 and r = 1 in (42) we arrive at the above result (44).

Variance (x) = $\mu_2 = \sum (x - \bar{x})^2 p(x)$ ... (45)

In the case of a continuous random variable with probability density function p(x), the above formula hold with the only difference that summation is replaced by integration over the values of the variable.

$\therefore$ Mean = $\int_R (xp(x)dx$ and variance = $\int_R (x - \bar{x})^2 p(x)dx$

where R is the range (finite or infinite) in which X may lie.

6.5.3 Mathematical Expectation

If X is a random variable which can assume any one of the values $x_1, x_2, ..., x_n$ with respective probabilities $p_1, p_2, ... p_n$, then mathematical expectation of X usually called the expected value of x and denoted by E(X) is defined as :

$$E(X) = x_1p_1 + x_2p_2 + ... + x_nP_n = \sum xp \quad ... (46)$$

where, $\sum p_i = p_1 + p_2 + ... + p_n = 1$ ... (47)

So, mathematical expectation of a random variable is nothing but its Arithmatic Mean.

Important results :

$E(C) = C$ where, C is a constant ... (48)

$E(CX) = CE(X)$ where, C is a constant ... (49)

Addition Law of Expectation :

If X and Y are random variables then $E(X + Y) = E(X) + E(Y)$ ... (50)

i.e., Expected value of the sum of two random variables is equal to the sum of their expected values. The result can be generalised to n variables. If $X_1, X_2, ..., X_n$ are n random variables, then

$$E(X_1 + X_2 + ... + X_n) = E(X_1 + E(X_2) + ... + E(X_n) \quad ... (51)$$

Corollary

$E(ax + by) = aE(x) + bE(Y)$ ... (52)

where, a and b are constants (using 49 and 50)

Multiplication Law of Expectation

If X and Y are independent random variables then

$E(X.Y) = E(X).E(Y)$ ... (53)

i.e., the expected value of the product of two independent random variables is equal to the product of their expected values.

In general if $x_1, x_2 ..., x_n$ are n independent random variables, then

$$E(X_1.X_2.X_3 ... X_n) = E(X_1).E(X_2).E(X_3)... E(X_n) \quad ... (54)$$

Variance of X in terms of Expectation

$$\sigma_x^2 = Var(X) = E[X - E(X)]^2$$

or, $\sigma_x^2 = E\{X^2 + [E(X)]^2 - 2X.E(X)\}$

or, $\sigma_x^2 = E(X^2) + [E(X)]^2 - 2E(X).E(X)$

or, $\sigma_x^2 = E(X^2) - [E(X)]^2$ ... (55)

For a probability distribution {x, p(x)}, we have

Mean = $E(X) = \sum x.p(x)$ ... (56)

and variance = $E(X^2) - [E(X)]^2 = \sum x^2.p(x) - \left[\sum x.p(x)\right]^2$ ... (57)

6.5.4 Joint and Marginal probability Distribution

Let X and Y be two discrete random variables. Let us suppose that X can assume m values x_1, x_2, ..., x_m and Y can assume n values y_1, y_2, ..., y_n. Let us consider the probability of the ordered pair (x_i, y_j), i = 1, 2, ...m; j = 1, 2, ... n, defined by

$p_{ij} = P(X = x_i \text{ and } Y = y_j) = P(x_i, y_i)$ ... (58)

The function p(x, y) defined in (58) for any ordered pair (x, y) is called the joint probability function of X and Y and is presented in a tabular form :

Joint Probability Function

y↓ \ x→	x_1	x_2	x_3	...	x_i	...	x_m	Total
y_1	p_{11}	p_{21}	p_{31}	...	p_{i1}	...	p_{m1}	p'_1
y_2	p_{12}	p_{22}	p_{32}	...	p_{i2}	...	p_{m2}	p'_2
y_3	p_{13}	p_{23}	p_{33}	...	p_{i3}	...	p_{m3}	p'_3
...	...	...	...	...	...	...	...	...
y_j	p_{1j}	p_{2j}	p_{3j}	...	p_{ij}	...	p_{mj}	p'_j
...	...	...	...	...	...	...	...	...
y_n	p_{1n}	p_{2n}	p_{3n}	...	p_{in}	...	p_{mn}	p'_n
Total	p_1	p_2	p_3	...	p_i	...	p_m	1

From the joint probability distribution of X and Y, the marginal probability function of X is given by :

$p_i = P(X = x_i) = p_{i1} + p_{i2} + p_{i3} + ... + p_{in}$ (i = 1, 2, ... m)

$= \sum_{j=1}^{n} p_{ij}$

The set of values $\{x_i, p_i\}$ gives the marginal probability distribution of X. Similarly,

$p'_j = P(Y = y_j) = p_{1j} + p_{2j} + p_{3j} + ... + p_{mj}$ (j = 1, 2, ... n)

$= \sum_{i=1}^{m} p_{ij}$

gives the marginal probability function of Y and the set of values (y_j, p'_j) gives the marginal probability distribution of Y.

Again, the probabilities in different cells in a column divided by the total of that column give the conditional distribution of X, given the value of Y as stated in the column. Similarly, the cell probabilities in any row when divided by the total of that row give the conditional distribution of Y, given the value of X as stated in that row.

The random variables X and Y are said to be independent if

$P(X = x_i, Y = y_j) = P(X = x_i).P(Y = y_j)$

i.e., $P_{ij} = P_i.P_j$ for all i, j.

Otherwise, the variables are dependent and there is some association between them. Correlation co-efficient will measure the linear association

$$r_{xy} = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Cov(X, Y) = E[{X – E(X)}{Y – E(Y)}] is the covariance of X and Y

or, Cov(X, Y) = E[XY – XE(Y) – YE(X) + E(X)E(Y)]

= E(XY) – E(X)E(Y) – E(Y)E(X) + E(X)E(Y)

= E(XY) – E(X)E(Y)

Corollary : when X and Y are independent, Cov(X, Y) = E(XY) – E(X)E(Y) = 0

$\therefore$ $r_{XY} = 0$ and E(XY) = E(X).E(Y)

Problem 12 : The monthly demand for transistors in known to have the following probability distribution :

Demand(X) :	1	2	3	4	5	6
Probability(P) :	0.10	0.15	0.20	0.25	0.18	0.12

Determine the expected demand for transistors. Also obtain the variance. Suppose that the cost(C) of producing(X) transistors is given by the rule C = 10,000 + 500X. Determine the expected cost.

Ans.

Demand(X)	1	2	3	4	5	6	Total
Probability(p)	0.10	0.15	0.20	0.25	0.18	0.12	1.00
X.P	0.10	0.30	0.60	1.00	0.90	0.72	3.62
X^2.P	0.10	0.60	1.80	4.00	4.50	4.32	15.32

$E(X) = \sum XP = 3.62$

Hence the expected demand for transistors is $3.62 \simeq 4$

Variation = $E(X^2) - [E(X)]^2$

$= 15.32 - (3.62)^2$

$= 15.32 - 13.10$

$= 2.22$

Cost function is given by $C = 10{,}000 + 500X$

$E(C) = E[10{,}000 + 500X]$

$= 10{,}000 + 500E(X)$

$= 10{,}000 + 500 \times 3.62 = 11{,}810$

Problem 13 : Let (X, Y) be a pair of discrete random variables each taking three values 1, 2 and 3 with the following joint distribution :

Y \ X	1	2	3
1	5/27	4/27	2/27
2	1/27	3/27	3/27
3	3/27	4/27	2/27

obtain the marginal probability distribution of X and Y and hence find E(X), E(Y) and E(X + Y). Also find Var (X) and Var(Y)

Ans.

Marginal Dist. of X Marginal Dist. of Y

X	P(X)	XP(X)	$X^2P(X)$	Y	g(Y)	Yg(Y)	$Y^2g(Y)$
1.	$\frac{5+1+3}{27}=\frac{9}{27}$	$\frac{9}{27}$	$\frac{9}{27}$	1.	$\frac{5+4+2}{27}=\frac{11}{27}$	$\frac{11}{27}$	$\frac{11}{27}$
2.	$\frac{4+3+4}{27}=\frac{11}{27}$	$\frac{22}{27}$	$\frac{44}{27}$	2.	$\frac{1+3+3}{27}=\frac{7}{27}$	$\frac{14}{27}$	$\frac{28}{27}$
3.	$\frac{2+3+2}{27}=\frac{7}{27}$	$\frac{21}{27}$	$\frac{63}{27}$	3.	$\frac{3+4+2}{27}=\frac{9}{27}$	$\frac{27}{27}$	$\frac{81}{27}$
Total		$\frac{52}{27}$	$\frac{116}{27}$		Total	$\frac{52}{27}$	$\frac{120}{27}$

$E(X) = \sum x.p(x) = \frac{52}{27} = 1.93$

$E(Y) = \sum y.g(b) = \frac{52}{27} = 1.93$

$E(X^2) = \sum x^2p(x) = \frac{116}{27} = 4.30$

$E(Y^2) = \sum y^2.g(b) = \frac{120}{27} = 4.44$

$Var(X) = E(x^2) - [E(X)]^2$

$= 4.30 - (1.93)^2 = 4.30 - 3.72$

$= 0.58$

$Var(Y) = E(Y^2) - [E(Y)]^2$

$= 4.44 - (1.93)^2 = 4.44 - 3.72$

$= 0.72$

$E(X + Y) = E(X) + E(Y) = 1.93 + 1.93 = 3.86$

6.6. Probability Distribution – Discrete and Continuous

In this section we shall study the following univariate probability distributions.

(i) Binomial Distribution

(ii) Poisson Distribution

(iii) Normal Distribution

The first two are discrete probability distributions and the third is a continuous probability distribution.

6.6.1 Binomial Distribution

Binomial distribution is also known as the 'Bernoulli distribution' after the Swiss mathematician James Bernoulli who discovered it in 1700. This distribution can be used under the following conditions :

(i) The random experiment is performed repeatedly a finite and fixed number of times,

i.e., n number of trials is finite and fixed

(ii) The outcome of each trial may be classified into two mutually disjoint categories, called success (the occurrence of the event) and failure (the non-occurrence of the event).

(iii) All the trials are independent, i.e., the result of any trial is not affected in any way by the preceding trials and does not affect the result of succeeding trials.

(iv) The probability of success (happening of an event) in any trial is p and is constant for each trial. $q = 1 - p$, is then termed as the probability of failure (non-occurrence of the event) for each trial.

The problems relating to (i) tossing of a fair coin or throwing of a fair die n times (which is fixed and finite) and (ii) drawing of cards from a pack of cards with replacement will conform to Binomial distribution.

Probability Function of Binomial Distribution

If X denotes the number of successes in n trials satisfying the above conditions, then X is a random variable which can take the values 0, 1, 2, ..., n; since in n trials we may get no success (all failures), one success, two successes, ... or all the n successes.

The general expression for the probability of r successes in n independent trials is given by

$$p(r) = P(X = r) = {}^{n}c_{r} \cdot p^{r} \cdot q^{n-r}; \; r = 0, 1, 2, ..., n \quad ... (59)$$

Remarks : 1. Putting r = 0, 1, 2, ..., n in (59)

We get the probabilities of 0, 1, 2, ..., n successes respectively given by

q^n, ${}^{n}c_1 q^{n-1}p$, ${}^{n}c_2 q^{n-2}p^2$, ..., p^n

Since these probabilities are the successive terms in the binomial expansion $(q + p)^n$, it is called The Binomial Distribution.

2. The expression for P(X = r) in (59) is known as the probability mass function of the Binomial distribution with parameters n and p. The random variable X following the probability Law (59) is called a Binomial variate with parameters n and p and we write $X \sim B(n, p)$

The Binomial distribution is completely determined i.e., all the probabilities can be obtained if n and p are known. Obviously q is known when p is given because $q = 1 - p$.

3. Since the random variable X takes only integral values, Binomial distribution is a discrete probability distribution.

Constants of Binomial Distribution

Mean = np ... (60)

Variance = $\sigma^2 = npq \rightarrow \mu_2 = npq$ ... (61)

$\mu_3 = npq(q - p)$ ... (62)

$\mu_4 = npq[1 + 3pq(n - 2)]$ ... (63)

The moment co-efficient of skewness is :

$$\beta_1 = \frac{(q-p)^2}{npq} \text{ or, } \gamma_1 = +\sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} \quad \text{... (64)}$$

Co-efficient of kurtosis is given by :

$$\beta_2 = 3 + \frac{1-6pq}{npq} \quad \text{or } \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq} \quad \text{... (68)}$$

Remarks : 1. Since q is the probability (of failure), we always have $0 < q < 1$

$\therefore$ Variance = np.q < np $\rightarrow$ Variance < mean

Hence, for Binomial distribution, variance is less than mean.

2. Binomial distribution is symmetrical $\gamma_1 = 0$ if p = q = 0.5.

It is positively skewed if p < 0.5 and negatively skewed if p > 0.5

Mode of Binomial Distribution

Mode is the value of X which maximises the probability function :

Working rule to Find Mode of Binomial Distribution

Let X be a Binomial Variate with Parameters n and p.

Case I : When (n + 1)p is an integer = k(say). In this case, the distribution is bi-modal, the two modal values being X = k and X = k – 1

Case II : When (n + 1)p is not an integer.

Let $(n + 1)p = k_1 + f$

Where, k_1 is the integral part and f is the fractional part of (n + 1)p. In this case, the distribution has a unique mode at $X = k_1$, the integral part of (n + 1)p.

Fitting of Binomial Distribution

Suppose a random experiment consists of n trials, satisfying the conditions of Binomial Distribution and suppose this experiment is repeated N times. Then, the frequency of r successes is given by the formula :

$f(r) = N \times p(r) = N \times {}^nC_r\, p^r q^{n-r}$; r = 0, 1, 2, ..., n ... (69)

Putting r = 0, 1, 2, ..., n we get the expected or theoretical frequencies of the Binomial distribution as given below :

$Nq^n,\ N.{}^nC_1 q^{n-1}p,\ N.{}^nC_2 q^{n-2}p^2,\ ...,\ Np^n$ ... (70)

If p is not known and if we want to fit a Binomial distribution to a given frequency distribution, we first find the mean of the given frequency distribution by the formula $\bar{x} = \frac{\sum fx}{\sum f}$ and equate it to np, which is the mean of the Binomial probability distribution. Hence p can be estimated by the relation :

$$np = \bar{x} \rightarrow p = \frac{\bar{x}}{n} \qquad \ldots (71)$$

Then q = 1 – p. With these values of p and q, the expected or theoretical Binomial frequenceies can be obtained by using the expression (70)

6.6.2 Poisson Distribution (As a Limiting case of Binomial Distribution)

Poisson distribution was derived by the French mathematician Simon D. Poisson as a limiting case of Binomial probability distribution under the following conditions :

(i) n, the number of trials is indefinitely large i.e., $n \rightarrow \infty$

(ii) p, the constant probability of success for each trial is indefinitely small i.e., $p \rightarrow 0$

(iii) np = m (say) is finite.

Under the above three conditions the Binomial probability function (59) tends to be the probability function of the poisson distribution given below :

$$p(r) = P(X = r) = \frac{e^{-m}.m^r}{r!},\ r = 0, 1, 2, \ldots \qquad \ldots (72)$$

Where X is the number of successes (occurrences of the event), m = np and e = 2.71828

The proof is given below : $\lim \left(\frac{n}{r}\right) p^r q^{n-r}$

$$= \lim \frac{n(n-1)(n-2)...(n-r+1)}{r!} p^r (1-p)^{n-r}$$

$$= \lim \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right)}{r!} (np)^r \left(1-\frac{np}{n}\right)^{n-r}$$

$$= \lim \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right)}{r!} m^r \left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{r!}.\lim\left\{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right)\right\} \times \lim\left(1-\frac{m}{n}\right)^n \times \frac{1}{\lim\left(1-\frac{m}{n}\right)^r}$$

$$= e^{-m}\frac{m^r}{r!}, \text{ Since, } \lim_{n\to\infty}\left\{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{r-1}{n}\right)\right\}=1, \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m} \text{ and}$$

$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^r = 1$ for a given r.

The essence of this result is that when n is sufficiently large and p is quite small, but np(= m) has a moderate value, the binomial probabilities $\binom{n}{r}p^r q^{n-r}$ can be well approximated by the corresponding poisson probabilities $\frac{e^{-m}.m^r}{r!}$

Remarks : 1. Poisson distribution is a discrete probability distribution, since the variable X can take only integral values 0, 1, 2, ..., ∝

2. If we know m, all the probabilities of the poisson distribution can be obtained. m is, therefore called the parameter of the poisson distribution and X ~ P(m)

Importance of Poisson Distribution

We have given certain practical situations where poisson distribution can be used :

(i) Number of telephone calls received at a telephone switch board per minute

(ii) The number of defective materials say, bulbs in a packing manufactured by a producing firm.

(iii) The number of deaths due to say heart attack in a year.

(iv) Number of accidents taking place per day on a busy road.

(v) Number of printing mistakes per page in a book

Constants of Poisson Distribution

For the Poisson Distribution with parameter m, we have

Mean = Variance = m ... (73)

i.e., mean and variance are equal, each being equal to m.

Other constants : The moments, (about mean) of the poisson distribution are :

$\mu_1 = 0$, μ_2 = variance = m, $\mu_3 = m$, $\mu_4 = m + 3m^2$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}\ ;\ \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}} \qquad \text{... (74)}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3m^2 + m}{m^2} = 3 + \frac{1}{m} \qquad \text{... (75)}$$

$$\gamma_2 = \beta_2 - 3 = \frac{1}{m} \qquad \text{... (76)}$$

Mode of Poisson Distribution

Case (i) When m is an ingeter : If m is an integer, equal to k(say), then the poisson distribution is bimodal, the two modes being at the points X = k and X = k – 1.

Case (ii) When m is not an integer : If m is not an integer, then the distribution is unimodal. The unique modal value will be the integral part of m. For example, if m = 9.4, then mode = 9, the integral part of 9.4.

Fitting of Poisson Distribution

If we want to fit a poisson distribution to a given frequencey distribution, we compute the mean $\bar{x}$ of the given distribution and take it equal to the mean of the fitted (poisson) distribution, i.e.., we take m = $\bar{x}$.

Once m is know, the various probabilities of the poisson distribuion can be obtained, the general formula being.

$$p(r) = p(X = r) = \frac{e^{-m} \times m^r}{r!},\ r = 0, 1, 2 \ldots \qquad \text{... (77)}$$

If N is the total observed frequency, then the expected or theoretical frequencies of the poisson distribution are given by N × p(r)

6.6.3 Normal Distribution

Normal distriution is one of the most important continuous theoretical probability distribution in statistics. It was first discovered by English Mathematician De-Moivre in 1733 while dealing with problem arising in the game of chance and later developed by Karl Friedrich Gauss who used this distribution to describe the errors of measurements in the calculation of orbits of heavenly bodies.

Equation of Normal Probability Curve

If X is a continuous random variable following normal probability distribution with mean μ and standard deviation σ, then its probability density function (p.d.f) is given by :

$$p(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty \qquad \ldots (78)$$

$$\text{or } p(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \qquad \ldots (78a)$$

Where, Π and e are the constants given by :

$$\Pi = \frac{22}{7}, \sqrt{2\Pi} = 2.5066 \text{ and } e = 2.71828$$

Remarks :

1. The mean μ and standard deviation σ are called the parameters of the Normal distribution so, the normal distribution is defined symbollically us $X \sim N(\mu, \sigma^2)$

2. If X is a random variable following normal distribution with mean μ and standard deviation σ, then the random variable Z defined as follows :

$$Z = \frac{X - E(X)}{\sigma_x} = \frac{X - \mu}{\sigma} \qquad \ldots (79)$$

is called the standard normal variate for which

$E(Z) = 0$ and $\text{var}(Z) = 1 \rightarrow \sigma_z = 1$ ... (80)

i.e., the standard normal variate Z has mean 0 and standard deviation 1.

Hence, the probability density function of Z is

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < Z < \infty \qquad \ldots (81)$$

This we have derived by putting $x = Z$, $\mu = 0$ and $\sigma = 1$ in (78).

6.6.4 Relationship between Binomial and Normal Distributions :

Normal distribution can be viewed as a limiting case of Binomial Distribution under the following conditions :

(a) n, the number of trials is indefinitely large that is, $n \to \infty$

(b) Neither p nor q is very small.

Consider the variate $Z = \frac{x - np}{\sqrt{npq}}$ ... (82)

where x is a Binomial variate and np, $\sqrt{npq}$ are mean and standard Deviation of the Binomial Distibution. When x = 0, $Z = \frac{-np}{\sqrt{npq}} = -\sqrt{\frac{np}{q}}$

which tends to $-\infty$ as $n \to \infty$.

when x = n, $Z = \frac{n - np}{\sqrt{npq}} = \frac{n(1-p)}{\sqrt{npq}} = \frac{nq}{\sqrt{npq}}$

$= \sqrt{\frac{nq}{p}}$ which tends to ∞ as $n \to \infty$

As x changes from x to x + 1, the change in variate

$Z = \frac{x+1-np}{\sqrt{npq}} - \frac{x - np}{\sqrt{npq}} = \frac{1}{\sqrt{npq}}$

This change tends to zero as $n \to \infty$ i.e., $\frac{1}{\sqrt{npq}}$ is infinitesimally small.

Let it be denoted by dx. Thus Z varies from $-\infty$ to $+\infty$ and the change is infinitesimally small and hence x may be regarded as a continuous variate.The probability of x successes in a Binomial Distribution is given by

$P(x) = {}^{n}c_{x} q^{n-x} p^{x}$

Now, the limiting value of p(x) as $n \to \infty$ is the probability dp when x lies in the interval $x - \frac{dx}{2}, x + \frac{dx}{2}$

$dp = \lim_{n \to \infty} P(x) = \lim_{n \to \infty} {}^{n}C_{x} q^{n-x} P^{x}$

$$= \lim_{n\to\infty} \frac{n!}{(n-x)!x!} q^{n-x} p^x$$

Using stirling's approximation formula

$$n! = \sqrt{2\Pi} e^{-n} . n^{n+\frac{1}{2}}$$

$$\therefore dp = \lim_{n\to\infty} \frac{\sqrt{2\Pi}\, e^{-n} . n^{n+\frac{1}{2}} . q^{n-x} . p^x}{\sqrt{2\Pi}\, e^{-x} . x^{x+\frac{1}{2}} \sqrt{2\Pi}\, e^{-(n-x)} . (n-x)^{n-x+\frac{1}{2}}}$$

$$= \lim_{n\to\infty} \frac{n^{n+\frac{1}{2}} q^{n-x} p^x}{\sqrt{2\Pi} . x^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}}}$$

Multiplying the numerator and denominator by $\sqrt{npq}$ we get,

$$dp = \lim_{n\to\infty} \frac{n^{n+1} q^{n-x+\frac{1}{2}} p^{x+\frac{1}{2}}}{\sqrt{2\pi}\sqrt{npq} . x^{x+\frac{1}{2}} . (n-x)^{n-x+\frac{1}{2}}}$$

$$= \lim_{n\to\infty} \frac{n^{n+1}}{\sqrt{2\pi}\sqrt{npq}} \times \frac{1}{x^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}}} \times \frac{(nq)^{n-x+\frac{1}{2}}}{n^{n-x+\frac{1}{2}}} \times \frac{(np)^{x+\frac{1}{2}}}{n^{x+\frac{1}{2}}}$$

$$= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}\sqrt{npq}} \left(\frac{np}{x}\right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x}\right)^{n-x+\frac{1}{2}}$$

$$= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} . \frac{1}{\sqrt{npq}} \times \frac{1}{k}$$

where, $k = \left(\frac{x}{np}\right)^{x+\frac{1}{2}} . \left(\frac{n-x}{nq}\right)^{n-x+\frac{1}{2}}$

From (82) $Z = \dfrac{x - np}{\sqrt{npq}}$

or, $x = np + z\sqrt{npq}$

or, $\dfrac{x}{np} = 1 + z.\sqrt{\dfrac{q}{np}}$

Again, $n - x = n - np - z\sqrt{npq} = nq - z\sqrt{npq}$

or, $\dfrac{n-x}{nq} = 1 - z\sqrt{\dfrac{p}{nq}}$

$$\therefore k = \left(1 + z\sqrt{\frac{q}{np}}\right)^{x+\frac{1}{2}} . \left(1 - z\sqrt{\frac{p}{nq}}\right)^{n-x+\frac{1}{2}}$$

$$= \left(1 + z\sqrt{\frac{q}{np}}\right)^{np+z\sqrt{npq}+\frac{1}{2}} . \left(1 - z\sqrt{\frac{p}{nq}}\right)^{nq-z\sqrt{npq}+\frac{1}{2}}$$

[$\because x = np + z\sqrt{npq}$ and $(n - x) = nq - z\sqrt{npq}$]

Now taking log of both sides, we get,

$$\log k = \left(np + z\sqrt{npq} + \frac{1}{2}\right)\log\left(1 + z\sqrt{\frac{q}{np}}\right) + \left(nq - z\sqrt{npq} + \frac{1}{2}\right)\log\left(1 - z\sqrt{\frac{p}{nq}}\right)$$

But $z\sqrt{\dfrac{q}{np}}$ and $z\sqrt{\dfrac{p}{nq}}$ both are less than 1, since $n \to \infty$

$$\therefore \log k = \left(nq + z\sqrt{npq} + \frac{1}{2}\right)\left[z\sqrt{\frac{q}{np}} - \frac{z^2}{z}.\frac{q}{np} + \ldots\right]$$

$$+ \left(nq - z\sqrt{npq} + \frac{1}{2}\right)\left[-z\sqrt{\frac{p}{nq}} - \frac{z^2}{2}.\frac{p}{nq} + \ldots\right]$$

Collecting the terms in desending order of n

$$= \left[z\sqrt{npq} + \left(-\frac{z^2}{2}q + z^2 q \right) + \frac{z^3}{3}.\frac{q^{3/2}}{\sqrt{np}} - \frac{z^3}{2}.\frac{q^{3/2}}{\sqrt{np}} + \frac{z}{2}.\frac{\sqrt{q}}{\sqrt{np}} + \ldots \right]$$

$$+ \left[-z\sqrt{npq} + \left(-\frac{z^2}{2}.p + z^2 p \right) + \left(-\frac{z^3}{3}.\frac{p^{3/2}}{\sqrt{nq}} + \frac{z^3}{2}.\frac{p^{3/2}}{\sqrt{nq}} + \frac{z}{2}.\frac{\sqrt{p}}{\sqrt{nq}} \right) + \ldots \right]$$

$$= \frac{z^2}{2}(q+p) + \frac{1}{\sqrt{n}}\left[\frac{z^3}{3}\left(\frac{-q^{3/2}}{\sqrt{p}} + \frac{p^{3/2}}{\sqrt{q}} \right) + \frac{z}{2}\left(\sqrt{\frac{q}{p}} - \sqrt{\frac{p}{q}} \right) + \ldots \right]$$

$$= \frac{z^2}{2} + \frac{1}{\sqrt{n}}\left[\left(\frac{z^3}{3}.\frac{p^2 - q^2}{\sqrt{pq}} \right) + \frac{z}{2}\left(\frac{q-p}{\sqrt{pq}} \right) \right] + \text{terms containing higher powers of } \frac{1}{\sqrt{n}}.$$

Taking limits $n \to \infty \lim_{n\to\infty} \log k = \frac{z^2}{2}$ or, $k = e^{\frac{z^2}{2}}$ or, $\frac{1}{k} = e^{\frac{-z^2}{2}}$

$$\therefore\ dp = \frac{1}{\sqrt{2\pi}}.e^{\frac{-z^2}{2}}\,dz$$

If $np = \mu$; $\sqrt{npq} = \sigma$ then $z = \frac{x - np}{\sqrt{npq}} = \frac{x-\mu}{\sigma}$

$$dz = \frac{dx}{\sigma}$$

$$dp = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx$$

Putting $\mu = 0$, $\sigma = 1$ in $z = \frac{x-\mu}{\sigma}$, we get, $dp = \frac{1}{\sqrt{2\pi}}.e^{-\frac{z^2}{2}}\,dz$ of variate z which has mean '0' and standard deviation '1' and hence it is called the standard normal distribution. The variate z is known as the standard normal variate.

$dp = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$ with mean μ and standard deviation σ is called the general normal distribution and $y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ is known as Normal Probability curve.

Relation between Poisson and Normal Distribution

If X is a random variable following poisson distribution with parameter m, then the standard poisson variate $Z = \frac{X - E(X)}{\sigma_x} = \frac{X - m}{\sqrt{m}}$ tends to be a standard normal variate if $m \to \infty$. Thus Normal distribution may also be regarded as a limiting case of poisson distribution as the parameter $m \to \infty$.

6.6.5 Moments of Normal Distribution

(a) Odd order moments about mean are given by

$$\mu_{2n+1} = \int_{-\infty}^{\infty} (x-\mu)^{2n+1} f(x)dx$$

$$= \int_{-\infty}^{\infty} (x-\mu)^{2n+1} . \frac{1}{\sqrt{2\pi}\,\sigma} . e^{\frac{-(x-\mu)^2}{2\sigma^2}} .dx$$

Putting $\frac{x-\mu}{\sigma} = z$

or, $x - \mu = \sigma z$

$\therefore\ dx = \sigma dz$

$$\mu_{2n+1} = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} (\sigma z)^{2n+1} . e^{\frac{-z^2}{2}} . \sigma dz$$

$$= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} . \int_{-\infty}^{\infty} Z^{2n+1} . e^{-\frac{z^2}{2}} .dz = 0,$$ Since the integrand is an odd function

$\therefore\ \mu_{2n+1} = 0$; (n = 0, 1, 2, ...) i.e., $\mu_1 = \mu_3 = \mu_5 = ... = 0$

(b) Even order moments about mean are given by $\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} f(x)dx$

$$= \int_{-\infty}^{\infty} (x-\mu)^{2n} . \frac{1}{\sigma\sqrt{2\pi}} . e^{\frac{-(x-\mu)^2}{2\sigma^2}} .dx$$

Now, putting $\frac{x-\mu}{\sigma} = z$ or, $x - \mu = 6z$

$\therefore$ $dx = \sigma dz$

So, $\mu_{2n} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n} . e^{\frac{-z^2}{2}} .dz(\sigma)$

or, $\mu_{2n} = \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} . e^{\frac{-z^2}{2}} .dz$

The integral $z^{2n} . e^{\frac{-z^2}{2}}$ being even function, we may write

$$\int_{-\infty}^{\infty} z^{2n} . e^{\frac{-z^2}{2}} dz = 2\int_{0}^{\infty} z^{2n} . e^{\frac{-z^2}{2}} .dz$$

$$\mu_{2n} = \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_{0}^{\infty} z^{2n} . e^{\frac{-z^2}{2}} .dz$$

Now, we put $\frac{z^2}{2} = t$ or, $z = \sqrt{2}.\sqrt{t}$ or, $dz = \frac{\sqrt{2}}{2} . \frac{1}{\sqrt{t}} dt$

$$\therefore \mu_{2n} = \frac{2(\sigma)^{2n}}{\sqrt{2\pi}} . \int_{0}^{\infty} (2t)^n . e^{-t} . \frac{1}{\sqrt{2}} \frac{dt}{\sqrt{t}}$$

$$= \frac{2^n(\sigma)^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} t^{n-\frac{1}{2}} . e^{-t} .dt = \frac{2^n(\sigma)^{2n}}{\sqrt{\pi}} . \Gamma\left(n + \frac{1}{2}\right)$$

Putting (n – 1) in place of n we get, $\mu_{2n-2} = \frac{2^{n-1}(\sigma)^{2n-2}}{\sqrt{\pi}}\Gamma\left(n-\frac{1}{2}\right)$

$$\therefore \frac{\mu_{2n}}{\mu_{2n-2}} = \frac{2\sigma^2\Gamma\left(n+\frac{1}{2}\right)}{\Gamma\left(n-\frac{1}{2}\right)} = \frac{2\sigma^2\left(n-\frac{1}{2}\right)\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(n-\frac{1}{2}\right)}$$

$\therefore \mu_{2n} = \sigma^2(2n-1)\mu_{2n-2}$

$\mu_2 = \sigma^2(1)\mu_0 = 1\sigma^2 = \sigma^2$ ($\because \mu_0 = 1$)

$\mu_4 = \sigma^2(3)\mu_2 = 3\sigma^4$

$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ ($\because \mu_3 = 0$)

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$

$\gamma_1 = \sqrt{\beta_1} = 0$ and $\gamma_2 = \beta_2 - 3 = 0$.

Since the Normal distribution is symmetrical, the moment co-efficient of skewness is given by

$\beta_1 = 0 \Rightarrow \gamma_1 = 0$

i.e., the normal curve is not skewed at all. The co-efficient of Kurtosis is given by

$\beta_2 = 3 \Rightarrow \gamma_2 = 0$

i.e., the Normal curve is meso-kurtic

6.6.6 Mode of Normal Distribution

The mode of Normal distribution is the value of x for which $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is maximum. Taking log of both sides we get $\log f(x) = \log \frac{1}{\sqrt{2\pi}\,\sigma} + \log e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

$= \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2}$

Differentiating we get $\frac{1}{f(x)}.f'(x) = 0 - \frac{2(x-\mu)}{2\sigma^2}$

[Since, $\frac{d}{dx}\log\frac{1}{\sigma\sqrt{2\pi}} = 0$ because $\log\frac{1}{\sigma\sqrt{2\pi}}$ is constant.]

or, $f'(x) = \frac{-2(x-\mu)}{2\sigma^2}.f(x)$ ... (83)

$f'(x) = 0$ at $x = \mu$

Differentiating (83) again we get,

$f''(x) = -\left[\frac{x-\mu}{\sigma^2}f'(x) + \frac{1}{\sigma^2}f(x)\right]$

Putting $x = \mu$ we get,

$f''(\mu) = -\frac{1}{\sigma^2}f(x) = -\frac{1}{\sigma^2}.\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ = negative at $x = \mu$

$\therefore$ at $x = \mu$, $f(x)$ is maximum.

Hence the mode of Normal distribution is μ.

6.6.7 Median of Normal Distribution

Median M of the Normal distribution is the value of x such that $\int_{-\infty}^{M} f(x)dx = \frac{1}{2}$

or, $\int_{-\infty}^{M} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$

or, $\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\mu} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx + \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu}^{M} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$ ... (84)

Consider the integral $I = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$

Putting $\frac{x-\mu}{\sigma} = z$ we get, $dx = \sigma dz$

$I = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} e^{\frac{-z^2}{2}} \sigma dz$

Let us put $z = -t$ or, $dz = -dt$

$\therefore I = \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{0} e^{\frac{-t^2}{2}} \sigma(-1)dt = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-z^2}{2}} dz$

$= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} . \frac{\sqrt{2\pi}}{2} = \frac{1}{2}$

Putting this value of I in equation (84) we get, $\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$

$= I + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$

or, $\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} . \int_{\mu}^{M} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{2}$

or, $\frac{1}{\sigma\sqrt{2\pi}} . \int_{\mu}^{M} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 0$

Hence $M = \mu$ (i.e., Median = Mean = Mode)

6.6.8 Points of Inflexion of Normal Curve

At the point of inflexion of a curve we know, the conditions are $f'(x) = 0$ and $f''(x) = 0$

The probability density function of a Normal curve is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Taking log of both sides we get, $\log f(x) = \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2}$

Differentiating both sides we get, $\frac{f'(x)}{f(x)} = \frac{-2(x-\mu)}{2\sigma^2} = -\frac{(x-\mu)}{\sigma^2}$

$\therefore f'(x) = \frac{-(x-\mu)}{\sigma^2}.f(x)$ ... (85)

Differentiating the above equation we get,

$f''(x) = -\left[\frac{(x-\mu)}{\sigma^2}.f'(x) + f(x) \times \frac{1}{\sigma^2}\right]$

$= -\left[\frac{(x-\mu)}{\sigma^2}.\left\{-\frac{(x-\mu)}{\sigma^2}.f(x)\right\} + f(x) \times \frac{1}{\sigma^2}\right]$ [Using equation (85)]

$= -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$

$\therefore f''(x)$ will be = 0 when $1 - \frac{(x-\mu)^2}{\sigma^2} = 0$

or, $(x-\mu)^2 = \sigma^2$ or $x = \mu \pm \sigma$

Hence the points of inflexion are at $x = (\mu \pm \sigma)$

6.6.9 Properties of Normal Distribution

The normal probability curve with mean μ and standard deviation σ is given by :

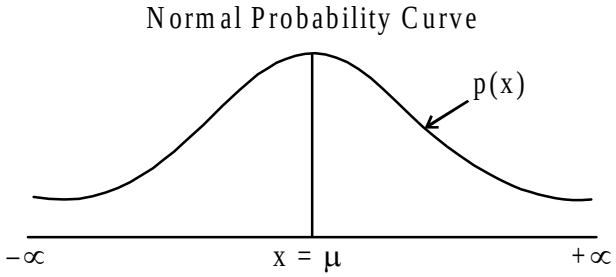
$p(x) = \frac{1}{\sqrt{2\pi}.\sigma}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$

The standard normal probability curve is given by the equation :

$$\phi(z) = \frac{1}{\sqrt{2\pi}}.e^{\frac{-z^2}{2}}, -\infty < z < \infty$$

It has the following properties :

1. The graph p(x) is the famous bell shaped curve as shown below. The peak point lies directly above the mean μ.



2. The curve is symmetrical about the line $X = \mu$, ($Z = 0$), i.e., it has the same shape on either side of the line $X = \mu$ (or $Z = 0$)

3. Since the distribution is symmetrical, mean, median and mode coincide. Thus,

Mean = Mode = Mode = μ

4. Since Mean = Median = μ, the ordinate at $X = \mu$, ($Z = 0$) divides the whole area into two equal parts. Further, since total area under normal probability curve is 1, the area to the right as well as to the left of the ordinate at $X = \mu$ (or $Z = 0$) is 0.5.

5. Also by virtue of symmetry, the quartiles are equilistant from median (μ) i.e., Q_3 – Median = Median – Q, or $Q_1 + Q_3 = 2$ Median = 2μ

6. Since the distribution is symmetrical, the moment co-efficient of skewness is given by

$\beta_1 = 0$ or $\gamma_1 = 0$ which indicates no skewness.

7. The co-efficient of Karlosis is given by

$\beta_2 = 3$ or $\gamma_2 = 0$ which indicates Mesokurtic Kurtosis.

8. Theoretically the range of the distribution is from $-\infty$ to $+\infty$. But Practically, the range is $\mu \pm 6\sigma$

9. As x increases numerically i.e., on either side of $X = \mu$, the value of p(x) decreases rapidly, the maximum probability occurrence falls at $x = \mu$ and is given by

$$[p(x)]_{max} = \frac{1}{\sqrt{2\pi}\ \sigma}$$

10. Since the distribution is symmetrical, all moments of odd order about the mean are zero.

$\therefore\ \mu_{2n+1} = 0$, (n = 0, 1, 2, ...) i.e., $\mu_1 = \mu_3 = \mu_5 = ... = 0$

11. The moments about mean of even order are given by $\mu_{2n} = (2n-1)\sigma^{2n}$; (n = 1, 2, 3, ...)

Putting n = 1 and 2 we get $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$

12. x-axis is an asymptote to the curve.

13. If x_1 and x_2 are independent normal variates with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively. Then,

(i) $x_1 + x_2$ is a normal variate with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$

(ii) $x_1 - x_2$ is a normal variate with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 - \sigma_2^2$

Therefore the sum as well as difference of independent normal variates is also a normal variate.

15. Mean Deviation about mean or median or mode ($\because$ Mean = Median = Mode) is given by

$$\text{Mean Deviation} = \sqrt{\frac{2}{\pi}}.\sigma = 0.7979\,\sigma \simeq \frac{4}{5}\sigma$$

16. Quartiles are : $Q_1 = \mu - 0.6745\sigma$ and $Q_3 = \mu + 0.6745\sigma$

17. Quartile Deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = 0.6745\sigma \simeq \frac{2}{3}\sigma$$

18. Area under the normal curve

Mean $\pm\ \sigma$ = 68.26%

Mean $\pm\ 2\sigma$ = 95.45%

Mean $\pm\ 3\sigma$ = 99.73%

6.6.10 Fitting a Normal Distribution to an Observed Distribution

To examine if a normal distribution will fit an observed distribution, we have to find the expected frequencies for different class-intervals and then compare the two series of frequencies.

The expected frequency for the class interval A to B, where B > A, is

$$N\left[\phi\left(\frac{B-\mu}{\sigma}\right)-\phi\left(\frac{A-\mu}{\sigma}\right)\right]$$

where, N is the total observed frequency.

When the values of μ and σ are not specified, we first estimate these by the method of moments. Equating μ and σ with the mean $\bar{x}$ and standerd deviation s, respectively of the observed distribution, we get estimates of μ and σ as follows :

$\hat{\mu} = \bar{x}$ and $\hat{\sigma} = s$

To draw the fitted curve over the histogram of the observed distribution, we compute the ordinates.

$$N \times \frac{1}{\sigma\sqrt{2\pi}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

as the class-boundaries of the observed distribution.

6.6.11 Importance of Normal Distribution

1. If X is a normal variate with mean μ and variance σ^2, then

$P(\mu - 3\sigma < \times < \mu + 3\sigma) = P(-3 < Z < 3) = 0.9973$

Thus, in all probability we should expect a standard normal variate to lie between the limits ± 3. This property of the normal distribution forms the basis of the large sample theory.

2. Most of the discrete probability distribution (e.g., Binomial, Poisson) tend to normal distribution as $n \to \propto$ (n = no of trials).

3. The entire theory of small sample tests viz., t, F, χ^2 tests etc is based on the assumption that the parent population from which samples have been drawn follows Normal Distribution.

4. Most important application of Normal distribution lies in the central limit theorem which states that "If X_1, X_2, ... X_n is a random sample of size n from any population with mean μ and variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{\sum x}{n}$ is asymptotically normal (as $n \to \propto$) with mean μ and variance $\frac{\sigma^2}{n}$".

5. Normal distribution is used in Statistical Quality Control in industry for the setting of control limits.

6.6.12 Some Worked out examples on Probability Distribution

Example 14 : A manufacturing process turns out articles that are on the average 10% defective. Compute the probability of 0, 1, 2 and 3 defective articles that might occur in a sample of 3 articles.

Ans. P = Probability of a defective article = 10% = 0.10

$q = 1 - P = 1 - 0.10 = 0.9$

The probability of x defectives in a sample of 3 articles is given by the Binomial Probability law by :

$P(x) = {}^3C_x\, P^x.\, q^{3-x} = {}^3C_x(0.1)^x.(0.9)^{3-x}$

where, x = 0, 1, 2, 3 ... (i)

Putting x = 0, 1, 2, 3 in (i) we get the probability of 0, 1, 2, 3 defective articles respectively :

$P(0) = (0.9)^3 = 0.729$

$P(1) = {}^3C_1\,(0.1)(0.9)^2 = 3 \times 0.1 \times 0.81 = 0.243$

$P(2) = {}^3C_2\,(0.1)^2(0.9) = 3 \times 0.01 \times 0.9 = 0.027$

$P(3) = (0.1)^3 = 0.001$

Example 15 : With the usual notations, find ϕ for a binomial random variable X if n = 6 and if $9p(x = 4) = p(x = 2)$

Ans. For the binomial random variable x with parameters n = 6 and p, the probability mass function is, $p(r) = p(x = r) = {}^6C_r p^r q^{6-r}$; r = 0, 1, 2, ..., 6 ... (a)

We are given : $9.p(x = 4) = p(x = 2)$

or, $9 \times {}^6C_4\, p^4q^2 = {}^6C_2p^2q^4$ [from (a)]

or, $9p^2 = q^2$ [$\because {}^6C_4 = {}^6C_{6-4} = {}^6C_2$ and $p \neq 0$]

or, $9p^2 = (1 - p)^2 = 1 + p^2 - 2p$

or, $8p^2 + 2p - 1 = 0$

$$\therefore p = \frac{-2 \pm \sqrt{4 + 4 \times 8}}{2 \times 8} = \frac{-2 \pm \sqrt{36}}{16} = \frac{-2 \pm 6}{16} = \frac{-8}{16} \text{ or, } \frac{4}{16} = -\frac{1}{2} \text{ or, } \frac{1}{4}$$

Since the value of probability can't be negative, $p = -\frac{1}{2}$ is rejected. Hence the value of $p = \frac{1}{4}$.

Example 16 : In a certain factory turning out blades, there is a small chance $\frac{1}{500}$ for

any one blade to be defective. The blades are supplied, in packets of 10. Use poisson distribution to calculate the approximate number of packets containing no defective, one defective, two defective, three defective blades respectively in a consignment of 10,000 packets. [Given that $e^{0.02} = 0.9802$]

Ans. Given N = 10,000; n = 10

p(Probability of a defective blade) = $\frac{1}{500}$

$$m = np = 10 \times \frac{1}{500} = \frac{1}{50} = 0.02$$

Let the random variable x denote the member of defective blades in a pack of 10. Then by Prisson probability rule, the probability of r defective blades in a packet is given by

$$p(x = r) = \frac{e^{-0.02}(0.02)^r}{r!} = \frac{0.9802 \times (0.02)^r}{r!}$$

Hence is a consignment of 10,000 packets the frequency (number) of packets containing r defective blades is :

$$NxP(X = r) = \frac{10,000 \times 0.9802 \times (0.02)^r}{r!} \qquad \text{... (a)}$$

Putting r = 0, 1, 2 and 3 in equation (a) we get : No. of packets containing no defective blade = 10,000 × 0.9802 = 9802

No. of packets containing 1 defective blade is : $\frac{10,000 \times 0.9802 \times (0.02)^1}{1}$

= 9802 × 0.02 = 196

No. of packets containing 2 defective blades is :

$$\frac{10,000 \times 0.9802 \times (0.02)^2}{2!} = \frac{196.04 \times 0.02}{2} = 1.9604 \simeq 2$$

No. of packets containing 3 defective blades is :

$$\frac{10,000 \times 0.9802 \times (0.02)^3}{3!} = \frac{1.9604 \times 0.02}{3} = 0.0130 \simeq 0$$

$\therefore$ No. of packets : 9802, 196, 2, 0

Example 17 : If 2% of electric bulbs manufactured by a certain company are defective, find the probability that in a sample of 200 bulbs (i) less than 2 bulbs (ii) more than 3 bulbs are defective. (Given : $e^{-4} = 0.0183$)

Ans. Given that n = 200

p = probability of a defective bulb = 2% = 0.02

m = mean number of defective bulbs in the box = np = 200 × 0.02 = 4.

Using Poisson approximation to the Binomial distribution, the probability of x defective bulbs in the box is given is :

$$p(x = x) = p(x) = \frac{e^{-m}m^x}{x!} = \frac{e^{-4}.4^x}{x!} \quad \text{... (i)}$$

(i) Required probability is given by

$p(x < 2) = p(0) + p(1) = e^{-4}.4^0 + e^{-4}.4^1 = e^{-4}(1 + 4) = e^{-4}.5$

$= 0.0183 \times 5 = 0.0915$ (using equation (i)

(ii) Required probability is given by

$p(x > 3) = 1 - p(x \leq 3) = 1 - \{p(0) + p(1) + p(2) + p(3)\}$

$$= 1 - e^{-4}\left(1 + 4 + 8 + \frac{32}{3}\right) \quad \text{[using (i)}$$

$= 1 - 0.0183 \times 23.6667$

$= 1 - 0.4331 = 0.5669$

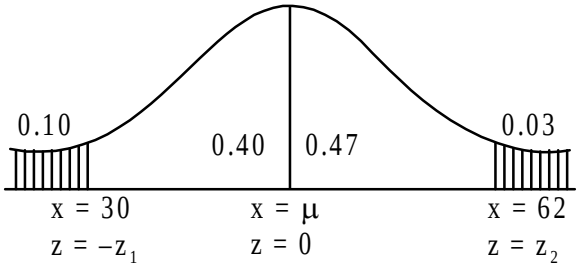
Example 18 : At a certain examination 10% of the students who appeared for the paper in statistics got less than 30 marks and 97% of the students got less than 62 marks. Assuming the distribution to be normal, find the mean and the standard deviation of the distribution.

Ans. Let x denotes the marks in statistics and let $x \sim N(\mu, \sigma^2)$. Then from the given information we may write

$P(X < 30) = 0.10$ and $P(X < 62) = 0.97$

$$\text{when } x = 30,\ z = \frac{30 - \mu}{\sigma} = -z_1 \text{ (say)} \qquad \text{... (i)}$$

NB : z_1 is negative as it falls to the left of μ where z = 0



when $x = 62$, $z = \dfrac{62 - \mu}{\sigma} = z_2$ (say) ... (ii)

z_2 is positive as it falls to the right of $x = \mu$ where $z = 0$. From the diagram it can be seen that

$p(0 < z < z_2) = 0.47$

From the Normal Table, $z_2 = 1.88$

Again, $p(-z_1 \leq z \leq 0) = 0.40$

By symmetry, $p(0 \leq z \leq z_1) = 0.40$

where, from the Normal Tables we get $z_1 = 1.28$. Substituting the values of z_1 and z_2 in (i) and (ii) we get,

$30 - \mu = -1.28\sigma$ ... (iii)

$62 - \mu = 1.88\sigma$ ... (iv)

Substracting (iii) and from (iv) we get,

$32 = 3.16\sigma$ or, $\sigma = \dfrac{32}{3.16} = 10.13$

Substituting this in (iii) we get,

$\mu = 30 + 1.28 \times 10.13 = 30 + 12.97 = 42.97 \simeq 43$

$\therefore$ Mean = 43 and S.D = 10.13

Example 19 : The following rules are followed in a certain examination. A candidate is awarded a first division if his aggregate marks are 60% or above, a second division if his aggregate marks are 45% or above but less than 60% and a third division if the aggregate marks are 30% or above but less than 45%.

A candidate is declared failed if his aggregate marks are below 30%. A candidate is awarded distinction if his aggregate marks are 80% or above.

At such an exam if is found that 10% of the candidates have failed. 5% of them obtained distinction calculate the percentage of students who are placed in the second

division. (Assume Normal distribution of marks).

Given, the areas under the standard normal curve from z = 0 to Z = z are

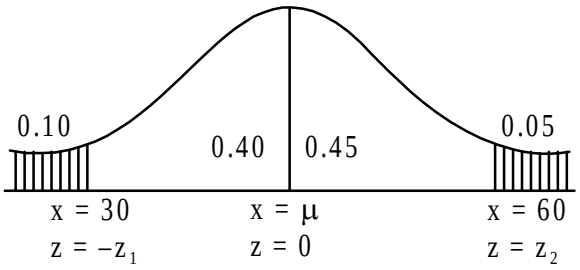
Z	1.28	1.64	0.41	0.47
area	0.4000	0.4500	0.1591	0.1808

Ans. Let the random variable X denote the aggregate marks (out of 100) obtained in the exam. Then a candidate fails if X < 30 and gets a distinction if x ≥ 80 and we are given that

$$\left.\begin{aligned} p(x < 30) &= 0.10 \\ p(x \geq 80) &= 0.05 \end{aligned}\right\} \quad \ldots (i)$$

when x = 30

$$\left.\begin{aligned} z &= \frac{30-\mu}{\sigma} = -z_1 \text{ (say)} \\ \text{when } x &= 60, \\ z &= \frac{60-\mu}{\sigma} = z_2 \text{ (say)} \end{aligned}\right\} \quad \ldots (ii)$$



From, the diagram we have,

$p(z < -z_1) = 0.10$ $\qquad$ $p(z > z_2) = 0.05$

or, $p(z > z_1) = 0.10$ (By symmetry) $\qquad$ $p(0 \leq z \leq z_2) = 0.5 - 0.05 = 0.45$

or, $p(0 \leq z \leq z_1) = 0.5 - 0.1 = 0.4$ $\qquad$ $\Rightarrow z_2 = 1.64$

$\Rightarrow z_1 = 1.28$

Now substituting these values of z_1 and z_2 in (ii) we get,

$30 = \mu - 1.28\sigma$ and $80 = \mu + 1.64\sigma$ $\qquad$ ... (iii)

Solving (iii) for μ and σ we get

$\mu = 51.9 \simeq 52$ and $\sigma = 17.12$

The percentage of students who are placed is the second division is given by :

$100 \times p(45 \le \times < 60) = 100 \times p(-0.41 \le z < 0.47)$

$= 100 \times [p(-0.41 \le z \le 0) + p(0 \le z < 0.47)]$

$= 100 \times [p(0 \le z \le 0.41) + p(0 \le z < 0.47)]$

$= 100[0.1591 + 0.1808) = 100 \times 0.3399 = 33.99 = \simeq 34$

6.7 Moment Generating Function (MGF)

The moment generating function (MGF) of a random variable x is defined as the mathematical expectation of the exponential function, e^{xt}, for any real value of t.

If x is a discrete random variable with probability distribution given by $P(x = x_i) = p_i(x)$; i = 1, 2 ... n then,

$$M_x(t) = Ee^{xt} = \sum_{i=1}^{n} e^{x_i t} p_i(x)$$

On the other hand, if x is a continuous, random variable with probability distribution given by f(x)dx, a < x < b then,

$$M_x(t) = Ee^{xt} = \int_a^b e^{xt} f(x)dx$$

The MGF exists is finite, if and only if the integral or the series on the right is convergent.

6.7.1 MGF of Binomial Distribution

For binomial distribution with parameters n and p, the MGF is

$$M_x(t) = E(e^{tx}) = \sum e^{tx} p(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} = (q + pe^t)^n$$

The moments about zero are obtained as

$$\left[\frac{d^r}{dt^r} M_x(t) \right]_{t=0} = \mu'_r(0)$$

μ = mean = $\mu_1'(0) = \left[\frac{d}{dt}M_x(t)\right]_{t=0} = [n(q + pe^t)^{n-1}.pe^t]_{t=0}$

$= n(q + p)^{n-1}.p = np \quad [\because\ p + q = 1]$

$\mu_2'(0) = \left[\frac{d^2}{dt^2}M_x(t)\right]_{t=0} = \left[\frac{d}{dt}\left\{\frac{d}{dt}M_x(t)\right\}\right]_{t=0}$

$= \left[\frac{d}{dt}\left\{np(q + pe^t)^{n-1}.e^t\right\}\right]_{t=0}$

$= [np\{(n-1)(q + pe^t)^{n-2}.pe^t.e^t + (q + pe^t)^{n-1}.e^t\}]_{t=0}$

$= np\{(n-1)p + 1\} = n(n-1)\,p^2 + np$

$\therefore\ \sigma^2 = Var(x) = \mu_2 = \mu_2' - \mu_1'^2 = n(n-1)p^2 + np - n^2p^2$

$= -np^2 + np = np(1-p) = npq$

MGF about mean :

$M_{np}(t) = \sum_0^\infty e^{(x-np)t}\,p(x)$

$= \sum e^{(x-np)t}.{}^nC_x q^{n-x}.p^x$

$= \sum e^{xt}.e^{-npt}.{}^nC_x q^{n-x}.p^x$

$= e^{-npt}\sum {}^nC_x q^{n-x}.p^x.e^{xt}$

$= e^{-npt}\sum {}^nC_x q^{n-x}(pe^t)^x$

$= e^{-npt}(q + pe^t)^x$

$= [qe^{-pt} + pe^{t(1-p)}]^n = [qe^{-pt} + pe^{qt}]^n$

$= \left[q\left(1 - pt + \frac{p^2t^2}{2!} + \ldots\right) + p\left(1 + qt + \frac{q^2t^2}{2!} + \ldots\right)\right]^n$

$= \left[(q+p) + qp(p+q)\frac{t^2}{2!} + pq(q-p)\frac{t^3}{3!} + \ldots\right]^n$

$$= \left[1 + qp\frac{t^2}{2!} + pq.(q-p)\frac{t^3}{3!} + \ldots\right]^n$$

$$= 1 + n\left[pq\frac{t^2}{2!} + pq(q-p)\frac{t^3}{3!} + \ldots\right] + \frac{n(n-1)}{2!}\left[pq\frac{t^2}{2!} + pq(q-p).\frac{t^3}{3!} + \ldots\right]^2 + \ldots$$

$$= 1 + npq.\frac{t^2}{2!} + npq(q-p)\frac{t^3}{3!} + \ldots$$

μ_2 = Co-efficient of $\frac{t^2}{2!} = npq$

μ_3 = Co-efficient of $\frac{t^3}{3!} = npq(q-p)$ and so on

Thus μ_4 = Co-efficient of $\frac{t^4}{4!} = 3n^2p^2q^2 - npq(1-6pq)$

(skewness) $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq}$ and (Kurtosis) $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq}$

6.7.2 MGF of Poisson Distribution

$$M_x(t) = \sum_{x=0}^{\infty} e^{tx}p(x) = \sum e^{tx}.\frac{e^{-m}.m^x}{x!}$$

$$= e^{-m}\sum_{x=0}^{\infty}\frac{(me^t)^x}{x!} = e^{-m}.e^{me^t} = e^{m(e^t-1)}$$

Moments about zero is given by $\mu_r'(0) = \left[\frac{d^r}{dt^r}M_x(t)\right]_{t=0}$

$$\therefore\ \mu_1'(0) = \left[\frac{d}{dt}e^{m(e^t-1)}\right]_{t=0} = e^{m(e^t-1)me^t} = m$$

$$\mu_2'(0) = \left[\frac{d^2}{dt^2} e^{m(e^t-1)}\right]_{t=0} = \left[\frac{d}{dt}\left(e^{m(e^t-1)}, me^t\right)\right]_{t=0}$$

$$= m\left[e^{m(e^t-1)}.m(e^t)^2 + e^{m(e^t-1)}.e^t\right]_{t=0}$$

$= m(m + 1)$ $\therefore$ mean $= \mu_1'(0) = m$

variance $= \mu_2' - \mu_1'^2 = m(m + 1) - m^2 = m$

MGF about mean is given by

$$M_m(t) = \sum e^{t(x-m)} p(x)$$

$$= \sum e^{t(x-m)}.\frac{e^{-m}.m^x}{x!}$$

$$= e^{-m-mt}\sum \frac{e^{tx}.m^x}{x!} = e^{-m-mt}\sum \frac{(me^t)^x}{x!}$$

$$= e^{-m-mt} e^m e^t = e^{m(e^t-t-1)}$$

Again, $e - t - 1 = \left(1 + t + \frac{t^2}{2!} + \ldots\right) - t - 1$

$$= \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots$$

$$M_m(t) = e^{m\left(\frac{t^2}{2!}+\frac{t^3}{3!}+\ldots\right)}$$

$$= 1 + m\left(\frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right) + \frac{m^2}{2!}\left(\frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^2 +$$

$$= 1 + m\frac{t^2}{2!} + m\frac{t^3}{3!} + (m + 3m^2)\frac{t^4}{4!} + \ldots$$

Co-efficient of $t = 0$ $\therefore$ $\mu_1 = 0$

Co-efficient of $\frac{t^2}{2!} = m$ $\therefore$ $\mu_2 = m$

Co-efficient of $\frac{t^3}{3!} = m \ \therefore \ \mu_3 = m$

Co-efficient of $\frac{t^4}{4!} = m + 3m^2 \ \therefore \ \mu_4 = m + 3m^2$

6.7.3 MGF of Normal distribution

$$M_a(t) = \int_{-\infty}^{\infty} e^{t(x-a)} dp = \int_{-\infty}^{\infty} e^{t(x-a)} \times \frac{1}{\sqrt{2\pi}.\sigma} . e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

Putting $\frac{x-\mu}{\sigma} = z$ and $dx = \sigma dz$

$$M_a(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^t (\mu + 6z - a) . e^{\frac{-z^2}{2}} . dz$$

$$= \frac{e^{t(\mu-a)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t\sigma z} . e^{\frac{-z^2}{2}} . dz$$

$$= \frac{e^{t(\mu-a)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-z^2}{2} + t\sigma z} . dz$$

But $-\frac{z^2}{2} + t\sigma z = -\frac{1}{2}(z^2 - 2t\sigma z)$

$$= -\frac{1}{2}(z^2 - 2t\sigma z + t^2\sigma^2) + \frac{t^2\sigma^2}{2}$$

$$= -\frac{1}{2}(z - t\sigma)^2 + \frac{t^2\sigma^2}{2}$$

$$\therefore \ M_a(t) = \frac{e^{t(\mu-a)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2 + \frac{t^2\sigma^2}{2}} dz$$

$$= \frac{e^{t(\mu-a)+\frac{t^2\sigma^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t\sigma)^2} dz$$

Again by putting $z - t\sigma = u$ and $dz = du$ we get

$$= \frac{e^{t(\mu-a)+\frac{t^2\sigma^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-\mu^2}{2}} du$$

$$= \frac{e^{t(\mu-a)+\frac{t^2\sigma^2}{2}}}{\sqrt{2\pi}} \times \sqrt{2\pi}$$

$$\therefore \; M_a(t) = e^{t(\mu-a)+\frac{t^2\sigma^2}{2}}$$

MGF about origin $M_0(t) = e^{t\mu+\frac{t^2\sigma^2}{2}}$ [by putting a = 0]

Moment generating function about mean is obtained by putting $a = \mu$.

$$M_\mu(t) = e^{\frac{t^2\sigma^2}{2}}$$

Cumulants of Normal Distribution

Cumulants of Normal distribution are obtained from

$$KM_0(t) = \log M_0(t) = \log e^{\mu t+\frac{t^2\sigma^2}{2}} = \mu t + \frac{t^2\sigma^2}{2} \qquad \text{... (i)}$$

The first cumulant k_1 = Co-efficient of t in (i)

$= \mu$ = Mean

The second cumulant k_2 = Co-efficient of $\frac{t^2}{2} = \sigma^2$ = Variance

Similarly third and fourth cumulants k_3 and k_4 are equal to zero.

6.8 Summary

Theory of probability deals with uncertainly in decigion making. Whenever the decision maker is in the middle of probabilistic dilemma, testing of hypothesis provides a very strong statistical tool which helps him arrive at a statisfactorily close to an accurate solution. This hypothesis testing is totally dependent on the theory of probability.

There are three approaches to probability (a) classical approach (b) Impirical approach and (c) Axiomatic approach.

The outcome of a random experiment is called an event.

Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others.

All outcomes are equally probable which indicates there is no bias in the dice or a coin. Events are said to be independent of each other if the happening of any one of them is not affected by any one of the others.

Permutation means 'arrangement' whereas combination means 'selection'.

In the classical probability theory, probability of an event A is defined as

$$p(A) = \frac{\text{Favourable member of cases to A}}{\text{Exhaustive number of cases}}$$

Set theoretical notations are helpful in understanding the theory of probability. Union of two events will denote the occurrence of at least one of the events. Intersection will denote the occurrence of both the events. Probability of an event cannot exceed 1.

If A and B are dependent events, then the conditional probability of 'A', given the fact that 'B' has already been occurred, is defined as

$$P(A/B) = \frac{n(A \cap B)}{n(B)} \text{ provided } n(B) \neq 0$$

P(A/B) is the conditional probability, $n(A \cap B)$ is the number of occurrence of both A and B, n(B) is the number of occurrence of event B.

If A and B are two independent events, then we have $P(A \cap B) = P(A) \times P(B)$.

Bayes' Theorem introduces the rule of the inverse probability.

Under Empirical definition of probability, an event A occurs m times in N repetitions of a random experiment, then

$$P(A) = \lim_{N \to \infty} \frac{m}{n}$$

Under the Axiomatic definition of probability, if $A_1, A_2, \ldots A_n$ is any finite squence

of events which are essentially disjoint events, then

(1) $P\left(\bigcup_{i=1}^{n} Ai\right) = \sum_{i=1}^{n} P(Ai)$; i = 1, 2, ... n.

(2) P(S) = 1

(3) $1 \geq P(A) \geq 0$

Random variable is different from the ordinary variable in the sense that the occurrence of it is associated with a definite probability. If the random variable x assumes only a finite set of values it is known as discrete random variable. Whereas if the random variable x assumes infinite set of values it is said to be continuous. Number of students in a college is discrete while the age or height of the students in a class will be continuous.

We have discussed the nature of two types of discrete probability distribution viz., Binomial distribution and poisson distribution. On the other hand the most important continuous probability distribution has been studied at length is the Normal probability distribution.

Moment Generating Function is unique and it determines completely the distribution of a random variable. If two random variables have the same moment generating function, they have the same distribution. However the moment generating function will exist if and only if the integral or the series on the right hand side is convergent.

6.9 Questions for Self Assessment

1. (A) True or False

   (i) The probability of an event may exceed unity.

   (ii) Probability of obtaining 3 or 5 in throwing a fair die is $\frac{1}{4}$

   (iii) Two mutually exclusive events with positive probabilities must be mutually independent.

   (iv) If x and y are independent, then E(xy) = E(x). E(y)

   (v) The p.d.f of standard normal variable t is $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$, $-\infty < t < +\infty$.

2. (a) What is classical definition of probability?

   (b) In a single throw of two dice, what is the probability of obtaining a total of nine.

(c) The possibility that a certain question can be solved by A is $\frac{1}{3}$ and by B is $\frac{1}{4}$. What is the probability that the question will be solved by any of them.

(d) A dice is tossed thrice. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of successes.

(e) If a person gains or loses an amount equal to the number appearing when a blanced die is rolled once, according to whether the number is even or odd, how much money can he expect per game in the run?

(f) (i) what is conditional probability?
(ii) Explain Bayes' Theorem.

(g) A coin is tossed three times. Let x denote the number of heads. Compute the variance of x.

(h) A coin is tossed until a head appears. What is the expected number of tosses?

3. Prove that poisson distribution is the limiting form of Binomial distribution.

4. Describe the properties of a Normal curve.

5. What is Moment Generating Function? Derive the mean, variance, measures of skewness and kurtosis from the moment generating function of a binomial distribution.

6. Find out the points of inflexion of a Normal Curve.

7. Find out the Mode of a Normal Distribution.

8. With the help of the moment generating function find out μ_1, μ_2, μ_3 and μ_4, the notations having their usual meaning.

9. The following table gives frequencies of occurrence of a variate x between certain limits.

Variate x	**f**
Less than 40	30
40 or more but less than 50	33
50 and more	37
	100

The distribution is normal. Find mean and S.D.

10. A normal curve has $\bar{x} = 20$ and $\sigma = 10$. Find the area between $x_1 = 15$ and $x_2 = 40$.

6.10 References

1. Goon, Gupta & Dasgupta : Fundamentals of statistics vol I & II, The World Press, Cal, 1983.
2. A. L. Nagar and Sharma P.D. : Statistical Methods and Econometric Analysis, S. Chand & Co. New Delhi, 1987.
3. John Mounsey : Introduction to statistical calculations, English Universities Press, 1964.
4. Das, N.G. : Statistical Methods Part I & II, M. Das and Co. Cal, 1977
5. Bowen and Starr : Basic Statistics for Business & Economics, McGraw Hill Co., 1982.