

**M.A. Semester – II  
ECONOMICS**

**Course Code: ECON 122  
Course Credit : 06 (DSC)**

# **Basic Statistics**

**Units: 1 to 20**

**By: Dr. K. Kaushik  
Prem Prakash**



**Centre for Distance and Online Education  
Himachal Pradesh University  
Summer Hill, Shimla, 171005**

# CONTENTS

UNIT NO.	TOPICS	PAGE NO.
1	Measures of Central Tendency	3
2	Measures of Dispersion and Skewness	28
3	Correlation and Regression	53
4	Fitting of Regression Equation and Standard Error of Estimate	77
5	The General Linear Regression Model: Matrix Formulation and Solution-I	86
6	The General Linear Regression Model: Matrix Formulation and Solution-II	97
7	Multiple and Partial Correlation-I	106
8	Multiple and Partial Correlation-II	115
9	Probability Theory and Concepts of Probability Distribution	124
10	Mathematical Expectation	137
11	Statistical Hypothesis	146
12	Non-Parametric Methods	155
13	Chi-square Test	164
14	Standard Error of Mean and Student's t Distribution	176
15	Testing Homogeneity of Several Independent Estimates of Population Variance	188
16	Analysis of Variance	200
17	Index Numbers	215
18	Tests for Consistency of Index Numbers	228
19	Analysis of Time Series	239
20	Measures of Non-linear Trends	250

# **M.A. (ECONOMICS) COURSE—VI**

## **BASIC STATISTICS**

***Maximum Marks: 100***

### **Unit -I Measures of Central Tendency**

Measure of Central Tendency, Dispersion, Skewness and Kurtosis. Correlation; meaning and methods of measuring Correlation, Karl Pearson's method, Spearman's Rank Correlation Coefficient, Limitation's of correlation analysis. Linear Regression; relation between correlation coefficient and regression coefficients, Fitting of regression equations, Standard error of estimate.

### **Unit - II: The General Linear Regression Model**

An introduction to the matrix formulation and solution of the General Linear Regression Model. Solution for a model with one dependent and two independent variables. Prediction for simple regression models of demand, supply, production and cost. Multiple and partial correlations, and regressions. Relationship between the measures of multiple correlation and measures of partial-correlation, Beta Coefficients.

### **Unit - III : Elements of Probability Theory :**

The Concept of Probability distribution and a density function. Mathematical Expectation Binomial distribution, the normal distribution some properties of the normal distribution. Sampling and sample designs : Simple random sampling; stratified random sampling, systematic sampling and cluster sampling. Large samples. Test of significance, Limitations of sampling, Procedure of testing hypothesis : Regions of acceptance and rejection; Two tailed and one, tailed test. Type 1 and Type 2 errors. [Non-parametric Tests: The sign test, rank sum test. The Mann-Whitney U test, advantages and limitations of non parametric test.

### **Unit -IV: Tests of Significance**

Standard error of mean Student's 't' distribution and its properties; Use of 't' distribution to test hypothesis of population means Chi-square general features of chi-square, chi-square as a test of goodness of fit, chi-square as a test of independence. Contingency table and date's connection for continuity. Testing homogeneity of several independent, estimates of population variance. Analysis of variance; meaning, assumptions and techniques of analysis of variance; one way and two way analysis of variance problem, Inter-relationship between 't',  $X^2$  (Chi sq.) and F test.

### **Unit - V : Analysis of Time Series :**

Meaning and components of time series. Methods of estimating trend—the semi average method, the moving average method and the least squares method; Fitting of straight line, second and third degree equations. Fitting of the mollified exponential curve, the Gompertz curve and the logistic curve. Measurement of seasonal cyclical and irregular variations.

Index Numbers : meaning, problems in the construction of index numbers. Classification of index numbers : unweighed price index numbers, relative of aggregate method and average of price relatives, weighted price index numbers. Laspeyres. Paache's and fisher's ideal index-numbers. Time reversal test and factor reversal test fixed and chain base index numbers. Uses and limitations of index numbers.

**Instructions For Candidates For This Courses Shall be as Follows ; —**

1. Question paper will consist of eleven questions in all. The first question (at serial No. 1) will consist of 10 short-answer type questions which will, cover the entire syllabus uniformly and will be based on concepts and definitions only. This question will carry 20 marks in all and each short-answer *type* question with answer about five lines (fifty words) will carry 2 marks each. The test of the ten questions (from Serial No. 2 to 11) will be such that there will be two essay type -questions' each 'from the five units of the syllabus, which will carry 16 marks each.
2. Candidates are required to attempt six questions in all. The question number one (with 10 parts) is compulsory and rest five questions in such a way choosing one questions each from the five' waits.

**Suggested Reading**

1. Taro Yamane, *Statistics*, Harper International.
2. M. "R. Spiegel. *Theory and Practice of Probability and Statistics*, Schaum's outline series, McGraw Hill.
3. A. L. Nagar and R. K. Das. *Basic Statistics*, Oxford University Press, New Delhi.
4. George Snedecar and W.G. Chockrane *Statistical Methods*, Oxford & IBH, New Delhi.
5. F.E. Croxton, D. J. Cowdea and Sidney Klein. *Applied General Statistics*, Prentice Hall of India, New Delhi.
6. S.P. Gupta, *Statistical Methods*, Sultan Chand & Sons, New Delhi.
7. S.P. Singh. (1996). *Statistic*, S. Chand & Company, New Delhi.
8. B.L. Agarwal, (1997), *Basic Statistics*, New Age International Limited, New Delhi.
9. H. M. Walker and J. Lev. (1953): *Statistical Inference*, Holt, Rimehart and Winston, Oxford and IBH Publishing Company, Calcutta.
10. Damodar N. Gujarati, *Basic Econometrics*, Second Edition, McGraw Hill Book Company, New York.
11. L.R. Klein, *Introduction to Econometrics*.

\*\*\*\*\*

# LESSON-1

## MEASURES OF CENTRAL TENDENCY

Dear Students,

Let me welcome you to the course on basic statistics which you will be studying through a course of a limited number of lessons. The basic attempt underlying these lessons is to make statistics as simple and clear as possible. Before I start explaining to you the various measures of central tendency and dispersion. I would like you to introduce to few basic concepts. “Statistics’ consists’ of two parts descriptive ‘statistics and statistical inference. Descriptive statistics deals with the collection, organization and presentation of data, while Statistical inference deals with generalization from a part to the whole. Statistical /inference deals with the development of methods as well as their use.

A population can be defined as the totality of all possible observations on measurements or outcomes, thus we may have human population, cattle population, population of students enrolled in Himachal Pradesh University etc. A population may be either finite or infinite.

Related to the concept of a population is the concept of sample, which is a set of measurements or outcomes selected from the population. An important type of probability sample is the random sample.

Both populations and samples can be described by, stating their characteristics. Numerical characteristic of a population are called parameters; the characteristics of a sample, given in the form of some summary measure; are called statistics.

With respect to a phenomenon that can be measured, is known as a variable which means, a homogeneous quantity that can assume different values at different points of observation. If a phenomenon can only be counted but not measured, we speak of an attribute. The definition of a variable, stresses the possibility of variation at different point of observation. On the other hand, a quantity that cannot vary from one observation to another is called constant A continuous variable is a variable that can assume any value on the numerical axis or a part of it. In contrast to a continuous variable, a discrete variable is one that can assume only some specific values on the numerical axis.

The final concept to be introduced at this stage is that of distribution. In the case of a sample we have & frequency distribution, while in the case of population we speak of a probability distribution. A distribution which confines to one ‘variable—is known as an univariate distribution, whereas when it deals with two or more variables, we have a bivariate and multivariate distribution.

### **Measure of Central Tendency**

With this unit, we begin our formal discussion of the statistical ‘methods for summarizing and describing numerical methods for summarizing and describing numerical data. The objective here is to find one representative value which can be used to locate and summarize the entire set of varying values. This one value can be used to make many decisions concerning the entire set. We can define measures of central tendency (or location) to find some central value around which the data tend to cluster.

Measures of central tendency, enable us to get an idea of the entire data. For example, it is impossible to remember, individual earnings of crores of earning people in India. But if the average income is 'obtained, a single value will represent the entire, population. These measures 'also enable us to compare two or more sets of data to facilitate comparison. For example, the average production figures of one particular year may be compared with the production figures of previous years.

A good measure of central tendency should possess, as far as possible, the following-properties.

- (i) It should be easy to understand
- (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.

Following are some of the important measures of central, tendency which are commonly used.

Arithmetic mean

Weighted Arithmetic mean

Median

Mode

Geometric Mean

Harmonic Mean

Arithmetic Mean

The arithmetic mean (or mean or average) is the most commonly used and readily understood measure, of central tendency. In statistics the term average refers to any of the measures of central tendency. The arithmetic mean is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations. Symbolically, it can be represented as :

$$\bar{X} = \frac{\sum x}{n}$$

where  $\sum x$  indicates the sum of the values of all the observations, and  $N$  is the total number of observations. For example, Let us consider the monthly sale of ten firms in Rs. lakh

10, 18, 20, 28, 30, 25, 40, 30, 20, 10.

If we compute the arithmetic then)

$$\begin{aligned}\bar{X} &= \frac{10+18+20+28+30+25+40+30+20+10}{10} \\ &= \frac{221}{10} = \text{Rs. } 22.10 \text{ lakhs.}\end{aligned}$$

Therefore the average monthly sale is Rs 22.10 Lakhs.

## The Arithmetic Mean

**Example I :** 100 crates of golden delicious ‘large’ variety of apples are sold at Rs. 40. big case (of 18 kg) another 250 crates of ‘medium’ variety at Rs. 30 per big case and 400 cases of ‘small’ variety at Rs. 25 per case, in the Simla market on a particular day.

Then

$$\begin{aligned}\bar{X} &= \frac{(100 \times 40) + (250 \times 30) + (400 \times 25)}{100 + 250 + 400} \\ &= \frac{21500}{500} = \text{Rs. } 28.67 \text{ kg. per case.}\end{aligned}$$

symbolically, if  $x_1, x_2, x_3, \dots, x_n$  are the values of a variable, the mean is computed by the formula

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + X_3 + \dots + X_n}{N} \\ &= \frac{\sum_{i=1}^n X_i}{N} \dots \dots \dots (1)\end{aligned}$$

Where  $\Sigma$  = the sum of  $\bar{X}$  = the mean of values,  $x_1$  = the value of variables,  $N$  = Number of values.

## Weighted Arithmetic Mean

In computing the arithmetic mean we give equal importance to each item in the series. This equal importance may be misleading if some items in a distribution are more important than others. In such cases ‘various items’ are given proper weights. The weights assigned to each item being proportional to the importance of the item in the distribution. For example, if we want to have an idea of the change in the cost of living of a certain group of people, then the simple mean will not give the correct result, because the commodities to be selected will, not be of equal importance. It is, therefore, necessary to calculate weighted mean in such cases.

So that *weighted average or weighted arithmetic mean*

$$\begin{aligned}\bar{X} &= \frac{W_1 X_1 + W_2 X_2 + \dots + W_n X_n}{W_1 + W_2 + W_3 + \dots + W_n} \\ &= \frac{\Sigma WX}{\Sigma W} \dots \dots \dots (2)\end{aligned}$$

where  $W_1, W_2, W_3, \dots, W_n$  stand for the weights of the items  $X_1, X_2, X_3, \dots, X_n$  respectively.

**Example 2 :** Table 1.1 shows the heights of the students in a class.

**Table 1.1**  
**Heights of Students in Economic Class**

Heights in Inches	Number of Students		
$X$	$f$	$fx$	
59 ( $X_1$ )	2 ( $f_1$ )	118	$f_1X_1$
60 ( $X_2$ )	5 ( $f_2$ )	300	$f_2X_2$
61( $X_3$ )	12 ( $f_3$ )	732	$f_3X_3$
62 ( $X_4$ )	20 ( $f_4$ )	1240	$f_4X_4$
63 ( $X_5$ )	35 ( $f_5$ )	2205	$f_5X_5$
64( $X_6$ )	48 ( $f_6$ )	3072	$f_6X_6$
65 ( $X_7$ )	30 ( $f_7$ )	1950	$f_7X_7$
65 ( $X_8$ )	18 ( $f_8$ )-	1188	$f_8X_8$
67 ( $X_9$ )	7 ( $f_9$ )	469	$f_9X_9$
68( $X_{10}$ )	3( $f_{10}$ )	204	$f_{10}X_{10}$
	$\Sigma f=180$	$\Sigma fX= 11.478$	

Arithmetic Mean of the series

$$\begin{aligned}\bar{X} &= \frac{f_1X_1 + f_2X_2 + f_3X_3 + \dots + f_nX_n}{f_1 + f_2 + f_3 + \dots + f_n} \\ &= \frac{\Sigma fX}{\Sigma f = N} \quad \dots\dots\dots (3) \\ &= \frac{11.478}{180} = 63.77'' \text{ app.}\end{aligned}$$

**Short-cut Method**

We can find the arithmetic mean by another method, where ‘we assume one of the heights to be the mean and then add a collection factors, with the help, of the following formula :

$$\bar{X} = \bar{X}_a + \frac{\Sigma D}{N} \quad \dots\dots\dots (4.1)$$

(for data without any weights or frequencies)

$$\text{or } \bar{X} = \bar{X}_a + \frac{\Sigma fD}{\Sigma f (= N)} \text{ (for data with frequencies)} \quad \dots\dots\dots (4.2)$$

and correspondingly

$$\bar{X} = \bar{X} + \frac{\Sigma wD}{\Sigma w (= N)} \text{ (for data with weight)} \quad \dots\dots\dots (4.3)$$



Where  $\bar{X}_a$  stands for the assumed mean

D stand for the value of the deviation of the variate from

the assumed mean =  $X - \bar{X}_a$ ,

$f$  stands for frequencies,

$w$  stands for the weights.

Let us apply this method to the data of Table 4.1. We assume the mean to be 64". Normally, the assumed mean should be the value of the variate with the highest frequency or the one, which lies in middle, In case of & symmetrical distribution or mildly skewed series, both these would tend to coincide. Our assumed mean 64" has the highest frequency (48) and is almost in the middle of the series.'

**Worksheet I based on Table 1.1** for computing Mean Height of Students in Economics Class

Height (X) in inches (1)	Frequency ( $f$ ) (2)	Deviations from $\bar{X}_a - 64$ $D = X - \bar{X}_a$ (3)	$f \times D$ (4) = (2), (3)
59	2	— 5	— 10
60	5	— 4	— 20
61	12	— 3	— 36
62	20	— 2	— 40
63	35	— 1	— 35
64	48	0	0
65	30	+1	+ 30
66	18	+2	+ 36
67	7	+3	+ 21
68	3	+4	+ 12
	$\Sigma f = N = 180$		-141+99 = -42

$$\bar{X} = \bar{X}_a + \frac{\Sigma fD}{N} = 64 + \frac{-42}{180}$$

$$= 64 - 0.23 = 63.77''.$$

In this method, we avoid multiplication of big number.

### The Arithmetic Mean from Grouped Date : Long Method

When dealing with a frequently distribution, we do not ordinarily have the original date from which the frequency distribution was made.

Sales Group (Rs.)	Number of Sales Zones $f$	Midpoint of the group (Rs.) $X$	$fX$ $X$
1,100—4,500	1	2,800	2,800
4,600— 8,000	4	6,300	25,200
8,100—11,500	10	9,800	98,000
11,600—15,000	11	13,300	1,46,300
15,100—18,500	6	16,800	1,00,800
18,600 —22,000	4	20,300	81,200
22,100—30,000	2	26,050	52,100
30,100—38,000	2	34,050	68,100
	40		5,74,500

$$\bar{X} = \frac{\sum fX}{\sum f (= N)} = \frac{574500}{40} = \text{Rs. } 14,362.50$$

Thus the average sale of the 40 sale zones is Rs. 14,362.50. It is obvious that for finding out the mean of grouped data, we have first to find out the mid value with its corresponding frequency, sum up these products and divide it by the sum 'of frequencies. Whenever the two values for  $\bar{X}$  do not agree, it is due to inadequacy of the mid-value assumptions. "It is almost always true' that *none of the mid values is actually the true concentration point of its class*. For groups to the left of the group of maximum frequency the mid value of the group to the right of the group of maximum, frequency, the mid-value of the a group frequently exceeds the mean of the group. Although all the mid value assumptions, are usually incorrect, there is a definite tendency for the error to offset each other provided the distribution is approximately symmetrical."

### Shortcut method for computing mean

Let us now compute the value of the mean with the help of an assumed mean.

$$\text{Mean } i \dots\dots \bar{X} = \bar{X}_a + \frac{\sum fD}{N}$$

Where  $\bar{X}_a$  stands for the assumed mean D stands for the deviations of the mid values, from the assumed mean

$$i.e. = X - \bar{X}_a$$

$$N = \sum f.$$

Let us assume the value of fee mean to be Rs. 13,000, the mid value with the maximum frequency, and calculate the mean, as shown in work sheet No. 3.

**Worksheet No.3 for Calculating the mean (by Short-Cut Method)**

Sales group (Rs)	Mid-value of the group (X)	Frequency (f)	Deviation from $\bar{X}_a = 13,300$ (D)	$f \times D$
1,100-4,500	2,800	1	—10,500	—10,500
4,600—8,000	6,300	4	—7,000	—28,900
8,100—11,500	9,800	10	—3,500	—55,000
11,600—15,000	13,300	11	0	0
15,100—18,500	16,800	6	+ 3,400	+ 21,000
18,600—22,000	20,000	4	+ 7,000	+ 28,000
22,100—30,000	26,050	2	+ 12,750	+ 25,500
30,000— 38,00	38,050	2	+ 20,750	+ 41,500
		40		—73,500
				+ 116,000
				+ 42,500

\* FE. Croxton, D.J. Cowden and Sidney Klein : Applied General Statistics. Third Edition. 1971. p. 159.

$$\begin{aligned}
 X &= X_a + \frac{\Sigma fD}{N} \\
 &= 13,300 + \frac{42,500}{40} = 1,300 + 1062.50 \\
 &= \text{Rs. } 14,362.50 \text{ (the same as found out with the direct, method)}
 \end{aligned}$$

Another short-cut method for computing the mean, goes for further simplification: by dividing the deviation (D's) by the highest common factor (i); among the class interval sizes,  $i$  would equal the common class interval when, the size of all class intervals in a frequency distribution is uniform. However, when the size of class intervals varies, it will stand for the highest common factor, among, them. Let us show its working sheet No.4.

Here,

$$\bar{X} = \bar{X}_a + \frac{\Sigma fD}{N} \times i \quad \dots\dots\dots (4.4)$$

Where  $D' = \frac{D}{i}$  stands for deviation of mid values from the assumed mean in term of  $i$ .

Worksheet No. 4 for Calculating the Mean

Sales group (Rs)	Mid-values of the group (X)	Frequency (f)	Deviation $D = (X - \bar{X}_a)$	$D' = \frac{D}{i}$ where $i = 250$	$fD'$
1,100—4,500	2,800	1	— 10,500	— 42	— 42
4,600—8,000	6,300	4	— 700	— 28	— 112
8,100—11,500	9,800	10	— 35,000	— 14	— 130
11,600—15,000	13,300	11	0	0	0
15,100—18,500	16,800	6	+ 35,000	+ 14	+ 84
18,600—22,000	20,300	4	+ 7,000	+ 28	+ 112
22,100—30,000	26,050	2	+ 12,750	+ 51	+ 102
30,100—38,000	34,050	2	+ 20,750	+ 83	+ 166
		40			— 294
					+ 464
					+ 170

Now 
$$\bar{X} = \bar{X}_a + \frac{\sum fD}{N} \times i$$

$$= 13,000 + \frac{170}{40} \times 250 + 13,300 + 1062.50$$

$$= \text{Rs. } 14362.50$$

Since when classes vary in width, the distribution is invariably skewed, and as skewness, increases our mid value assumptions offset each other *less closely* and therefore, the mean worked out from a frequency distribution with unequal class intervals may differ markedly from the mean computed from the unclassified data.

### 1.3. Prostrates of the Arithmetic Mean

1. An important property of the mean is that is algebraic some of the deviations of fee various values from the mean is equal zero, *i.e.*

$$\sum d = 0$$

Where  $d$  stand for deviations of the values from the mean =  $X - \bar{X}$

2. The sum of the square of fee deviations of a set of number  $X$ , from any number  $a$  is a minimum if and only if  $a = \bar{X}$ .
3. If  $f_1$  numbers have, mean  $m_1$ ,  $f_2$  numbers have mean  $m_2$ ,..... $f_k$  numbers have mean  $m_k$ , then the mean of all the numbers combined is given by

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \dots + f_k m_k}{f_1 + f_2 + \dots + f_k}$$

## 1.4 The Median

1.3.1 An important characteristic of the mean is that it is affected by all the values, especially by extreme values, Assume 5 groups with the following income:

Rs. 20,000, Rs. 25,000, Rs. 27,000, Rs. 28,000, Rs. 1,50,000.

The mean is

$$\bar{X} = 1/5 (20,000 + 25,000 + 27,000 + 28,000 + 1,50,000) = \text{Rs. } 50,000.$$

The mean of the first four incomes is Rs. 25,000, but with the inclusion of the extremely high income the mean jumps sharply to Rs. 50,000. It is obvious, that this mean of Rs. 50,000 does not adequately represent values in the frequency, distribution.

In such cases where the frequency distribution is skewed and has extreme values a measure of central location or tendency called the *median*, is in many, cases, more appropriate. Median is affected by the position, but not by the size of the Items.

The *median is positional* measure of central tendency; it divides the series into two equal halves when presented as an array. The median is easily found out by arranging the data in the form of an array and locating the observation, that has just as many observations above it, as below it, if the number of observations in the series is odd, e.g. 37 or 75, then the Median Value of  $\frac{n+1}{2}$ th observation.

If the number of observations in the series is an even number there, will be no one value that divides them into two equal parts. In such a case, the median = arithmetic mean of the two middle values  $\frac{N}{2}$ th and  $\left(\frac{N}{2} + 1\right)$ th. and the median value, may not coincide with an actual value in the series.

Before going on to consider the computation of the median for grouped, data, let us calculate the value of the median for the, sales of the 40 Divisional Sales Managers HIMGU Products arrayed in Table 1.1: We want to find the value which is so located that  $\frac{40}{2} = 20$  items will be on either side of it. This is, of course, the value of the arithmetic mean of the 20th and 21st items and counting from either side reveals that the value of the median is 12,950.

### The Median From Grouped Data

For computing the value of the median for a frequency distribution, we count half, of the frequencies from either end of the distribution, in order to determine the value, on either side of which half of the frequencies fall.

We locate the median, by interpolation, with the help of the following formula:

$$\text{Median} = l_1 + \frac{i}{f_{\text{median}}} \left( \frac{N+1}{2} - cf_1 \right) \quad \dots\dots (5.1)$$

where  $l_1$  = lower class boundary of median class (i.e. the class containing the median).

$i$  = size of the median class interval.

$f_{\text{median}}$  = frequency of median class.

$N$  = Number of items in the data ( $=\Sigma f$ );

$Cf_1$  = cumulative frequency of the group preceding the median class.

Geometrically, the median is the value of  $X$ . (abscissa), corresponding to that vertical line which divides a histogram of frequency curve into two parts having equal areas. This value of  $X$  is sometimes denoted by  $X_m$ .

Let us now compute the value of the median for our sales in 40 zones.

**Worksheet 9, for computing the value of the median**

Sales Group (Rs.) (1)	Sales Group class Boundaries (2)	Sale Zones Frequency $f$ (3)	Cumulative frequency (4)
1,100—4,500	7,050—4,550	1	1
4,600—8,800	4,550—8,050	4	5
8,100—11,500	8,650—11,550	10	15
11,500—15,000	11,550—15,050	11	26
15,100—18,500	15,050—18,450	6	32
18,600—22,000	18,550—22,050	4	36
22,100—30,000	22,050—30,050	2	38
30,100—38,000	30,050—38,050	2	40
		40	

$$\text{Median Sales Zone} = \frac{40+1}{2} = 20.5\text{th zone.}$$

If we take column 2, 3, and 4 together; we can describe the frequency table as under :

In one zone, sales are less than Rs. 4,550 ; in 5 zones, sales are below Rs. 8,050 ; in 15 zones are below 11,550; The fourth sales group Rs. 11,550 to 15,050 has 11 sales zones, starting from 16th to 26th. Hence, our median sales zone i.e. 20.5th. zone lies in this group. Having located the median class, we can now very easily calculate 'the value of the median, by inter-polation.

$$\text{Median} = l_1 + \frac{i}{f_{\text{median}}} \left( \frac{N+1}{2} - cf_1 \right)$$

$$\text{Now } l_1 = 11,550, \quad i = 3,500, \quad f_{\text{median}} = 11$$

$$\frac{N+1}{2} = \frac{40+1}{2} = 20.5 \text{ and } Cf_1 = 15$$

$$\begin{aligned} \text{Hence Median} &= 11,550 + \frac{3,500}{11} (20.5 - 15) \\ &= 11,350 + 1750 = \text{Rs. } 13,300. \end{aligned}$$

We would get exactly the same result if we start from the other end of the median group.

$$\text{Median} = l_2 - \frac{i}{f_{\text{median}}} \left( cf_m \frac{N+1}{2} \right) \quad \dots\dots (5.2)$$

where  $l_2$  = upper real limit of the median group

and  $Cf_m$  = the cumulative frequency of the median group.

$$\text{Hence Median} = 15,050 - \frac{3.500}{11} (26 - 22.5)$$

$$= 15,050 - 1750$$

$$= \text{Rs. } 13,300 \text{ (the same as found with the help of formula 5.1)}$$

The value of the median obtained from the frequency distribution Rs. 13,300 is in fairly close agreement with that of 12,950.00 found from the array. “Unless the data contains gaps or irregularities; we can expect rather close agreement when dealing with, a continuous variable, and likewise for a discrete variable if the data are not broken.”

We have now computed the value of the mean and the median for the frequency distribution of zonal sales of HIMCU Products. The mean was Rs. 13,362.40 and the median Rs. 13,300. The mean exceeds the median because the distribution is skewed to the right. If a distribution is symmetrical, the mean and median are equal. The median is not affected by the presence of unequal or open-end-class intervals.

The median of a frequency distribution can also be located graphically, with the help of the ogive in the following way:

- (a) Compute  $\frac{N+1}{2}$  and locate, this point on the vertical scale.
- (b) At this point draw a perpendicular to the Y = axis and extend it, so that it intersects the ogive.
- (c) Drop a perpendicular to the X-axis from the point of intersection and read the value of the median on the X-axis'. The value of the median, located graphically, will be approximately equal to the value computed arithmetically.

#### 4. The Quartiles, Quintiles, Deciles and Percentiles

The median, we have, seen, divides a set of data into two equal parts. An extension of this idea is dividing the set into four (quartiles), five (quintiles), ten (deciles) or hundred (percentiles) equal parts.

These values of *quartiles*, denoted by  $Q_1$ ,  $Q_2$  and  $Q_3$  are called the first (or lower), the second, (or middle), and third (or upper) *qualities* respectively, the value  $Q_2$  being equal to the median.

Similarly, the *deciles* are denoted by  $D_1, D_2, \dots, D_9$ , and the *percentiles* are denoted by  $P_1, P_2, \dots, P_{99}$ .

$$D_5 = P_{10} = Q_2 = \text{Median.}$$

Similarly,  $P_{25}$  and  $P_{75}$  correspond to the lower and upper quartiles respectively.

Collectively quartiles, deciles, percentiles and other values obtained by equal subdivision of the data are called *quintiles*, Quintiles are computed in the same way as the median,

1. We first divide the total number of observations by, the number of equal parts into which the series is to be divided, by 4, 5, 10 or 100 for quartiles, deciles and percentiles respectively.’
2. Multiply by the order of this quantile *i.e.* by 1 in case, of first quartile, decile etc. by 3 in case of third quartile, by 7 in case of 7th decile or percentile; and by 56 in case of  $P_{56}$ .
3. Determine with the help of cumulative frequency column the class interval in which our positional measure lies, and
4. Interpolate the value of the quartile, in the same way as the median.

Let us compute the value of some, of these positional measures, for the data on zonal sales, with the help of worksheet No 6 (given earlier).

$$Q_1 = l_1 + \frac{i}{fq_1} \left( \frac{N}{4} - Cf_1 \right) \quad \dots\dots\dots (6)$$

where  $l_1$  = lower limit of the first quartile group.

$i$  = the size of the first quartile class interval.

$fq_1$  = frequency of the first quartile group.

$Cf_1$  = cumulative frequency of the group preceding the first quartile group.

Now, for the zonal sales data

The 10th zonal sales lie, in the group, 8,100-11,500 (where 6th to 15th zonal sales lie) whose *real* ; class limits are 8,050 and 11,550.

Hence  $l_1$  = 8,050,  $i$  = 3,500  $fq_1$  = 10.

and  $Cf_1$  = 5.

$$\begin{aligned} \text{Therefore, } Q_1 &= 8,050 + \frac{3,500}{10} (10 - 5) \\ &= 8,050 + 1,750 = \text{Rs. } 6,800. \end{aligned}$$

Similarly;

$$Q_3 = l_1 + \frac{i}{fq_3} \left( \frac{3N}{4} - Cf_1 \right) \quad \dots\dots\dots (7)$$

Here we first locate  $\frac{3N}{4} = \frac{3+40}{4} = 30$  zonal sales and we find that it lies in the group, 15,100—18,500 whose *real* class limits are 15,050 and 18,550.

Hence  $l_1$  = lower boundary of the third quartile group = Rs. 15,050

$i$  = 3,500,  $fq_2$  = 6, and  $Cf_1$  = 26.

$$\begin{aligned} \text{Therefore, } Q_3 &= 15,050 + \frac{3,500}{6} (30 - 26) \\ &= 15,050 + 2,333.33 \\ &= \text{Rs. } 17,383.33 \end{aligned}$$

---

\*Ibid p.-866



We may also, similarly, compute the value of 7th decile ( $D_7$ ) and 56 percentile ( $P_{56}$ ).

$$D_7 = l_1 + \frac{i}{fD_7} \left( \frac{7N}{10} - Cf_1 \right) \quad \dots\dots\dots (8)$$

The 7th decile item i.e.  $\frac{7N}{10}$  th = 28th item lies in the group 15,100 — 18,500, whose class boundaries are 15,050 and 18,550 (the same as for  $Q_3$ ).

Now  $l_1 = 15,050$ ,  $l = 3,500$ ,  $fD_7 = 6$ ,  $Cf_1 = 25$ .

Hence

$$\begin{aligned} D_7 &= 15,050 + \frac{3,500}{6} (28 - 26) \\ &= 15,050 + 1,157.14 \\ &= \text{Rs. } 16,207.14 \end{aligned}$$

$$P_{56} = l_1 + \frac{i}{f_{P_{56}}} \left( \frac{56N}{100} - Cf_1 \right) \quad \dots\dots\dots (9)$$

$\frac{56N}{100} = \frac{56 \times 40}{100} = 22.4$ th item lies in the group, 11,600 — 15,000, with class boundaries as 11,550 and 15,050 (the same as median group).

Now  $l_1 = 11,550$ ,  $i = 3,500$ ,  $f_{P_{56}} = 11$ ,  $Cf_1 = 15$ .

Hence.

$$\begin{aligned} -P_{54} &= 81,550 + \frac{3,500}{11} (22.4 - 15) \\ &= 11,550 \times 2,354.54 = \text{Rs. } 13,904.54. \end{aligned}$$

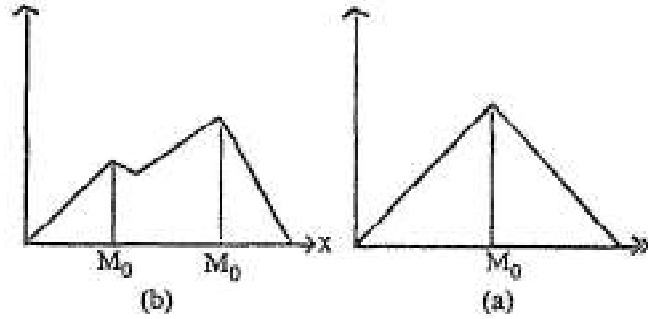
The quartiles can also be located on an ogive in the same way as the median.

## 5. The Mode

Another measure of location of a frequency distribution is the mode. The *mode* of a distribution is any value at which the frequency density is at a maximum. This implies that it is any value of the variable that occurs most frequently. In simple terms we may say that mode is the most popular value of the variable, i.e. a value which occurs maximum number of times, it further implies that if the frequency curve has one peak (i.e.) one maximum as in the figure 4.1 (a), there is only one mode, whereas if the frequency curve has, two (or more) peaks (i.e. two or more maxima), as in the figure 4.1 (b) the distribution has two (or more) modes. On the other hand, if we have a rectangular distribution\* there is no mode,

---

\*If a rectangular distributor, is plotted on a graph paper, the curve is a straight line parallel to X—axis.



**Fig. 1.1**

**Example 4.** Let us suppose a student took 5 tests in a semester with the results :

58, 70, 70, 75, 85.

Then the mode is there but not so well defined 70 occurs twice while all other scores occur only once. If, however, the scores are:

59, 70, 73, 75, 85, there is one mode.

Since the mode is the most typical value of a series of values, it is not at all affected by the presence of one or a few extreme values. Further, the mode cannot be readily located in an ungrouped data or even an array.

For grouped data, the mode may be located more readily. While there are several ways of computing the mode, it is usually sufficient for practical purposes to use the midpoint to the *modal group*. In some cases, it may suffice to say that the mode lies between such and such limits. If, however, we have to determine the modal value of the point of maximum concentration, this can be done with the following formula:

$$M_0 = l_1 + \frac{f_m + f_1}{(f_m - f_1) + (f_m - f_2)} \times i \quad (...10)$$

where  $l_1$  = lower limit of the modal group.

$f_m$  = modal or maximum frequency.

$f_1$  = frequency of the group preceding the modal group.

$f_2$  = frequency of the group following the modal group.

$i$  = size of the modal class interval.

For our zonal sales data, the modal group will be 11,600—15,000, since it has the highest frequency (11). The class boundaries of this group are 11,550 and 15,050.

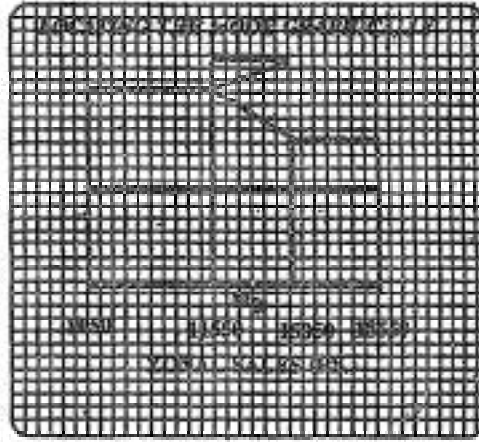
Hence  $l_1 = 11,550$ ,  $i = 3,500$ ,  $f_m = 11$ ,  $f_1 = 10$ ,  $f_2 = 6$ .

$$\text{Therefore, } M_0 = 11,550 + \frac{11-10}{(11-10) + (11-6)} \times 3,500.$$

$$= 11,550 + \frac{1}{6} \times 3,500, = 11,550 \times 583.83$$

$$= \text{Rs. } 12,133.33.$$

The interpolation of mode shown above can also be located graphically, as in fig. 1.2.



The mode is easy to compute and may be applied to qualitative as well as quantitative data. One may be investigating, for example, consumer preferences for six brands of tea, A, B, C, D, E, and F. Let the preferences be

Here the modal preference is tea D. Suppose a garments store wishes to stock men's shoes. An investigation shows that size 6 has the greatest demand. This is the modal value of the distribution of shoe sizes.

## 6. Comparison of the Mean, Median, and Mode

When a distribution is symmetrical, the mean, median and mode coincide. When a distribution is skewed to the right, then Fig. 1.3 (b). Mean > Mode.

For example, income distribution is frequently skewed to the right, where the majority of families have incomes concentrated at lower levels and then the number of families tapers off as the income goes up. In this case, the mean is pulled up by the extreme high incomes and the relation among the three measures is as stated above *i.e.*

17



$$\bar{X} = \frac{\Sigma X}{N}; \Sigma X = N \cdot \bar{X} \quad \text{and} \quad N = \frac{\Sigma X}{\bar{X}}.$$

Further, if we know a series of arithmetic means, we can compute the arithmetic mean of the combined series, using appropriate weights.

Thus  $NX = N_1X_1 + N_2X_2 + N_3X_3 + \dots + N_nX_n$

$$\text{or} \quad \bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 + \dots + N_n\bar{X}_n}{N = (N_1 + N_2 + N_n)}$$

- (iv) The arithmetic mean may be calculated from ungrouped data, from arrayed data, from the frequency distribution or as noted in (iii) above, merely, from our knowledge of  $\Sigma X$  and  $N$ . Further, the mean from the grouped data tends closely to approximate that from ungrouped data, the closeness being greater, the more symmetrical is the distribution.

Median cannot be computed from ungrouped data, without first being put in the form of an array. Its value, as computed from a frequency distribution, will agree approximately with that from a ray, if the items in the *median class* are distributed regularly.

The mode is easily located in a frequency distribution, but with some difficulty in an array. Further, the interpolation of the mode in the modal class is at best only an approximation.

- (v) When *skewness is present* in a distribution, the *class intervals*, generally, *cannot be of uniform size*. Thus, lack of uniformity in the size of class intervals tends to make the mean from the grouped data deviate from that computed from ungrouped data. Since most of the frequency distributions are skewed to the right, the value of  $\bar{X}$  from these distributions, tends to exceed that from ungrouped data.

The median generally remains unaffected by the variation in the width of the class interval. The upper quartile and the higher deciles and percentiles may, however, get affected by this phenomenon, making the values concerned, less reliable.

The mode can be located fairly satisfactory, if the class intervals, following and preceding the modal class, have the same width as that of the modal group. Otherwise, modal's value becomes less reliable:

As stated earlier, the *presence of skewness affects the mean, the most, and the median*, less. The mode is hardly affected by it. Further the skewness pulls the mean and median in the direction of the excess tail *i.e.*, of skewness.

- (vi) The existence of a few *extreme values* affects the mean greatly, the median is only (slightly affected and the mode is not at all influenced. Because of this characteristic, mean is not a suitable measure of central tendency for income distribution, especially if a few persons, at the top have incomes far exceeding those of most people.

Similarly, wherever we have reason to suspect heterogeneity in data, the median shall be preferable to the mean.

Again, whenever we have a series in which we know the number, but not the exact value of the extreme items, the median and the mode can be determined, but 'not the mean.'

- (vii) The presence of *open end classes* makes the value of mean, largely conjectural because the mid value of these classes cannot be determined.

Since these open-end classes occur at the beginning or at the end of a frequency distribution they hardly affect the median.

Open-end classes do not affect the determination of the mode, unless the distribution is of J-type or J-reverse type in which case mode cease to be a measure of *central tendency*.

- (viii) The arithmetic mean is hardly a suitable measure of central tendency for a frequency distribution if the data, are irregular or broken, because it would tend to deviate for from the mean from ungrouped data! The median, for its location required that data to be regularly distributed over the median class; hence, if the class in which median lies, has gaps or the data are irregularly distributed median would not be an appropriate measure of central tendency.

If the data around the mode has gaps, the mode is not likely to be well defined. But irregularity of gaps elsewhere do not affect the mode.

As described above, the choice of the measure of central tendency depends upon (a) the nature of the distribution of the data and (b) the concept of central tendency, desired for a particular purpose.

With symmetrical or near symmetrical distributions any of the three measures may be used. With skewed distributions, the mean, not being a typical value, the median or the mode are to be preferred.

Other things being, equal the mean is preferable to other measures.

Besides the three measures of central tendency, the mean, the median, and the mode, we use two other measures occasionally in business and economics. These are the geometric mean and the harmonic mean. Of these, *the geometric, mean* is more, important and is used *for averaging rates of change* and constructing index numbers.

### The Geometric Mean

The geometric, mean (G) may be defined as the  $n$ th root of the product of  $n$  items *i.e.*

$$\begin{aligned} G &= n \sqrt[n]{(X_1 X_2 X_3 \dots X_n)} \\ &= (X_1 X_2 X_3 \dots X_n)^{1/n} \end{aligned} \quad \text{..... (11)}$$

Logarithms, are used to calculate the  $n$ th root

$$\text{Thus } \log G = \frac{1}{n} [\log X_1 + \log X_2 + \dots + \log X_n]$$

$$\text{or } \log G = \frac{1}{n} \Sigma \log X \quad \text{..... (12.1)}$$

$$\text{or } G = \text{Antilog} \left\{ \frac{1}{n} \Sigma \log X \right\} \quad \text{..... (12.2)}$$

For frequency distributions, each logarithm must be multiplied by the corresponding frequency so that

$$\text{Log } G = \frac{f_1 \log X_1 + f_2 \log X_2 + \dots + f_n \log X_n}{N (= \Sigma f)}$$

$$= \frac{\Sigma f \log X}{N} \dots\dots\dots (13)$$

Geometric mean is especially useful in the construction of index number it is an average most suitable when large weights have to be given to small values of observations and small weights to large, value of observations.

Here, as in the case of the arithmetic mean, we first find out the mid values of the classes look up the logarithms of these mid values, multiply each logarithmic mid value by the corresponding frequency, sum up these and divide by the number of items and then find but anti-logarithm of this quotient.

Thus log of geometric mean is equal to the arithmetic mean of the log, *values of* the variate in the unit disrate (frequency distribution or weighted mean incase of grouped frequency distribution.

**Application of Geometric Mean: Average Rates of change and the Compound Interest Formula.**

Let us assume that the output of an industry increased 20 percent in the first year, 40 percent in the second year, 30 percent in the third year and 10 percent in the fourth” year.

Then the index of industrial output will be:

At the end of	zero year	:	100	
	1st year	:	120	(20% increase)
	2nd year	:	168	(40% increase)
	3rd year	:	218.4	(30% increase)
	4th year	:	240. 2	(10% increase)

What is the average rate of increase during, these four years ? We can clearly see that the output in 1st year was 1.20, times that of the zero year, that in the 2nd year 1.40 times of that of the 1st year, and so on. The geometric mean (G), would show the average rate of increase.

Thus  $G = \sqrt[4]{(1.20 \times 1.40 \times 1.30 \times 1.10)}$

or  $\log G = \frac{1}{4} (\log 1.20 + \log 1.40 + \log 1.30 + \log 1.10)$

$$= \frac{1}{4} (0.0792 + 0.1461 + 0.1139 + 0.0414)$$

$$= \frac{1}{4} (0.3806) = 0.09515.$$

Hence  $G = \text{Anti log } (0.09515)$   
 $= 245$

This implies that the average rate of change of output over 4 years was 24.5 percent  $(1.245 \times 100 - 100)$ .

If  $r$  is average rate of change, then

$$P_n = P_0 (1+r)^n \dots\dots\dots (14.1)$$

where  $P_0$  is initial and  $P_n$ , the figure after  $n$  years of change. This is the familiar compound interest formula.

This may also be written as.

$$r = \left( \frac{P_n}{P_0} \right)^{1/n} - 1 \quad \text{..... (14.2)}$$

**Example 5.** India's net national product was Rs. 13,294 crores in 1960-61; it rose to Rs. 18,755 crores in 1970-71, at 1960-61 prices. What is the average rate of change over the decade?

Using formula (14.1) and (14.2), we have

$$18,755 = 13,294 (1 + r)^{10}$$

$$\text{or} \quad r = \left( \frac{18755}{13294} \right)^{1/10} - 1$$

$$\text{i.e. } 1 + r = \left( \frac{18755}{13294} \right)^{1/10}$$

Applying logarithms, we get

$$\begin{aligned} \log(1 + r) &= \frac{1}{10} (\log 18755 - \log 13294) \\ &= \frac{1}{10} (4.2731 - 4.1237) \\ &= \frac{1}{10} (0.1494) \\ &= 0.01494 \end{aligned}$$

Hence,  $1 + r = \text{Anti log } (0.01494)$

$$= 1.035$$

$$r = 0.035 - 1$$

$$= 0.035 \text{ or } 3.5 \text{ percent}$$

Thus India's N.N.R on the average, increased at a compound rate of 3.5 percent per annum over, the decade 1960-61 to 1970-71.

Another use of the geometric mean, through compound interest formula is *in discounting and capitalization*.

**Example 6:** Suppose, we have to choose between producing' electricity through a hydel project or a thermal project. It is estimated that the cost of construction of the hydel project would, be Rs.100 crores and its life would be 80 years and that of the thermal project. Rs. 40 crores and its life would be 25 years. Both Projects are expected to yield an annual net income of Rs. 10 crores. Which of the projects should be chosen?

For comparing the relative net returns versus costs of the two Projects, we shall have to compute the present value of returns of both projects and compare these to the costs. Present value can be found out by discounting the future returns at, the rate of 'discount,  $r$  (let us say, 5 percent (and by using the compound interest formula:



$$P_0 = \sum \frac{P_n}{(1+r)^n} \quad \text{..... (14.3)}$$

$$\begin{aligned} P_{0\text{hydro}} &= \frac{10}{(1+r)} + \frac{10}{(1+r)^2} + \frac{10}{(1+r)^3} + \dots + \frac{10}{(1+r)^{80}} \text{ crores} \\ &= 10 \left[ \frac{1}{1+r} + \frac{1}{(1+r)^2} + \frac{1}{(1+r)^3} + \dots + \frac{1}{(1+r)^{80}} \right] \text{ crores} \end{aligned}$$

Now applying the formula for the sum of a geometric series we have, the sum of the present value of net returns over 80 years\*

$$\begin{aligned} &= 10 \left\{ \frac{\frac{1}{1+r} \times \left[ 1 - \left( \frac{1}{1+r} \right)^{80} \right]}{1 - \frac{1}{1+r}} \right\} = 10 \left\{ \frac{\frac{1}{1+r} \times \left[ 1 - \left( \frac{1}{1+r} \right)^{80} \right]}{\frac{1+r-1}{1+r}} \right\} \\ &= 10 \left\{ \frac{1 - \left( \frac{1}{1+r} \right)^{80}}{r} \right\} \\ &= \frac{10}{r} \left\{ 1 - \left( \frac{1}{1+r} \right)^{80} \right\} = \frac{10}{0.00} \left\{ 1 - \left( \frac{1}{1+0.05} \right)^{80} \right\} \\ &= 200 \left\{ 1 - \frac{1}{1.05^{80}} \right\} = 200 - 200 (1.05)^{-80} \end{aligned}$$

Let  $x = (1.05)^{-80}$

Hence log  $x = -80 \log 1.05$   
 $= -80(0.002166)$   
 $= -0.173280$   
 $= 1.82672$

$x = \text{Anti-log } 1.82672 = 0.671$

$P_{0\text{hydro}} = 200 - 200 (1.05)^{-80}$ , therefore, becomes.  
 $= 203 - 200 (0.671) = 200 - 134.2$   
 $= \text{Rs. } 65.8 \text{ crores.}$

---

\* $S_n = \frac{a(1-r)^n}{1-r}$  where  $S_n$  is the sum of a geometric series for  $n$  terms,  $a$  is the first terms of the series, and  $r$  is the common ratio  $[1/(1+r)]$  in our present case].

Applying the above formula, to the thermal projects we get

$$P_{0\text{thermal}} = 200 [1-(1.05)^{-25}]$$

Now let  $x = (1.05)^{-25}$

then  $\log x = -25 \log 1.05$   
 $= -25 (0.002166)$   
 $= -0.054150$   
 $= \bar{1}.94585$

Therefore,  $x = \text{Anti-log} (\bar{1}.94585)$   
 $= 0.88277,$

Hence  $P_{0\text{therma}} = 200 (1-0.88277)$   
 $= 200 (0.11723) = \text{Rs. } 23,446 \text{ crores.}$

Now Returns/Cost ratios for the two projects, will be:

- (i) Returns/Cost ratio for Hydro Project =  $65.8/100 = 0.658$ .
- (ii) Returns/Cost ratio for me Thermal Project =  $23.44/40 = 0.586 \text{ app.}$

Clearly, at a rate of discount of 5 percent, the hydro project would yield a higher net return/cost ratio, and maybe preferable.

The use of geometric mean; for averaging, index number will be studied, when we take up the study of Indian price index numbers. The geometric mean is preferred for this purpose, because its assigns equal weight to equal ratio of change.

The geometric mean is, very useful in averaging ratios and percentages. It also helps in determining the rates of increase and disease: It is also capable of further, algebraic treatment so that a combined geometric; mean can easily be computed. Geometric mean cannot be computed if any observation has either a value zero or negative.

### The Harmonic Mean

The Harmonic mean H is the reciprocal of the arithmetic mean of the reciprocal of the values. The harmonic mean is a measure of central tendency for the data expressed, as rates such as kilometers per hour, tones per day, kilo meters per liter etc.

Thus

$$H = \frac{1}{\frac{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}}{N}}$$

$$= \frac{1}{\frac{\sum \frac{1}{X}}{N}} \quad H = \frac{N}{\sum \frac{1}{X}} \quad \dots\dots\dots (15.1)$$

For purpose of computation, (15.1) may be put as

$$\frac{1}{H} = \frac{\sum \frac{1}{X}}{N} \quad \dots (15.2)$$

Having found the value of  $\frac{1}{H}$ , we can get H; as its reciprocal.

**Example 7.** Assume that me 3 grades of, Golden delicious apples are being quoted in Shimla Market as follows:

**Large :** 4 per rupee; **Medium :** 6 per rupee and **small :** 11 per rupee.

The arithmetic mean  $\bar{X} = \frac{4+6+11}{3} = 7$  per rupee i.e 14.3 paise per apple.

This is the price we must pay *if we spend equal amounts of money for each grade*. Paying 14.3 p. for each of the 21 apples we shall spend Rs. 3/- for the lot.

The harmonic mean gives different results.

$$= \frac{1}{\frac{\sum \frac{1}{X}}{N}} H = \frac{N}{\sum \frac{1}{X}}$$

= 5.91 app. for Re 1/- or 1692 paise per apples

This is the price we shall pay if *equal number of apples are bought at each price*.

The harmonic mean for frequency distributions is hardly ever used. In this case.

$$H = \frac{N = \sum f}{\sum \frac{f}{X}} \quad \dots (15.3)$$

$$\text{or} \quad \frac{1}{H} = \frac{\sum \frac{f}{X}}{N} \quad \dots (15.4)$$

The harmonic mean hardly adds much to the information furnished by the arithmetic mean. It may, however, be useful when, data are generally quoted in terms of problems solved per minute, miles or kilometers covered per hour, units sold per rupee etc.

If, in case of a highly skewed distribution, the plotting of the reciprocals of the data (or class mid values), results in an approximately normal curve, harmonic mean could be useful. But these instances are rather unusual. 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.

Another measure of central tendency, which is more, of theoretical rather than practical interest is the *root mean square* or the *quadratic mean*.

$$\text{Root mean Square (R.M.S) - } \sqrt{\left(\frac{\sum X^2}{N}\right)} \quad \dots (16)$$

This type of average is used sometimes in physical applications.

### What Have We Learned :

1. Mean, medium, mode, geometric mean and harmonic mesa are parameters which attempt to characterize the central tendency of the distribution.
2. (a) The arithmetic mean is the sum of all observations in the series divided by the number of observations.  
 (b) Formula for ungrouped data  

$$\bar{X} = \frac{\sum X}{N}$$
  
 (c) Formula for grouped data  
 (i) Long method:  $\bar{X} = \frac{\sum fX}{N}$ .  
 (ii) Short-cut method :  $\bar{X} = \bar{X}_a + \frac{\sum fD'}{N}$ .  
 (iii) Short-cut method (in terms of highest common factor) :  $\bar{X} = \bar{X}_a + \frac{\sum fD'}{N} \times i$ .
3. (a) The median is a positional measure of central tendency dividing the series when arrayed, into two equal halves.  
 (b) For grouped data, Median =  $l_1 + \frac{i}{f} \left( \frac{N+1}{2} - Cf_1 \right)$ .  
 (c) Quantiles divide a series when arrayed or a frequency distribution into  $n$  equal parts.
4. (a) The mode is the most frequent or typical value or a series. It cannot be generally easily located for ungrouped data.  
 (b)  $M_0 = l_1 + \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_1)} + i$   
 (a) For a symmetrical distribution,  
 Mean = Median = Mode.  
 (b) (i) For positively skewed distributions  
 Mean > Median > Mode.  
 (ii) When the distribution is slightly skewed  
 Mean - Mode = 3 (Mean - Median).  
 (c) The mean possesses a number of advantages over the median and mode, but is less suitable when data are irregular or broken, highly skewed, with unequal or open end class intervals.

- (a) The geometric mean is the  $n$ th root of the product of the value of  $n$  items. It is useful to measure the average rate of change, for discounting and capitalization, and for constructing price indices.

(b)  $G = \text{Anti} - \log \left( \frac{\sum \log X}{N} \right)$

or  $\text{Anti} - \log \left( \frac{\sum f \log X}{N} \right)$  for grouped data.

7. The harmonic mean is the reciprocal of the arithmetic means of the reciprocal, of the values

$$H = \frac{1}{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} \dots \frac{1}{X_N}}$$

### Suggested Readings

1. J. Kmenta : *Elements of Econometrics*, Macmillan, 1971.
2. R E. Croxton, D. Coriden and Klein : *Applied General Statistics*, Pentice Hall, 1967.
3. T. Yamane : *Statistics : An Introductory Analysis*, Harper and Row, 1973.

\*\*\*\*\*

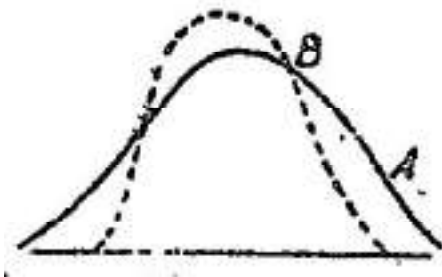
## LESSON-2

### MEASURES OF DISPERSION AND SKEWNESS

**Dear Student,**

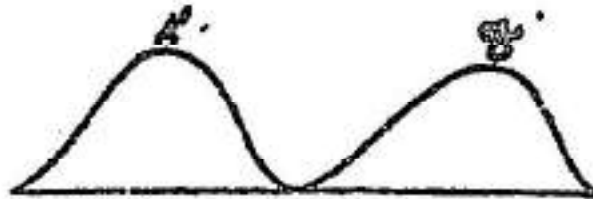
In the previous lesson, we were concerned with various measures that are used to provide a single, representative value of a given set of data. This single value alone cannot adequately describe a set of data. Therefore, in the present lesson we shall study two more important characteristics of a distribution. First we shall discuss the concept of dispersion-variation and later the concept of skewness.

**2.1.0. Dispersion:** Two groups of students Show' the same mean height, 64 inches. However, the height of the first group ranges from 58" to 74", whereas that of the other, lies between 60" and 69". It is obvious, 'that the heights of the first group are more scattered; whereas the second group shows a greater uniformity. In fig 2.1 (a), we have shown frequency curve A, with a greater spread or dispersion than frequency curve B, though both have the same mean. Fig. 2.1 (b) shows, on the other hand, two frequency curves having different means, but having, the same dispersion.



**Fig. 2.1 (a)**

Two Frequency Curves A & B,  
with different Means but same.



**Fig. 2.1 (b)**

Two Frequency Curves A & B  
with Same mean, but Different

#### Dispersions

When we know the dispersion of a variate besides its central tendency, we may speak with greater confidence, about the dependability of the mean. You have all, I believe heard the story of a man who on an enquiry found that the mean or average depth of the river was 5 feet. Since he was 5' - 6" tall," he decided on the basis of this information that he could safely cross across the river. Unluckily he was drowned, because he did not know swimming and he met, on the way, depths over 12 feet had he know about the dispersion of the river depth, he would surely have thought Better, not to risk wading through the river.

Assume that two manufactures of fluorescent tubes claim an average life 2500 hrs. for their products, Further enquiry showed, that tubes; of company B lasted from 500 to 5,000 hours, whereas, those of company T, lasted a minimum of 1500 and a maximum of 3500 hours, it is obvious that products of company T have greater uniformity than those of B.

Dispersion, therefore refers to the variability in the size of items. It indicates that the size of items in a series is not uniform *i.e.*, the value of various items differs from each other. If the variation is substantial, dispersion is said to be considerable and if the variation is little, dispersion is insignificant.

The term dispersion not only gives a general impression about the variability of a series, but also a precise measure of this variation. Generally in a precise study of dispersion, the deviations of size of items from a measure of central tendency are found out and then these deviations are averaged to give single figure representing the dispersion of the series. This figure, can be compared with, similar figure representing other series, and this it is possible to make a comparison between the averages and the dispersion of two or more series.

### **Significance of Measuring Variation**

Measuring Variation is significant for some of the following purposes.

- (i) Measuring variability, determines the reliability of an average by pointing out as to how far an average is representative of the entire data.
- (ii) Another purpose of measuring variability is to determine the nature and cause of variation in order to control the variation itself.
- (iii) Measures of variation enable comparisons of two or more distributions with regard to their variability.
- (iv) Measuring variability is of great importance to advanced statistical analysis. Sampling and statistical inference is essentially a problem in measuring variability.

### **Absolute and Relative dispersion :**

If we calculate dispersion of a series relating to the income of a group of person in absolute figures, it will be expressed in the unit in which the original data or say Rupees. Thus when we say that in income of a group of persons is Rs. 120 p.m. and the dispersion is Rs. 30. This is called Absolute Dispersion.

When dispersion is, measured as a percentage or ratio of the average it is called Relative Dispersion. It is not expressed in the unit of original data, in the above example the average income would be Rs. 30 p.m. and the Relative dispersion  $30/120$  or 25.00 percent.

#### **2.1.1 Measures of Absolute Dispersion**

We have several measures of dispersion available for use for different purposes. The most important of these are:

(i) The range, (ii) The 10-90 Percentile Range, (iii) The Quartile Deviation, (iv) The Mean Deviation, and (v) The Standard deviation.

#### **The Range**

This is the simplest but a crude measure of dispersion. It is the difference between the values at the extreme items of a series. Suppose we are told that the average grade of two groups of students, A and B is 65 points. However, the highest and lowest grades of group A are 90 and 25 points respectively whereas those of group B, range from 60 to 70 points *i.e.* only 10 points.

Now the range of Group A  $90 - 25 = 65$  points and that of Group B  $70 - 60 = 10$  points. The range, as may be seen, tells us only about the two extreme values and we do not know anything more

about the rest of the data whether they are concentrated around the mean or scattered widely. It is unfit for purpose of comparison, if the distributions are in different units. If range is divided by the sum of the extreme items the resulting figure is called the “ratio” of the range” or the “coefficient of scatter”.

**2.1.2. The 10-90 Percentile Range:** This measure excludes the lower 10 percent and the upper. It percent of the items and concentrates attention on the middle 80 per cent. Thus it steers clear of the, extreme values. It, however, does not use the value of all the items. Its only concern is the values of  $P_{90}$  and  $P_{10}$ , and is not at all affected by the arrangement of values within and outside this range. It is obvious that the 10-90 percent tile range.

### 2.1.3. The Quartile Deviation or the Semi-Inter Quartile Range

This measure of dispersion, as should be obvious from its name, is based on the values of the lower and upper quartiles. It is given by

$$Q = \frac{Q_3 - Q_1}{2} \quad \text{..... (1)}$$

In a symmetrical series, the lower and upper quartiles lie at equal distances from the median, so that the median  $\pm Q$  “should en-compass 50 percent of the items. The *quartile deviation* like the 10-90. Percentile range is not affected by extreme values. However, this also fails to consider the values of all the items.

### 2.1.4. The mean of Average Deviate

A measure of dispersion that includes the variability of all the Stems is the mean deviation. It is the mean or average of absolute deviations from the mean of median (i.e. ignoring sign).

Mean Deviation

$$\text{or} \quad \delta \bar{x} = \frac{\sum |dx|}{N} \quad \text{..... (2.1)}$$

about mean

Where  $\delta \bar{x}$  is the mean deviation from the mean,  $dx$  are the deviations of the items forms  $\bar{x}$ , it is the symbol for ignoring sign. Since we are interested in the amount of variability, *i.e.* in the *distance of the deviations*, the minus “signs” are diregarded when finding the mean variability.

Mean Deviation

$$\text{or} \quad \delta \text{Med} = \frac{\sum |dx|}{N} \quad \text{..... (2.2)}$$

about Median

where  $|dx|$  = deviations of the  
values from the median  
(ignoring the signs).

Because the sum of the *absolute value* deviations is a minimum when taken round the median, sometimes, the mean deviation is computed *in* relation to *the* median. In practice, however, it is the mean which is generally used, and if the series is symmetrical, the resulting mean deviation will be the same; whether computed in relation to the mean or the median.



Mean deviation for a frequency distribution is

$$\delta \bar{x} = \frac{\sum f |dx|}{N} \quad \dots (2.3)$$

where  $\delta \bar{x}$  stands for the mean, deviation in relation to the mean.

$|dx|$  stands for deviations of the mid-points of class intervals, or the class average from the mean (signs ignored).

$f$  stands, for the frequency.

Similarly, the mean, deviation in relation to the median for grouped data, will be

$$\delta \text{ Med.} = \frac{\sum f |dx|}{N} \quad \dots (2.4)$$

where  $|dx|$  stands for *absolute value* of deviation around the median.

In case of normal distribution, 57.5 percent of the items lie within the range  $\bar{Y} \pm \delta \bar{Y}$ . Where the distribution is moderately skewed, this will be approximately true.

Let us compute the quartile deviation and the mean deviation for the monthly per capita consumer expenditure for rural India, as given in table below.

**Table 2.1 Consumer Expenditure in Rupees per person for 30 days by Monthly per capita. Expenditure Classes:**

Rural India						
Mid-value X	Monthly Per Capita Expenditure Classes in (Rs.)	Percentage of persons in the Expenditure Class (f)	Average Consumer Expenditure (Rs.)	Cumulative Percentage of person under upper limit of group	$ dx $ From Median	$f dx $
1	2	3	4	5	6	7
4	8.00	7.28	6/62	7.28	1243	90.49
9.5	8.11	13.79	9.72	21.07	6.93	95.56
12	11.13	10.95	12.02	32.07	4.43	48.51
14	13.15	11.12	13.94	43.14	2.43	27.02
16.5	15.18	14.43	16.50	57.57	0.07	1.01
19.5	18.21	10.69	19.72	68.26	3.07	32.82
22.5	21.24	7.3	22.51	76.09	6.07	47.55
26	24.28	6.72	25.98	82.81	9.57	64.52
31	28.34	8.09	30.09	90.90	14.57	117.31
38.5	34.43	4.68	38.37	95.58	22.07	103.29
49	43.55	2.32	50.74	97.30	32.57	75.56
*88-13	55 & above	2.10	88.13	100.00	71.70	150.57
	All levers	100.00	20.03			854.54

\* Source : Data for the columns 2—4 from NSS, 15th Round, July 59—June 60, No. 98 New Delhi 1965.

Mid-value for this class is assumed to be equal to the class average.

$$Q_3 = L + \frac{i}{f} \left( \frac{3N}{4} - C \right)$$

Now the upper quartile  $Q_3$  = value of  $\frac{3N}{4}$  th items.

$$\text{i.e.} = \text{value of person } \frac{3}{4} \times 100 = 75\%$$

The percentage lies in the group Rs. 21-24.

Hence  $L = 21$ ,  $i = 3$ ,  $f = 7.83$ , and  $C = 68.26$ .

$$\begin{aligned} \therefore Q_3 &= 21 + \frac{3}{7.73} (75 - 68.26) = 21 + \frac{3}{7.84} (6.74) \\ &= \text{Rs. } 23.58 \end{aligned}$$

Similarly,  $Q_1$ , the lower quartile = value of  $\frac{N}{4}$  th

i.e.  $\frac{100}{4} = 25$ th item. This lies in the group. Rs. 11-13.

So  $L = 11$ ,  $i = 2$ ,  $f = 10.95$ , and  $C = 21.07$ .

$$\begin{aligned} \text{Hence } Q_1 &= 11 + \frac{2}{10.95} (25 - 21.07) \\ &= 11 + \frac{2 \times 3.93}{10.95} = \text{Rs } 11.72. \end{aligned}$$

$$\begin{aligned} \text{Now quartile deviation, QD} &= Q = \frac{Q_3 - Q_1}{2} \\ &= 23.58 - 11.72 \\ &= \frac{11.86}{2} = \text{Rs. } 5.93. \end{aligned}$$

The median per capita monthly expenditure is the value of the  $\frac{N}{2} = \frac{100}{2} = 50$ th percentage person. This lies in the group Rs. 15-18.

Here,  $L = 15$ ,  $i = 3$ ,  $f = 14.43$ ,  $C = 43.14$ .

$$\begin{aligned} \text{Median} &= L + \frac{i}{f} \left( \frac{N}{2} - C \right) \\ &= 15 + \frac{3}{14.43} (50 - 43.14) \\ &= 15 + 1.43 = \text{Rs. } 16.43. \end{aligned}$$

Now mean deviation in relation to median

$$\delta \text{ mad} = \frac{054.54}{100} \text{ Rs. 8.55 app.}$$

Similarly, for this series, the mean is  $= 20.03 = \bar{x}$ .

For computing the mean deviation in relation to the mean we get the deviation etc. as follows:

**Work sheet table 2.1.1 for computing the Mean Deviation**

Class midpoint	$ dx $ from 20.03	Class Average	$f$	$f dx $
x(1)	(2)	x(3)	(4)	(5)
4	16.03	6.62	7.28	116.70
95	10.53	9.72	13.79	145.21
12	8.03	12.07	10.95	87.93
14	6.03	13.94	11.12	67.05
16.5	3.53	16.50	14.43	50.9'4
19.5	0.53	19.72	10.69	5.67
22.5	2.47	22.51	7.83	1862
26	5.97	25.98	6.72	40.12
31	10.97	30.09	8.09	88.75
38.5	18.47	38.37	4.68	86.44
49	28.97	50.47	2,32	67.21
*88.13	63.10	88.13	2.10	143.08

\*Mid-point of the last class assumed  $\Sigma f|dx| = 917.65$  to be equal to the class average.

$$\delta \bar{x} ; \text{ using mid-points of classes} = \frac{\Sigma f|dx|}{N} = \frac{917.65}{100} \text{ Rs. 9.18}$$

we can clearly see that the mean deviation from the Median (Rs. 8.55) is less than that from the mean (Rs. 9.18).

The mean deviation for grouped data is very rarely used. We therefore, take up the discussion of the most important measure dispersion viz. The standard deviation.

### 2.1.5 The Standard Deviation

The standard deviation is regarded as the most important measure of dispersion because this has desirable mathematical properties. We overcome the people of the negative sign are deviations, not by ignoring it, but by squaring the deviation. Standard Deviation is the square-root of the arithmetic average of the squares of the deviation measured from the mean. For ungrouped data we compute the measures as follows :

$$\sigma^2 = \left[ \frac{\Sigma dx^2}{N} \right] = \left[ \frac{\Sigma (x_1 - \bar{x})^2}{N} \right] \quad \text{..... (3.1)}$$

where  $\sigma^2$  stands for the *variance* (which is the square of the standard deviation and  $dx$  stands for  $(x_1 - \bar{x})$ ).

Hence standard Deviation  $\sigma$  (read as ‘sigma’)

$$= \sqrt{\left\{ \left[ \frac{\Sigma(x_1 - \bar{x})^2}{N} \right] \right\}} = \sqrt{\left[ \frac{\Sigma dx^2}{N} \right]} \quad \dots\dots (3.2)$$

**Table 2.1.1 Computation of standard deviation for Annual rainfall in Himachal Pradesh**

Year	Annual Rainfall (cms) $x$	$dx$ from $\bar{x} = 138.5$	$dx^2$	$X^2$	$D$ from $\bar{x}_a = 140$	$D^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1950	146	+ 7.5	56.25	21316	+ 6	36
1955	161	+ 22.5	506.25	25921	- 21	441
1960	120	- 18.5	342.25	14400	- 20	400
1961	161	+ 22.5	506.25	25921	+ 21	441
1965	109	- 29.5	870.25	11881	- 31	961
1966	149	+ 10.5	110.25	22201	+ 9	81
1967	157	+ 81.5	3421.25	24469	+ 117	289
1168	105	+33.5	1122.25	11025	- 35	1225
Total	1108	+ 33.5	3856.00	15731	+74 - 86 = 12	3874

**Source:** Statistical outline of Himachal Pradesh 1970, pp. 92-93 approximation to, cms, done by the author.

$$\bar{x} = \frac{\Sigma x}{N} = \frac{1108}{8} = 138.5 \text{ cms.}$$

$$\sigma^2 = \sqrt{\left( \frac{\Sigma dx^2}{N} \right)} = \sqrt{\left( \left[ \frac{3856.00}{8} \right] \right)} = \sqrt{(482.00)}$$

We pointed out in the last lesson that  $\Sigma d^2$  is a minimum when taken round the arithmetic mean. Therefore, the standard deviation is always computed in relation to the mean. The steps in the above computation are :

- (i) Determine the deviation  $dx$  of each item from the  $X$  i.e. find  $x_1 - \bar{x}$ ,  $\bar{x}$ ,
- (ii) Square these deviations i.e. compute each  $(x_1 - \bar{x})^2$ ,
- (iii) Sum these squared deviations i.e. compute  $\Sigma(x_1 - \bar{x})^2$ ,
- (iv) Divide, this sum by  $N$ . This gives us the value of the *variance* i.e.  $\sigma^2$ .
- (v) Take the square root of the variance.

The above procedure involves computation of deviation ( $x$ ) for every item, and this would be quite lengthy and cumbersome, if the number of items,  $N$  is quite large. The value of standard deviation can also be found out, without doing this procedure, by means of the formula :

$$\sigma^2 = \sqrt{\left\{ \frac{\Sigma x^2}{N} - \left( \frac{\Sigma x}{N} \right)^2 \right\}} \quad \text{..... (3.3)}$$

$$\begin{aligned} &= \sqrt{\left\{ \frac{157314}{8} - (138.5)^2 \right\}} \\ &= (19664.25 - 19182.25) \\ &= (482.00 = 21.95 \text{ cms.}) \end{aligned}$$

It is obvious from the calculation of  $\bar{x}$  in col. 5 of Table 2.2.1 that this method is more suitable to machine calculation and simpler.

By combining formulas (3.2) and (3.3), we get

$$\begin{aligned} \sigma^2 &= \frac{\Sigma(x_i - \bar{x})^2}{N} \\ &= \frac{\Sigma x^2 - \{(\Sigma x)^2 / N\}}{N} \quad \text{..... (3.4)} \end{aligned}$$

or

$$\begin{aligned} \sigma^2 &= \left[ \frac{\Sigma x^2 - \{(\Sigma x)^2 / N\}}{N} \right] \\ &= \sqrt{\left[ \frac{157314 - (1108^2 / 8)}{8} \right]} = \sqrt{\left\{ \frac{157314 - (12227664 / 8)}{8} \right\}} \\ &= \sqrt{\left[ \frac{157314 - 153458}{8} \right]} = \sqrt{\left( \frac{3856}{8} \right)} \end{aligned}$$

We have a short-cut method also available, which allows us to compute the value of the standard deviation, by taking the deviations from an assumed; mean, rather than, from the true mean and make the necessary correction. This formula is:

$$\text{S.D.} = \sigma = \sqrt{\left[ \frac{\Sigma D^2}{N} - \left\{ \frac{\Sigma D}{N} \right\}^2 \right]} \quad \text{..... (3.4.1)}$$

$$\sigma = \sqrt{\left[ \frac{\Sigma D^2 - \{(\Sigma D)^2 / N\}}{N} \right]} \quad \text{..... (3.5.2)}$$

where D stands for deviation of the items from an assumed, mean

$$\text{i.e. } D = x_f - \bar{x}_0 .$$

Let us choose 140 cms as our assumed mean and calculate the standard deviation, as shown in cols. (6) and ,(7) of Table 2.2.1.

Now according to formula 3.4.1., we have

$$\sigma = \sqrt{\left[ \frac{\Sigma D^2}{N} - \left\{ \frac{\Sigma D}{N} \right\}^2 \right]} = (484.25 - 2.25)$$

$$= \sqrt{\frac{3378}{8} - \left\{ \left( \frac{-12}{8} \right) \right\}^2} = 3(484.25 - 2.25)$$

$$= (482.00) = 21.95 \text{ cms.}$$

Similarly, with formula 3.4.2, we have

$$\sigma = \sqrt{\frac{\Sigma D^2 - \{(\Sigma D)^2 / N\}}{N}} = \sqrt{\left[ \frac{3874 - \{(-12)^2 / 8\}}{N} \right]}$$

$$= \sqrt{\left\{ \frac{3874 - 18}{8} \right\}} = \sqrt{\left\{ \frac{3856}{8} \right\}} = \sqrt{(482)}$$

$$= 21.95 \text{ cms.}$$

### The standard deviation for frequency Distribution

For a frequency distribution, where the values of the individual items are not known, such as in Table 2.1, a formula that gives the value for the, standard deviation of the distribution is as follows :

$$\sigma = \sqrt{\left( \frac{\Sigma f dx^2}{N} \right)} = \sqrt{\left( \frac{\Sigma f (x_1 - \bar{x})^2}{N} \right)} \quad \dots\dots (3.5)$$

where  $dx = x_1 - \bar{x}$ , stands for the deviation of the mid values ( $x_1$ 's) for mean. It may be noted that  $x_1$ 's here stand for class mid values as distinguished individual values for ungrouped data.

### Worksheet Table 2.3.1 for computing the Standard Deviation for Consumer's Expenditure Data of Table 2.1

Monthly per capita expenditure classes	Class mid-values ( $x$ )	Percentage of persons in expenditure true class ( $f$ )	Deviations for the mean = 19.74 dx <sup>2</sup> ( $dx$ )		$fdx^2$
(1)	(2)	(3)	(4)	(5)	(6)
0-8	4	7.28	-15.74	248.27	1847.41
8-11	9.5	13.79	-10.24	104.86	1146.02
11-13	12	10.95	-7.75	59.91	656.01
13-15	14	11.12	-5.74	32.95	366.40
15-18	16.5	14.43	-3.24	10.50	151.52
18-21	19.5	10.69	-0.24	0.06	0.64
21-24	22.5	7.83	+2.76	7.62	59.66
24-28	26	6.72	+6.26	39.12	263.36
28-34	31	8.09	+11.26	126.79	1025.73
34-43	38.5	4.68	+18.76	351.94	1647.08
43-55	49	2.32	+29.26	856.15	1986.27
55 & above	88.13	2.10	+68.39	4677.19	9822.10
All levels		100.00			9232.20

\* Mean compound through the use of class in d - point = 19.74.

\* Class mid-value assured to be equal to class average,

$$\sigma = \sqrt{\left( \frac{\sum f dx^2}{N} \right)} = \sqrt{\left( \frac{19232.28}{100} \right)} = \sqrt{192.32.20}$$

$$= \text{Rs. } 13.87$$

Computation of the standard deviation, through this long method is extremely laborious and cumbersome. We may therefore use deviations from an assumed mean and make the necessary correction for this by the use of the formula.

$$\sigma = \sqrt{\left[ \frac{\sum f D^2}{N} - \left\{ \frac{\sum f D}{N} \right\}^2 \right]} \text{ or } \sqrt{\left[ \frac{\sum f D^2 - \{ \sum f D \}^2 / N}{N} \right]} \dots\dots (3.6)$$

where  $D = x_1 = \bar{x}_a$

Let us compute the standard deviation by the short-cut method 3.6 for the foregoing frequency distribution.

**Worksheet Table 2.3.2 for combating the standard Deviation for Consumer's Expenditure**  
**Data of table 2.1.**  
**(by short-cut method)**

Class mid-values (x)	Frequency (f)	Deviations from $\bar{x} = 20$ (D) = $x - \bar{x}_a$	f.D (2) $\times$ (3)	f.D <sup>2</sup> (5) -(3) $\times$ (4)
(1)	(2)	(3)	(4)	(5)
4	7.28	- 16	-116.48	1863.68
9.5	13.79	- 10	-144.80	1520.40
12	10.95	- 8	-87.60	700.80
14	11.12	- 6	-66.72	400.32
16.5	14.45	- 3.5	-50.05	176.75
19.5	10.69	- 0.5	-5.34	2.67
22.5	7.83	+ 2.5	+40.38	248.93
26	6.72	+ 6	+40.32	241.92
31	8.09	+ 11	+88.99	978.89
38.5	4.68	+ 18.5	+86.58	1601.73
49	2.32	+ 29	+67.28	1951.12
88.13	2.10	+ 68.13	+143.07	9747.36
	100.00		- 471.44	19234.59
			+ 445.82	
			- 25.62	

$$\begin{aligned}\sigma &= \sqrt{\left( \frac{\Sigma fD^2 - \frac{(\Sigma fD)^2}{N}}{N} \right)} = \sqrt{\left( \frac{19234.59 - \left( \frac{-25.62}{100} \right)^2}{100} \right)} \\ &= \sqrt{\left( \frac{19234.59 - \left( \frac{656.28}{100} \right)}{100} \right)} = \sqrt{\left( \frac{19234.59 - 6.56}{100} \right)} \\ &= \sqrt{\left( \frac{19228.03}{100} \right)} = \sqrt{192.28} = \text{Rs. } 13.87.\end{aligned}$$

A further implication of computation can be achieved by taking the deviation in tens of the common class interval or common factor. The formula then, becomes

$$\sigma = i \times \sqrt{\left( \frac{\Sigma fD'^2 - \frac{(\Sigma fD')^2}{N}}{N} \right)} \text{ where } D' = \frac{D}{i}.$$

$i$  = common class-interval or common factor.

**Example:**

**Table 2.4.1 for computing Standard Deviation**

Percentage of mark	Mid-value (x)	Number of Student (f)	$D' = (x - \bar{X}_a)/i$ $\bar{X}_a = 05x \quad i = 20$	$fD'$	$fD'^2$
0-20	10	7	-2	-14	28
20-40	30	15	-1	-15	15
40-60	50	0	0	0	
60-80	70	12	1	12	12
80-100	90	9	2	18	36
				-29	
				+30	91
				=1	

$$\sigma = 20 \times \sqrt{\left[ \frac{91 - \frac{(1)^2}{100}}{100} \right]}$$



$$= 20 \times \left[ \left( \frac{90.9999}{100} \right) \right]$$

= 19.08 marks.

Formula 3.5 can also be 'written in the form

$$\sigma = \sqrt{\left\{ \frac{\sum fx^2}{N} - \left( \frac{\sum fx}{N} \right)^2 \right\}} \quad \dots (3.8)$$

### Properties of the standard Deviation.

The standard deviation and its square viz. *Variance* are, by *far*, the most important and most frequently used *absolute* measures of dispersion. Its use in sampling for determining the areas under the normal curve and for various types of skewed distribution is extremely common. It is also used in testing the reliability of certain statistical measures, in correlation and in business cycle analysis.

1. The standard deviation may be defined as :

$$\sigma = \sqrt{\left( \frac{\sum (A - a)^2}{N} \right)}$$

where  $a$  is an average besides the arithmetic mean. Of all such  $\sigma$ 's, the minimum is that for which  $a = \bar{x}$ . As. stated in the last lesson, the sum, of the squares of the deviations from the mean  $\bar{x}$ , is the minimum. Hence, the definition of the standard deviation is

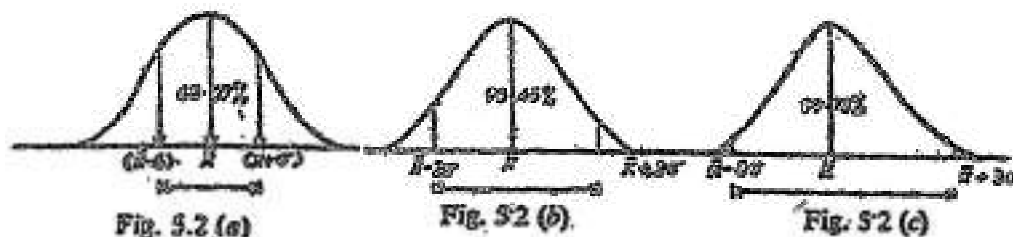
$$\sigma = \sqrt{\left( \frac{\sum (X_i - \bar{X})^2}{N} \right)}$$

2. For normal distributions (single central-peaked *i.e.* unimodal, symmetrical about this central peak and tapering off to both direction from the central peak (to be discussed in a later lesson), it turns out that:

- 68.27% of the observations lie between  $\bar{X} \pm \sigma$  range *i.e.* one standard deviation distance from the mean on either side;
- 95.45 percent of the observations are within the range  $\bar{X} \pm 2\sigma$  and  $\bar{X} + 2\sigma$ , *i.e.* two standard deviation distance of the mean, on either side; and
- 99.73 percent of the observations are within the range  $\bar{X} \pm 3\sigma$  *i.e.* three standard deviation distance from the mean, on either side. These are shown in the diagram 5.2 (a), 5.2 (b) and 5.2 (c) below. For moderately skewed 'distributions, the above percentages may hold approximately.

### Diagram

#### Percentage of Area under the Normal Curve



For the consumption expenditure distribution of Table 2.1  $\bar{X} \pm \sigma = 19.744 \pm 1370$  is Rs. 6.04 and 33.44. To find out the percentage of persons, lying between these two limits, we first determine the percentage included in the range two limits, we first determine the percentage included in the range Rs. 6.04 and Rs. 8 (the upper limit of the first group) and in the range, Rs. 28 and Rs. 33.44. The percentage of persons in the classes, 8-11, 11-13, 13-15, 15-18, 18-21, 21-24, and 24-28, all lie within the limit of  $X \pm \sigma$  or: Assuming regular distribution of frequencies in the classes, 0 to 8, and 28 to 34, the estimated, frequency lying between, the specified limits are:

$$\text{Rs. 6.04 and 8} = \frac{8 - 6.08}{8} \times 7.78 = 1.75\% \text{ app.}$$

$$\text{and Rs. 28 and Rs. 33.4} = \frac{33.4 - 28}{6 = (34 - 28)} \times 8.09 = 7.28\% \text{ app.}$$

Hence, the percentage between Rs. 6.04 and 33.44 = 1.75 + 13.79 + 10.95 + 11.12 + 14.43 + 10.69 + 7.81 + 6.72 + 7.28 + 84.56%.

Similarly, the percentage of persons with per capita monthly consumption expenditure, lying within the range  $\bar{X} + 2\sigma$  i.e... Rs. 19.73 + 2 (13.70) - Rs 47.13 an 0\*, may no calculation be found to be 96.28% and those within  $X \pm 3\sigma$ , 99.82%. It is obvious from these observed percentages that our distribution diverges sharply from a normal curve and is highly skewed.

When we study the normal curve, we shall be interested in determining the limits within which 90%, or 95% or 99% of the observation lie, or rather in the proportion lying beyond certain specified limits, such as 1 percent or 5 percent etc. We can 'always translate the difference between the mean and an individual value into units of standard deviations. We say the deviation  $X_1 - \bar{X}$  have been *standardized of normalized*.

In general, when the variable X is divided by its standard deviations,

$$\frac{X}{\sigma} : \frac{X_1}{\sigma} : \frac{X_2}{\sigma} \dots \dots \frac{X_n}{\sigma}$$

We say the variable X has been *standardized*. These deviations in units of standard deviations

i.e.  $\frac{X_1 - \bar{X}}{\sigma}$  are called Z's (more about these later).

The standard deviation of the distribution of a standardized variable:

$$\frac{X}{\sigma} : \frac{X_1}{\sigma} : \frac{X_2}{\sigma} \dots \dots \frac{X_n}{\sigma}$$

is unity, in theoretical statistics, distribution with mean = 0 and standard, deviation (or variance.) = 1 (called unit distribution are often used to facilitate analysis).

3. If we have two set, consisting of  $N_1$  and  $N_2$  numbers (or two frequency distribution with total frequencies'  $N_1$  and  $N_2$ ) have variances given by  $\sigma_1^2$  and  $\sigma_2^2$  respectively and the same mean =  $\bar{X}$  then the combined variance ( $\sigma^2$ ) of both sets (or both frequency distribution) is given by

---

\*Consumption expenditure cannot fall below zero  $X + \sigma$  i.e. Rs. 19.73 + 2 (13.73).

$$(i) \quad \sigma^2 = \frac{N_1\sigma_1 + N_2\sigma_2}{N = (N_1 + N_2)} \quad \dots (3.9.1)$$

$$\text{or} \quad N\sigma^2 = \sum_i^{N_1} (X_{1i} - \bar{X}_1)^2 + \sum_i^{N_2} (X_{2i} - \bar{X}_2)^2 \quad \dots (3.9.2)$$

(ii) Where, however,  $N_1$ ,  $\bar{X}_1$  and  $\sigma_1$  and  $N_2$ ,  $\bar{X}_2$  and  $\sigma_2$  are given of two Sets of variates, the variance for the composite set is given by :

$$N\sigma^2 = N_1 (\sigma_1^2 + d_1^2) + N_2 (\sigma_2^2 + d_2^2) \quad \dots (3.9.3)$$

where  $d_1 = \bar{X}_1 - \bar{X}$  and  $d_2 = \bar{X}_2 - \bar{X}$ , and where  $\bar{X}$  is ' combined mean for the two groups :

The combined variance  $\sigma^2$  can also be computed by the following methods.

$$(iii) \quad N\sigma^2 = N_1 (\sigma_1^2 + \bar{X}_1^2) + N_2 (\sigma_2^2 + \bar{X}_2^2) - N\bar{X}^2 \quad \dots (3.9.4)$$

$$(iv) \quad R\sigma^2 = N_1\sigma_1^2 + N_2\sigma_2^2 + \frac{N_1N_2}{(N_1 + N_2)} (\bar{X}_1 - \bar{X}_2)^2 \quad \dots (3.9.5)$$

These results can be generalized to 3 or more sets.

### Sheppard's correction for variance

Sheppard's correction is intended to make adjustments for error due to grouping, of data into classes (grouping error)

$$\text{Corrected } \sigma^2 = \text{Variance from grouped data } C^2/12. \quad \dots (3.10)$$

where  $C$  is the class interval size. The correction  $C^2/12$  is used, for distributions of continuous variables where the "tails" taper gradually to zero in both directions.

There is lack of agreement among statisticians when and where Sheppard's corrections should be applied, because of the belief that they often to *over correct* and thus replace old errors, by new errors. Hence, great caution is needed in their use.

### Empirical Relation between Measures of Dispersion

For moderately skewed distribution, the relationship between mean deviation, quartile deviation and standard, deviation is as under.

$$\delta\bar{x} = \frac{4}{5} \sigma$$

$$\text{and} \quad Q.D = \frac{4}{5} \sigma$$

These relationships result from the fact that for the normal *distribution*, the mean deviation and the quartile deviation are equal to  $0.797 \sigma$  and  $0.6745\sigma$ , respectively.

### 2.1.2 Measures of Relative Dispersion

We have so far been discussing measures of absolute dispersion, all of which are expressed in terms of the units rupees centimeters etc. When we want to compare the dispersion of spread of two or more series, it may not generally be appropriate to use such a measure. We may be faced with three possible types of situations while making such comparisons:

- (a) The series sought to be compared may be expressed in terms of the same units, and the means may be equal or nearly, equal. Hence absolute dispersion measures can serve fairly adequately for comparing the series.
- (b) The series to be compared may be expressed in terms of the same units, but the means may be different. Assume that two brands of electric bulbs have the following mean life and standard deviation :

Company A		Company B (in hours)	Company C
Mean life	1500	1800	2500
Standard deviation	50	90	100

Here the lowest absolute dispersion measure is 50 *i.e.* of Company A. However, we cannot say, for certain that the life of company. A's electric bulbs is relatively 'less spread or scattered. Here measures of *relative dis-persion* are a better guide.

### Co-efficient of Variation

$$V = \frac{\sigma}{x} \times 100 \quad \dots (3.11.1)$$

(expressed as a percentage)

The co-efficient of variation, expressed of a fraction is also sometimes, called the Co-efficient of Standard deviation and is equal to  $\frac{\sigma}{x}$ .

Thus, for the relative dispersion of the three/ brands of electric bulbs, we have

$$V_A = \frac{\sigma_1}{x_1} \times 100 = \frac{50}{1500} \times 100 = 3.33\%.$$

$$= \frac{\sigma_2}{x_2} \times 100 = \frac{90}{1800} \times 100 = 5\%.$$

$$V_c = \frac{\sigma_3}{x_3} \times 100 = \frac{50}{2500} \times 100 = 4\%.$$

It is evident that the series A is relatively less spread and hence, more concentrated around its mean.

- (c) The series to be compared may be expressed in different units, and we cannot therefore, compare the standard deviations.

Similar to the co-efficient of standard deviation are the co-efficient of quartile deviation and mean deviation.

$$\text{Co-eff. of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad \dots (3.12)$$

$$\text{and Co-eff. of Mean Deviation or} = \frac{\delta \bar{x}}{\bar{X}} \text{ or } \frac{\delta \text{Med}}{\text{Med}} \quad \dots (3.13)$$

Thus for any absolute dispersion measure, we can find out a relative dispersion measure,

$$\text{Relative Dispersion} = \frac{\text{Absolute Dispersion}}{\text{Average used}}$$

$$\text{Thus Co-eff. of Range} = \frac{X_h - X_l}{X_h + X_l} \quad \dots (3.14)$$

where  $X_h$  is the highest value of an item

and  $X_l$  is the lowest value of the items in the series.

$$\text{Co-efficient of 10-90 percentile Range} = \frac{P_{90} - P_{10}}{P_{90} + P_{10}} \quad \dots (3.15)$$

## 2.9 Skewness

Skewness is the degree of asymmetry or departure from symmetry of a distribution. Measures of skewness not only indicate the magnitude of skewness but also its *distribution*. A distribution is said to be skewed in the direction to the extreme values or in the direction of excess tail, for a frequency curve. Most of the skewed curves in social sciences are skewed to the right. It is only very rarely that we meet curves skewed to the left and even more rarely, do we find series characteristically skewed to the left *i.e. negatively skewed*.

A large number of series and frequency distributions are, however, characteristically skewed to the right (or *positively skewed*). Frequency distributions of wages, salaries, incomes, earnings, sales, output, consumption, and of a number of other socio, economic and business variables are likely to be sharply skewed to the right.

### Diagram

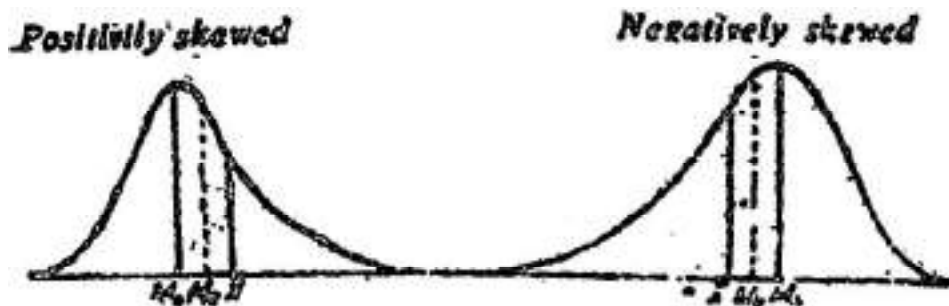


Fig. 5.3

Croxten, Cowden and Klein cite the distribution of ages at death of 371 American inventors, which happened to be characteristically skewed to the left. The plausible reasons cited are :

- (i) "Younger men do not often have enough inventions to their credit to be classified as "inventors"
- (ii) "a time factor is present—almost one fifth of the inventors included in this study were born before 1800."

We have seen in the preceding lesson that the per-centage of the extreme, values does not affect the mode, the median is affected by their position only and the arithmetic mean is influenced by their size. Karl Pearson used these characteristics of the mode and the mean to measure skewness. An absolute measure of skewness, then, would be

$$\text{Skewness} = \text{Mean} - \text{Mode} \quad \dots (3.16)$$

Since, in a moderately skewed distribution of a continuous variable, Pearson showed that the median travels  $\frac{2}{3}$  the distance from the mode towards the mean the value of mode could be written as

$$M_0 + \bar{X} - 3(\bar{X} - \text{Med.}) = 3 \text{ Median} - 2 \text{ Mean.}$$

Substituting the expression so the mode in the measure of skewness (3.16) we get

$$\begin{aligned} \text{Absolute Skewness} &= \bar{X} - [\bar{X} - 3(\bar{X} - \text{Med.})] \\ &= (\bar{X} - \text{Med}) \\ &= (\text{Mean} - \text{Median}) \quad \dots (317) \end{aligned}$$

However, measures of absolute skewness suffer, from certain obvious defects. First, these measures are expressed in terms of the units of the problems. Thus, heights would be measured in centimeters and weights, in kilograms. We cannot, obviously, compare the skewness in the distribution of Heights with that in the distribution of weights without bringing them to a common denominator.

This can be achieved by dividing the absolute measure by a measure of dispersion, such as the standard deviation. Hence,

$$S_k = \frac{\bar{X} - M_0}{\sigma} \quad \dots (3.18)$$

Since, in income distributions, mode is only an approximation, M can be more satisfactorily located and hence used for the purpose.

$$S_k = \frac{3(\bar{X} - \text{Med})}{\sigma} \quad \dots (3.19)$$

The above two measures are called, Pearson's *first and second Co-efficient of skewness*. Similarly, we know that in a symmetrically distributed, median lies exactly half-way between the two quartiles, as also between the 10th and 90th percentiles. Hence, when the median does not lie in this position, some asymmetry or skewness is surely present. Based on this position of the median vis-a-vis the quartiles and the 10th & 90th percentiles, we have quartile and percentile coefficients of skewness. These are

$$\begin{aligned} S_k &= \frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{Q_3 - Q_1} \\ &= \frac{Q_3 - Q_1 - 2 \text{ Med.}}{Q_3 - Q_1} \quad \dots (3.20) \end{aligned}$$

$$\text{and } S_k = \frac{P_{90} - P_{10} - 2 \text{ Med.}}{P_{90} - P_{10}} \quad \dots (3.21)$$

However, these measure of skewness are not as satisfactory as Pearson's 'co-efficient, because they suffer from obvious deficiencies of the quartiles and percentiles.

An important measure of skewness used is the *third moment* about the mean, expressed in dimension less form as follows :

$$\text{Moment Co-efficient of Skewness} = a_3 = \frac{m_3}{\sigma_3}.$$

$$\frac{m_3}{(\sqrt{m_2})^3} \quad \text{..... (3.22)}$$

where  $m_2$  and  $m_3$  are the second and third moments about the mean respectively.

Another measure of skewness is sometimes given by  $\beta_1 = \alpha_3^2$ . For perfectly symmetrical curves, such as the normal curve,  $\alpha_3$  and  $\beta$  are zero.

### Ginni's Mean Difference:

Carrado Ginni has suggested alternative method of studying dispersion. The method is :

$$\text{Ginni mean Difference} = \frac{D}{nq}$$

$$S_k = \frac{3(\bar{x} - Med)}{a}$$

Nq = Number of difference

$$= 1/2 n(n-1),$$

Following Example illustrate the method.

**Example:** Find Ginnis mean differences from the following items:

X : 8 10 12 14 16

**Solution :** Given items are 8, 10, 12, 14, 16

$$16-8=8 \quad 14-8=6 \quad 12-8=4 \quad 10-8=2$$

$$16-10=6 \quad 14-10=4 \quad 12-10=2$$

$$16-12=4 \quad 14-12=2$$

$$16-14=2$$

20	12	6	2
----	----	---	---

Now total of alt differences  $20 + 12 + 6 + 2 = 40$

$$\text{Total Number of differences} = \frac{1}{2} n (n-1) = \frac{1}{2} 5 (5-1) = 10$$

$$\text{Gini's Mean difference} = \frac{D}{nq} = \frac{40}{10} = 4 \text{ Ans.}$$

### Moments:

If  $X_1, X_2, \dots, X_n$ , are the values assumed by the variable  $X$ , then the  $r$ th moment is defined as.

$$X_r = \frac{X_1 + X_2^r + \dots + X_n^r}{N} = \frac{\sum X^r}{N} \quad \dots\dots (3.23)$$

The first moment (i.e. when  $r=1$ ) is the arithmetic mean  $\bar{X}$ .

The  $r$ th moment about the mean  $\bar{X}$  is defined as

$$\pi_r \text{ or } m_r = \frac{\sum (X - \bar{X})^r}{N} = \frac{\sum (dx)^r}{N} \quad \dots\dots (3.24)$$

where  $D_x = X - \bar{X}$  is the deviation of  $X$  from  $\bar{X}$  (the assumed mean), if  $\bar{X} = 0$ , (2.25) is often called,  $r$ th moment about zero.

The following relations exist between the moments, about the mean,  $m_r$  or  $p_r$  and moments about an arbitrary origin,  $v_r$  and  $p_r$

$$m_1 = \pi_1 = 0 \quad \dots\dots (3.26.1)$$

$$m_2 = \pi_2 = v_2 = v_1^2 \quad \dots\dots (3.26.2)$$

$$m_3 = \pi_3 = v_3 = 3v_1 v_2 + 2v_1^3 \quad \dots\dots (3.26.2)$$

where  $v_1, v_2$  and  $v_3$  are the 1st, 2nd, 3rd moments about an arbitrary origin  $m_2$  or  $\pi_3$  is a measure of absolute skewness.

The measure of relative skewness is

$$\beta = (a_2) = \frac{\pi_3^2}{\pi_2^3} \quad \dots\dots (3.27.1)$$

$$a_1 = \sqrt{\beta} = \sqrt{\frac{\pi_3}{\pi_2^3}} \quad \dots\dots (3.27.2)$$

For symmetrical series  $\beta = 0$ . The greater the value of  $\beta$ , the more skewness there is in the series.

### 5.4 Kurtosis

Kurtosis is the degree of peakedness of distribution, usually taken in relation to a normal distribution. A distribution having a relatively high peakedness as the curve of Fig. 5.4 (a) is called *leptokurtic* while the curve of Fig. 5.4 (b) (c) which is flat topped is called *platykurtic*. The normal distribution, Fig. 5.4 (b) which is neither, very peaked nor very flat topped, is called *mesokurtic*.

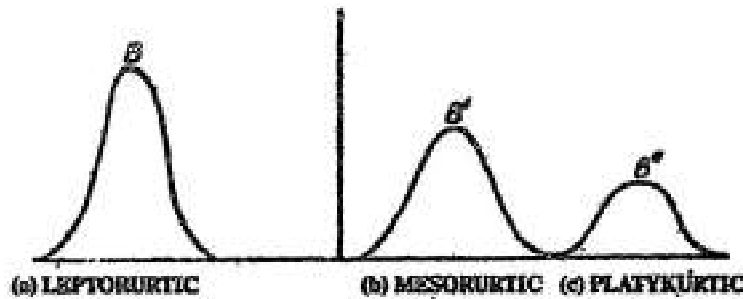


Fig 5.4 (a)

Fig. 5.4 (b)

\*Kurtic Means Jump backed : thus humped or un-normal. Lep to means slender narrow, Platy means broad, wide flat and Meso means in the middle.



The degree of kurtosis present in a series, may be measured by making use of the fourth moment about the mean.

$$\pi_4 \text{ or } m_4 = \frac{\sum dx^4}{N} \quad \dots (3.28)$$

Where  $d_x$  stands for the deviation of the items from the mean or for a frequency distribution:

$$\pi_4 \text{ or } m_4 = \frac{\sum fdx^4}{N}$$

The fourth, moment about the mean, can be expressed, in terms of the moments about any arbitrary origin, as under (in terms of class interval units).

$$\begin{aligned} \pi_4 = & \frac{\sum f(d')^4}{N} - 4 \frac{\sum f(d')}{N} \times \frac{\sum f(d')^3}{N} \\ & + 6 \left( \frac{\sum fd'}{N} \right)^2 \left( \frac{\sum f(d)^2}{N} \right)^2 - 3 \left( \frac{\sum f(d')}{N} \right)^4 \end{aligned}$$

However  $\pi_4$  is an absolute measure of kurtosis. The *moment Co-efficient of kurtosis* is given by

$$\beta_4 \text{ or } \alpha_4 = \frac{\pi_4}{\sigma^4} = \frac{\pi_4}{\pi_2^2} \quad \dots (3.30)$$

For the normal distribution  $\beta_2 = 3$

For the reason the kurtosis is sometimes defined which is positive for a leptokurtic distribution, negative for a platykurtic distribution and zero for the mesokurtic or normal distribution. Alternately, we can put this criterion as

Type of Curve	Value of
Leptokurtic	$\beta_2 > 3$
Mesokurtic	$\beta_2 = 3$
Platykurtic	$\beta_2 < 3$

$k_0$  another measure, of kurtosis, which is some-times used, is based on both quartiles. and precentiles and is given by  $\pi_4$ .

$$k \text{ (pronounced as Kappa)} = \frac{Q}{P_{90} - P_{10}} \quad \dots (3.31)$$

Where Q stands for the quartile deviation. This is sometimes referred to as percentile co-efficient of kurto-sis. For the normal distribution, it has the value 0.263.

**Example 1.** Let us compute the moments, moment coefficient of skewness and kurtosis for our data of percentage marks of students in Economics in Table 5.4.

Percentage Marks of Students Mid- value X	Number of Students $f$	$D'$	$fD'^1$	$fD'^1$	$fD'^1$	$fD'^4$
10	7	-2	-14	28	-56	115
30	15	-1	-15	15	-15	15
50	32	0	0	0	0	0
70	12	+ 1	+12	12	+ 12	12
90	9	+ 2	+18	+ 36	+72	144
			-29	91	-71	
	75		$\frac{+30}{+1}$		$\frac{+84}{+15}$	283

$$m_1' \text{ or } v_1 = \frac{\Sigma fD}{N} \times i = \frac{1}{75} i = 0.13 i = 0.13i.$$

$$m_2' \text{ or } v_2 = \frac{\Sigma fD'}{N} \times i^2 = \frac{91}{75} i^2 = 1.213i^2.$$

$$m_3' \text{ or } v_3 = \frac{\Sigma fD'^3}{N} \times i^3 = \frac{13}{75} i^3 = 0.173i^3.$$

$$m_4' \text{ or } v_4 = \frac{\Sigma fD'^4}{N} \times i^4 = \frac{283}{75} i^4 = 3.77i^4.$$

Where  $i$  stands for the size of the class interval, 20, in the present case,  $m_1$  or  $\pi_1$ , = 0.

$$m_2 \text{ or } \pi_2 = v_2 - v_1^2 = [1.213 - (0.13)^2] i^2 = 1.213 i$$

$$m_3 \text{ or } \pi_3 = v_3 - 3v_1v_2 + 2v_1^3 = (0.173 - 3(0.13)(1.213) + 2(.013)^3] i^3$$

$$= 0.173 - .047 + .000] i^3$$

$$m_4 \text{ or } \pi_4 = v_4 - 4v_1v_3 + 6v_1^2v_2 - 3v_1^4 = 3.738 i^4$$

$$\sqrt{\beta_2} = \alpha_3 = \frac{\pi_3}{\sqrt{\pi_2^3}} = \frac{0.126 i^3}{\sqrt{\{(1.213i^2)^3\}}} = .07$$

$$\beta_2 = \alpha_4 = \frac{\pi^4}{\sqrt{\pi_2^2}} = \frac{3.738}{1.471(i^2)^2} = 2.565$$

The values of  $\alpha_3$  and  $\alpha_4$  show mat the distribution is very slightly skewed positively and is tending to be platykurtic.

1. Before the representatives of a measure of, central tendency can be assessed, the dispersion or spread of the data needs to be known.

2. Of the Various measures of absolute dispersion, the standard deviation is, by far, the most important, and most frequently used.

(i). The range =  $X_n - X_1$  i.e. the difference between, the highest and lowest values of the Variable  $x_i$  is a very crude measure and uses only two extreme values-the lowest and the highest.

(ii). The 10—90 percentile range =  $P_{90} - P_{10}$ , though an improvement upon the rang, suffers broadly from the same defects as the range.”

(iii) The quartile deviation,  $Q = \frac{Q_3 - Q_1}{2}$ -encompasses the middle 5% of items. Like the 10—90 percentile range, it is not affected by ‘extreme values’. ‘However,’ like the range the 10 - 90 percentile range, it fails to consider the values of all, items.

(iv) The average or mean deviation.  $\delta = \frac{\sum |d_v|}{N}$  or  $\frac{\sum f |d_x|}{N}$  is usually computed in relation to the mean, even though,  $\sum |d_x|$  or  $\sum f |d_x|$  i.e the sum of deviations, neglecting the sign is the minimum in relation to the median, Despite the feet that it uses all the items in & series, it is very rarely used from frequently distributions.

(v) The standard deviation is the square root of fee mean of the squared deviations of the values from the arithmetic mean.

A number of formula are ‘available for computing the standard deviation,  $\sigma$  as follows :

(a) The formula by long method is,

$$\sqrt{\left\{ \frac{\sum (x - \bar{x}^2)}{N (= f)} \right\}} = \sqrt{\left( \frac{\sum dx^2}{N} \right)}$$

(b) The formula for the short-cut method (especial suited to machine calculation) is

$$\sqrt{\left\{ \left( \frac{\sum x^2}{N} \right) - \left( \frac{\sum X}{N} \right)^2 \right\}} \text{ or } \sqrt{\left\{ \frac{\sum x^2 - \left( \frac{\sum X}{N} \right)^2}{N} \right\}}$$

(c) The formula for the short method using deviations (D) from an arbitrary or assumed mean  $\bar{X}_a$  is

$$\sqrt{\left\{ \frac{\sum D^2}{N} - \left( \frac{\sum D}{N} \right)^2 \right\}} \text{ or } \sqrt{\left\{ \frac{\sum D^2 - \left( \frac{\sum D}{N} \right)^2}{N} \right\}}$$

(d) The variant of (c) above, especially when the distance between successive values of the variable are equal and greater than of less than 1, uses equal distance as common interval (i) and the deviations from the assumed mean, in terms, of this common interval(i);

$$\left( D' = \frac{D}{i} \right)$$

$$i \times \sqrt{\left\{ \frac{\Sigma D'^2}{N} - \left( \frac{\Sigma D'}{N} \right)^2 \right\}}$$

or

$$i \times \sqrt{\left\{ \frac{\Sigma D'^2 - \left( \frac{\Sigma D'}{N} \right)^2}{N} \right\}}$$

For grouped data these deviations are the mid values of the classes from the true or assumed mean, and frequencies are introduced and the 'formula are as follows

$$(a) \quad \sigma = \sqrt{\left\{ \frac{\Sigma f(X - \bar{X})^2}{N (= \Sigma f)} \right\}} = \sqrt{\left( \frac{\Sigma f dx^2}{N} \right)}$$

$$(b) \quad \sigma = \sqrt{\left\{ \frac{\Sigma fx^2}{N} - \left( \frac{\Sigma fx}{N} \right)^2 \right\}} \quad \text{or} \quad \sqrt{\left\{ \frac{\Sigma fx^2 - \left( \frac{\Sigma fx}{N} \right)^2}{N} \right\}}$$

$$(c) \quad \sigma = \sqrt{\left\{ \frac{\Sigma fD^2}{N} - \left( \frac{\Sigma fD}{N} \right)^2 \right\}} \quad \text{or} \quad \sqrt{\left\{ \frac{\Sigma fD^2 - \left( \frac{\Sigma fD}{N} \right)^2}{N} \right\}}$$

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

$$\text{or} \quad i \times \sqrt{\left\{ \frac{\Sigma fD'^2 - \left( \frac{\Sigma fD'}{N} \right)^2}{N} \right\}}$$

3. The Varaince =  $\sigma^2$ .

4. Properties of the standard deviation are:

(i) The sum of squared deviations is the minimum when the deviations are computed from the arithmetic mean, rather from any other origin.

(ii) For normal distribution the following percentage of observations lie within specified 'multiples' of standard deviation from the Mean, in both direction;

	Range	Percentage of item range of limits
X	$\bar{X} = \sigma$	68.27
	$\bar{X} = 2\sigma$	95.45
	$\bar{X} = 3\sigma$	99.73

For moderately skewed distributions, these percentages are approximately valid. This property helps us in computing the actual percentage of observations within specified standardized units range, and then compare these with the expected or theoretical frequencies for the normal curve, and thus find out; how far the distribution departs from the normal.

(iii) The deviations of the individual values from the mean i.e.  $X_i - \bar{X}$  have been *standardized* by dividing

$$\sigma, \text{ i.e. } Z_i = \frac{X_i - \bar{X}}{\sigma}.$$

(iv) We have can compute the *combined variance* (or standard deviation), if we know the  $\sigma$ ,  $s$  of the various sets, and the number of items in each set.

5. Sheppard's correction for variance of grouped data is  $\sigma^2 - C^2/12$ , where  $C$  is the size of the class intervals. This correction is however; a subject of controversy regarding its usefulness.

6. For moderately skewed distributions,

$$\sigma \bar{X} = \frac{4}{5} \sigma$$

$$Q = \frac{2}{3} \sigma.$$

7. Measures of absolute dispersion fail to guide us, if two or more series have different means or if these series are expressed in different units. In such cases, we have to find measures of relative dispersion, the most important among these being the coefficient variation;  $V = \frac{\sigma}{\bar{X}}$  expressed as a fraction or percentage.

### Percentage

In other measure of relative dispersion, too, we divide the measure of absolute dispersion by some sort of average.

$$\text{Relative dispersion} = \frac{\text{Absolute dispersion}}{\text{Average used}}$$

8. Skewness is the degree of departure, from symmetry of a distribution: it may be positive or negative. Most distributions in social sciences are skewed to the right. The coefficients of skewness are given by the following formulae:

$$\text{Karl Pearson's first co-efficient of skewness} = \frac{X - M_0}{\sigma}$$

(ii) Karl Pearson's Second co-efficient of skewness =  $\frac{3(\bar{X} - \text{Med.})}{\sigma}$  especially, where mode is not defined, and is only an approximation.

(iii) Quartile co-efficient of skewness =  $\frac{Q_3 - Q_1}{Q_2 + Q_3}$ .

(iv) Percentage co-efficient of skewness =  $\frac{P_{90} - P_{10}}{P_{90} + P_{10}}$ .

9. (i) The *r*th moment of variable X is  $\frac{\sum X^r}{N}$  the first moment being the mean.

(ii) The *r*th moment about the mean  $\bar{X}$ , is defined as

$$m_r = \pi_r = \frac{\sum (X - \bar{X})^r}{N} = \frac{\sum dx^r}{N}$$

(iii) The *r*th moment about any origin, is  $\bar{X}$

$$v_r \text{ or } \pi_1^r = \frac{\sum (X - \bar{X}_a)^r}{N} = \frac{\sum D^r x}{N}$$

moments about the mean can be expressed in terms of moments about an arbitrary origin.

10. Kurtosis is the degree of peakedness of a distribution.

Distributions may be leptokurtic (with narrow peaks); mesokurtic (with middle or intermediate peak) and platykurtic with flat peak).

Kurtosis is measured by the following co-efficient

(i) Moment co-efficient of kurtosis,  $\beta_2$  or  $\sigma_4 = \frac{\pi_4}{\sigma^4} = \frac{\pi_4}{\sigma_2^2}$  If  $\beta_2 = 3$  the distribution is normal, if  $\beta_2 < 3$ , it is leptokurtic and if  $\beta_2 > 3$ , then it is platykurtic.

(ii) Percentile measure of kurtosis.  $k = \frac{Q}{P_{90} + P_{10}}$

For normal distribution,  $k = 263$ .

### SUGGESTED READINGS

- Clark, T.C. and E. W Jordan : *Introduction to Business and Economic- Statistics*, Southern Western. Publishing Co., 1985.
- W.C. Merrill and K.A. Fox : *Introduction to Economic Statistics*, Wiley, 1970.

\*\*\*\*\*

## LESSON-3

### CORRELATION AND REGRESSION

Dear Friends,

We have so far been studying characteristics of series involving a single variable, consumption expenditure, heights, weights etc. In this lesson, we are going to study a technique, very frequently used in economics and business, research, and by applied statisticians, namely, that of correlation and regression analysis. In the early days Correlation analysis found widespread use in biological problems, but subsequently, it has been extensively used in economics, agriculture, and many other fields.

The term correlation or co-variation indicates the 'relationship between two such variable in which, with changes in the values' of one variable, the values of the other variables also change.

For example, we may be interested in finding out the relationship between the number of years of education completed and the income of adult males in a community or we may be interested in relating the crime rate, with the incidence of employment. In some of these problems, we are often not only to describe the *nature* of relationship between two variables so that we can predict (or estimate) the value of one variable, if we know the value of the other. For instance, we may want to predict a person's future income from his education. 'When we are, principally, interested in them exploratory task of finding out which variables are related to a given variable, we are likely to be mainly interested in measures of degree of relationship such as, correlation co-efficient. On the other hand, once we have found the significant variables', we are more likely to turn our attention to regression analysis. We may take up, correlation first and shall take up regression, in the next lesson.

Correlation can either be positive or negative. When the values of two variables move in the same direction so that an increase in the value of one variable is associated with an increase in the value of the other variable also, and a decrease in the value of one variable is associated with the decrease in the value of the other variable also, correlation is said to be positive or direct if the values of two variables move in different directions in such a way that with an increase in the value of one variable the value of the other variable decreases, and with a decrease in the value of the variable the value of the other variable increases, correlation is said to be negative.

In exact sciences, the study of correlation is easier because mathematical relationship can be established between values of two variables on the basis of experiments. The effect for example, of heat on density of air can be reduced to a mathematical formula disclosing the relationship between these two variables. Economic and social data are affected by a large number of causes. We cannot determine how much a particular cause contributes to a given effect. Increase in price will lead to increase in supply, and vice versa but supply is affected by many other factors also. We cannot make the other factors inoperative for the time being as we can do in experimental methods. Therefore, measurement of correlation is relatively more difficult in social sciences.

Even, when correlation has been established and measured, it does not mean that each and every item would confirm to the same pattern. Tall fathers will have tall sons on the average, but some tall fathers may have short sons.

### 3.2. Degree of Correlation :

When changes in the two Mated variables are exactly Proportional, there is perfect correlation. Correlation-will-men be said to be linear. If a 10% increase in price is every time accompanied by a 15% rise, in supply, the relationship between these two variables is linear and of the type  $y = a + bx$ . The Values of such series when plotted on a graph paper will fall exactly on a straight line. When such a straight line rises from left to right upward, correlation will be perfect and positive. When the straight line falls from left/to right downwards, correlation is perfect but negative.

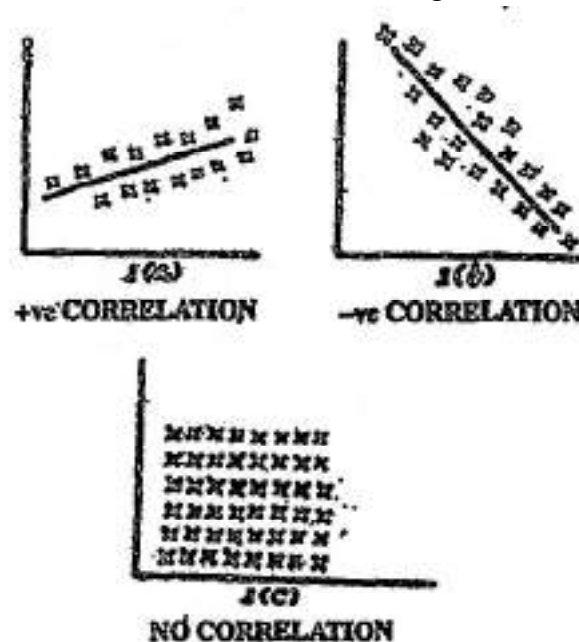
But such perfect linear-correlation are fare in social sciences. If changes in the two variables, are in the same direction but not in the same proportion, the correlation is positive but less than perfect if changes are in the opposite direction (and not in the same proportion), the correlation is negative and limited. Correlation between two series, may be studied by one of the following methods.

1. Scatter diagram.
2. Correlation graph.
3. Coefficient of Correlation.
4. Correlation table.

The third method of computing a coefficient is the most important method.

#### 3.3.1 Scatter Diagram:

Diagrams and graphs can be drawn to have an idea about the correlation, between two variables. Suppose we are given figures of ages of hundred wives and husbands. If the age of husbands is represented by  $x$  and the age of wives by  $y$ , we shall have hundred pairs, of  $x$  and.  $y$  values. Now plot variables, (age of husbands) on the horizontal axis and  $y$ , (age of wives) on the vertical, axis. We shall obtain, one point for each set of two values. When, we have plotted data about, these hundred wives and husbands we shall get hundred such points. On the horizontal scale each point represents the age of the husband and at the vertical scale, the age of wife. The diagram showing, these, hundred points would be called a Scatter Diagram.





If these points show definite tendency or trend upwards or downwards; the two variables are correlated. If the points lie around a diagonal band rising from left bottom to the right top the correlation is positive. If the points are scattered around a diagonal band from left to top to right bottom, the correlation is negative. Closer the points are to the diagonal higher is the degree of correlation, if the points, of scatter diagram be exactly on straight line, the correlation is perfect.

If the plotted points scattered widely over the whole graph so that no evidence of definite, trend or tendency in any direction is available, the correlation does not exist.

A scattered diagram can only provide rough idea, about the degree of correlation and its direction i.e. whether the correlation is positive or negative. It does not provide any exact, measure of correlation between two variables.

### **3.3.2. Correlation by graphic Method:**

Another way of detecting positive or negative correlation and also its extent is to draw correlation graphs and read the direction of the curves.

You have to be very careful, while, drawing a correlation graph about the choice of the series and the base line. They should be so chosen that the averages of the two variables on the vertical scale should be as close to one another as possible.

If the curves or graph representing two variables move in the same direction in the same ranges, the correlation is positive; if they move in opposite directions, the correlation is negative. And if the two curves show no particular pattern, in their movements, sometimes moving in the same direction and other times in opposite direction, sometimes one rising or falling and the other remaining constant correlation does not exist at all.

### **3.3.3.0 Coefficient of Correlation :**

Correlation exists in various degrees. It is perfect and positive when an increase (or decrease) in one variable is always followed by a corresponding and proportional increase (or decrease) in the other variable. Correlation is perfect but negative if an increase or decrease in one variable is always, followed by a corresponding and proportional decrease (or increase) in the other variable. Correlation; will be non-existent or zero if changes in one variable cannot be associated at all with changes in other variable. In between positive perfect correlation and no correlation; there shall be different degrees of positive correlation, and similarly in between perfect negative correlation and no correlation, there shall be different degrees of limited negative correlation.

Coefficient of correlation is calculated to study the extent or degree of correlation between two variables. Generally Karl Pearson's Coefficient 'r' of Correlation is used. This Coefficient varies between + 1 and —1. Perfect positive correlation is indicated by + 1. Perfect negative correlation by —1 and complete independence by 0. Limited degrees of positive correlation are indicated by values lying between 0 and +1-and limited; degrees of positive correlation by values lying between 0. and —1. All this, is true also of Spearman's Coefficient of Rank Correlation.

### **3.3.3.1. Karl Pearson's Coefficient of Correlation:**

The formula devised by Karl Pearson, the great biologist and statistician, measuring coefficient of correlation between two variables is the most satisfactory. The formula is as under:

$$r = \frac{\sum dx dy}{n \sigma_1 \sigma_2} \dots\dots\dots 3.1$$

Here r stands for the coefficient of correlation,  $dx$  for the deviations ,of values of, x variables from the arithmetic mean of x series,  $dy$  for the deviations of y values from the arithmetic mean of y series, n for the number of pairs of observations,  $\sigma_1$ , for the standard deviation of the x-series and  $\sigma_2$  for the standard deviation of the y-series.

In some books, you will find this formula written as:

$$r = \frac{\sum xy}{n \sigma_1 \sigma_2}$$

Here x and y are the same things as  $dx$  and  $dy$  in formula 3.1 above

Thus the coefficient of correlation between variables is obtained by dividing fee sum of the products of the corresponding deviations of the various items of two series from their respective mean by the product of their standard deviations and the number of pairs of observations.

$$r = \frac{\sum dx dy}{n \sigma_1 \sigma_2}$$

is the basic form of Pearson's formula. All me number forms are derived from this fundamentals form.

**Example 1.** The following table shows the marks obtained by ten students in Economics and Statistics. Find the coefficient of correlation;

Marks in Eco.	78	36	98	25	75	82	90	61	65	39
Marks in Stats.	84	51	91	60	68	62	86	58	53	47

**Solution:**

	I	II	III	IV	V	VI	VII
Sr. No.	Marks In Eco. $x$	Deviation from $dy=(65)$ $dx$	$d^2x$	Marks in Stats $y$	Donation from $=(66)$ $dy$	$d^2y$	$dxxdy$
1	78	+ 13	169	84	18	324	+ 234
2	36	- 29	841	51	-15	225	+ 435
3	98	+ 33	1089	91	+ 25	624	+ 825
4	25	- 40	1600	60	- 6	36	+ 240
5	75	+ 10	100	68	- 1	4	+ 20
6	82	+ 17	289	62	- 4	16	- 68
7	90	+ 24	625	86	+ 20	400	+ 500
8	72	- 3	9	58	- 8	64	+ 24
9	65	0	- 0	53	- 13	169	+ 0
10	39	- 26	676	47	- 19	361	+ 494
		0	5398	= 660		=0 2224	=2704
	$\sum x = 650$		$\sum d x^2$	$\sum y$		$\sum d y^2$	$\sum dxdy$

Marks in Economics ( $x$ ) and marks in statistics, ( $y$ ) are given in columns I and IV respectively.

$$\text{A.M. of } x \text{ series is } \frac{\Sigma x}{n} = \frac{650}{10} = 65$$

$$\text{A.M. of } y \text{ series is } \frac{\Sigma y}{n} = \frac{660}{10} = 66$$

Standard deviation of  $x$  series -

$$\sigma_1 = \sqrt{\left(\frac{\Sigma dx^2}{n}\right)} = \sqrt{\left(\frac{5398}{10}\right)} = \sqrt{(539.8)}$$

Standard Deviation of  $y$  series -

$$\sigma_2 = \sqrt{\left(\frac{\Sigma dy^2}{n}\right)} = \sqrt{\left(\frac{2224}{10}\right)} = \sqrt{(222.4)}$$

$$\begin{aligned} r &= \frac{\Sigma dx dy}{n \sigma_1 \sigma_2} = \frac{2704}{10(539.8) \times (222.4)} \\ &= \frac{2704}{10 \times 23.2 \times 14.7} = \frac{2704}{3456.8} = 0.78 \text{ approximately.} \end{aligned}$$

The algebraic sign of  $r$  will be the same as that of  $\Sigma dx dy$ , the number of pairs of observation cannot, be negative,  $\sigma_1$  and  $\sigma_2$  cannot be negative. The denominator of  $\frac{\Sigma dx dy}{n \sigma_1 \sigma_2}$  cannot be negative. If therefore,  $\Sigma dx dy$  is positive,  $r$  will also be positive if  $\Sigma dx dy$  is negative,  $r$  will be negative.

If can be proved that  $r$  or  $\frac{\Sigma dx dy}{n \sigma_1 \sigma_2}$  cannot exceed the numerical value of 1. We need not take the trouble of proving this. But we must always be aware that 1 (— 1) correlation is negative) is the maximum or minimum value that  $r$  can have. If in a problem your  $r$  comes to be greater than 1, it simply means that your calculations have been wrong somewhere.

The formula (3.1) above is

$$r = \frac{\Sigma dx dy}{n \sigma_1 \sigma_2}$$

If values of  $\sigma_1$  and  $\sigma_2$  inserted, the formula becomes

$$r = \frac{\Sigma dx dy}{\left(\frac{\Sigma dx^2}{n}\right) \times \left(\frac{\Sigma dy^2}{n}\right)} \quad \dots (3.2)$$

In the denominator, the two  $n$ 's within the under root sign's will cancel with the one a outside. Formula (3.2) will then be reduced to

$$r = \frac{\Sigma dx dy}{\sqrt{(\Sigma dx^2)} \times \sqrt{(\Sigma dy^2)}} \quad \text{..... (3.3)}$$

From formula (3.3) it should be clear that we need not *find* standard deviations in order to calculate *r*.

Let us find *r* in Example 1 by applying formula. (3.3).

$$r = \frac{\Sigma dx.dy}{\sqrt{(\Sigma dx^2)} \times \sqrt{(\Sigma dy^2)}}$$

$$r = \frac{2704}{\sqrt{(5598)} \times \sqrt{(2224)}} = \frac{2704}{3456.8}$$

$$= 0.78 \text{ approximately}$$

exactly the same as obtained earlier.

### 7.3.3.2. Short cat Method :

In example (1) above arithmetic means of both *x* and *y* variables are complete numbers and, therefore, finding deviations of value from these arithmetic means and processing, these deviations, mathematically is not difficult. If however, arithmetic means are in fractions finding deviations and processing them becomes a tedious job.

In such cases, certain short-cut-methods can be used where assumed average is used for calculating the coefficient of correlation. This is what we have done for calculating standard deviation, in lesson-2. One of the following shortcut formula can be used.

$$r = \frac{\Sigma Dx.Dy - \Sigma Dx.Dy / n}{\sigma_1 \sigma_2}$$

Here,  $\Sigma Dx.Dy$  = sum of products of deviations from assumed means,

$\Sigma Dx$  = Sum of deviations of *x* values value from assumed mean  $X_a$

$\Sigma Dy$  = sum of deviations of *y* valises from, assumed mean  $y_a$

$\sigma_1$  = a standard deviations of *X* series,

$\sigma_2$  = as standard deviations of *y* series,

If in formula (3.4), we insert the formula of  $\sigma_1$  and  $\sigma_2$  the formula becomes

$$r = \frac{\Sigma Dx.Dy / n - (\Sigma Dx / n)(\Sigma Dy / n)}{\sqrt{\left\{ \frac{\Sigma Dx^2}{n} - \frac{(\Sigma Dx)^2}{n} \right\}} \times \sqrt{\left\{ \frac{\Sigma Dy^2}{n} - \frac{(\Sigma Dy)^2}{n} \right\}}} \quad \text{..... (3.4)}$$

$$r = \frac{\Sigma Dx.Dy - \Sigma Dx \Sigma Dy / n}{\sqrt{\left\{ \Sigma Dx^2 - \frac{(\Sigma Dx)^2}{n} \right\}} \times \sqrt{\left\{ \Sigma Dy^2 - \frac{(\Sigma Dy)^2}{n} \right\}}} \quad \text{..... (3.5)}$$

$$r = \frac{\eta \Sigma Dx.Dy - \Sigma Dx \Sigma Dy / \eta}{\sqrt{\{\eta \Sigma Dx^2 - (\Sigma Dx)^2\} \times \sqrt{\{\eta \Sigma Dy^2 - (\Sigma Dy)^2\}}} \quad \dots\dots (3.6)$$

(Work out yourself how - outside the square root signs in the numerator has been cancelled).

Formula (3.1) is the most convenient and most commonly, used from of the formula for calculating Persons correlation coefficient.

You will recall that

$$\bar{X}_a = \bar{X}_a + \frac{\Sigma Dx}{n} \text{ where } \bar{x} \text{ is, is arithmetic mesh of } x \text{ values } x_a \text{ is the assumed mean and}$$

$\Sigma Dx$  is the sum of deviations from assumed mean.

$$\therefore \frac{\Sigma Dx}{n} = \bar{x} - \bar{x}_a = \text{or } \Sigma Dx = n (\bar{x} - \bar{x}_a).$$

Similarly,  $\Sigma Dy = n (\bar{y} - \bar{y}_a)$

Substituting these values of Dx and Dy in formula (7.5), we get

$$r = \frac{\Sigma Dx.Dy - n(\bar{x} - \bar{x}_a / n)(\bar{y} - \bar{y}_a) / n}{\left( \Sigma Dx^2 - \frac{[n(\bar{x} - \bar{x}_a)]^2}{n} \right) \left( \Sigma Dy^2 - \frac{[n(\bar{y} - \bar{y}_a)]^2}{n} \right)}$$

$$r = \frac{\Sigma Dx.Dy - n(\bar{x} - \bar{x}_a) + (\bar{y} - \bar{y}_a)}{\sqrt{[\Sigma Dx^2 - n(\bar{x} - \bar{x}_a)^2]} \sqrt{[\Sigma Dy^2 - n(\bar{y} - \bar{y}_a)^2]}}$$

**Example II:** Calculate the co-efficient of correlation between the values of x and y given below;

x	78	89	96	69	59	79	68	61
y	125	137	156	112	107	136	123	108

**Solution:**

x	y	(x-69) =Dx	(y— 112) =Dy	Dx <sup>2</sup>	Dy <sup>2</sup>	DxDy
78	125	920	13	81	625	117 500
89	137	27	44	729	1936	1188
96	156	27	44	729	1936	1188
69	112	0	0	0	0	0
59	107	-10	- 5	100	25	50
79	136	-10	24	100	576	240
68	123	- 1	11	1	121	-11
61	108	- 8	- 4	64	16	32
$\Sigma Dx = 47$		$\Sigma Dy = 108$	$\Sigma Dx^2 = 1475$	$\Sigma Dy^2 = 3468$		$\Sigma Dx.Dy = 2116$

Assumed mean of x series = 69

Assumed mean of y series =112

Applying Formula (3.5)

$$r = \frac{\Sigma Dx.Dy - \Sigma Dx. \Sigma Dy / n}{\sqrt{\left(\Sigma Dx^2 - \frac{(\Sigma Dx)^2}{n}\right)} \sqrt{\left(\Sigma Dy^2 - \frac{(\Sigma Dy)^2}{n}\right)}}$$

$$r = \frac{2116 - (47 \times 108 / 8)}{\sqrt{\left(1475 - \frac{(47)^2}{8}\right)} \sqrt{\left(3468 - \frac{(108)^2}{8}\right)}}$$

$$r = \frac{11852}{(2591)(16080)} = .954$$

Simplification of the terms here is a tedious job. You must have lot of practice if you want to avoid, waste of your precious time in the examination. You can also use logarithms for this simplification.

If zero is taken as the assumed mean for both  $x$  and  $y$  variables, the deviations will be nothing but the original  $x$  and  $y$  values themselves. Formula. (3.5) will then become.

$$r = - \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}} \quad \text{..... (3.7)}$$

With this formula we can find co-efficient, or correlation if we are given :

- (1) The number of pairs of observations.
- (2) Sum of the X values and sum of Y values.
- (3) Sum of the squares of X values and sum of the squares of Y values.
- (4) Sum of the products of values of X and Y. variables.

### 7.3.3.3. Calculation of Coefficients of correlation in continuous series.

If the values of the two variables are grouped and the frequency of different groups are given two way tabulation is necessary in order to find out the coefficient of correlation. Suppose the two variables are marks obtained in Economics ( $x$ ) and marks obtained in Statistics ( $y$ ) and they have, been grouped in class intervals. They are given below:

y Marks In Statistics I	X Marks in Economics				Total $f_y$
	5—15	15—25	25—35	35-45	
0—10	1	1	—	—	2
10—20	3	6	5	1	15
20—30	1	8	9	2	20
30—40	—	3	9	3	15
40—50	—	—	4	4	8
Total $f_x$	5	18	27	10	60

We want to find the coefficient of correlation between marks in Economics and in Statistics of these 60 boys. Formula (3.5) will assume the following shape when frequencies are also given

$$r = \frac{\Sigma fDxDy - (\Sigma fDx \cdot \Sigma fDy) / n}{\sqrt{\left(\Sigma fDx^2 - \frac{(\Sigma fDx)^2}{n}\right)} \sqrt{\left(\Sigma fDy^2 - \frac{(\Sigma fDy)^2}{n}\right)}}$$

This is the most convenient formula for use in grouped data:

In the data given above, we can find  $\Sigma fDx$ ,  $\Sigma fDy$ ,  $\Sigma fDx^2$ ,  $\Sigma fDy^2$ . Please find out these values; Remember that deviations are being taken from assumed means/If you have any doubt anywhere, revise the techniques of finding standard deviation.

Understand this table very clearly and carefully. In the cell against 0—10 and below 5—15 groups, “we find the-value 1. This is the frequency. It means that there is 1 student whose score in Statistics is between 0—10 and in Economics’ between 5—15. Similarly, for all other entries in these cells.

**Examples III:** Calculate coefficient of correlation from the grouped data given above :

**Solution :**

$MARKS (ECO.)$				5-15	15-25	25-35	35-45			
$X \rightarrow$				10	20	30	40			
$MIDPOINT$				-10	0	10	20			
$D_x$				-1	0	1	2	Total		
$MARKS (STAT.)$				$MIDPOINT$	$D_y$	$fD_y$	$fD_y^2$	$fD_y^3$	$fD_y^4$	$fD_y^5$
$Y \downarrow$										
0 - 10	5	-20	-2	1	1	-	-	2	- 4	8 2
					2	0				
10-20	15	-10	-1	3	8	5	1	15	-15	15 -4
					3	0	-5 -2			
10-30	25	0	0	1	3	9	2	20	0	0 0
					0	0 0	0			
30-40	35	10	1	—	3	9	3	15	15	15 15
						2 2	6			
40-50	45	20	2	—	—	4	4	8	16	32 24
						8	16			
Total $f_x$				5	18	27	10	60	12	70 37
$fD'x$				-5	0	27	20	42		
$fD'x^2$				5	0	27	40	72		
$fD'xD'y$				5	0	12	20	37		

$$\begin{array}{ll}\Sigma fD'x = 42 & \Sigma fD'x = 12 \\ \Sigma fD'x^2 = 72 & \Sigma fD'y^2 = 72 \\ \Sigma fD'xD'y = 37 & x = 60\end{array}$$

Formula (7.8) for grouped data is

$$\frac{\Sigma fD'xD'y - (\Sigma fD'x \cdot \Sigma fD'y) / n}{\sqrt{\left\{ \left( \Sigma fDx^2 - \frac{(\Sigma fD'x)^2}{n} \right) \left( \Sigma fDy^2 - \frac{(\Sigma fD'y)^2}{n} \right) \right\}}}$$

$$D'x = \frac{Dx}{i}; D'y = \frac{Dy}{J}$$

where  $i$  and  $j$  are common factors in the series of  $Dx$  and  $Dy$ .

All the values required for this formula have been calculated in the lengthy table above,  $x$  values are given, horizontally, and  $y$  values vertically. Step deviations of the  $x$  values are given us — 1, 0, 1, 2 and of the  $y$  values as — 2, —1, 0, 2. Once the step deviations are taken, farther adjustments at any stage are not necessary. This is true even if common factors in the step deviations of  $x$  and  $y$  variable are unequal. Frequencies corresponding to step deviations of  $y$  variable occur in the row marked  $fx$  immediately below the calls. To find  $fD'x$ , multiply the frequency ( $fx$ ) by the corresponding step deviation given in the row immediately above the frequency cells.

$fD'x$  values are given in the row below  $fx$ .

To find  $fD'x^2$  simply multiply  $fD'x$  by  $D'x$ . Summations of  $fD'x$  and  $fD'x^2$  gives us  $\Sigma fD'x$ .

Frequency corresponding to step deviations by  $y$  variable are given in the column  $fy$  immediately after the frequency cells.

By multiplying  $fy$  by step deviations we can get  $fD'y$  and multiplying  $fD'y$  further again by  $D'y$  we get value  $fD'y^2$  their summations gives  $\Sigma fD'y^2$  and  $\Sigma fD'x^2$ .

$n$  is the sum of frequencies.

$\Sigma fD'xD'y$  still remains to be found  $fD'xD'y$  is calculated in each frequency cell. For example, in the frequency cell against  $D'x = -1$  and  $D'y = -2$ , frequency  $f = 1$ ,  $\therefore fD'xD'y = 1$  and  $D'y = -1$   $f = 5$ ,  $\therefore fD'xD'y$  as  $5(1)(-1) = -5$ . All values of  $fD'xD'y$  written in the frequency cell have been underlined in order to distinguish them from the frequencies. These  $fD'xD'y$  values have been added in the last of columns as well as last of rows.

This sum (whether obtained vertically from the column or horizontally from the row) will be  $\Sigma fD'xD'y$ . In this example  $\Sigma fD'xD'y = 37$ .

This is a lengthy table. You will have to exert considerably to understand it at once. Its sheer length and breadth should not frighten you into leaving it altogether. From the table :

$$\begin{array}{ll}n = 60 & fD'y = 12 \\ fD'x = 42 & fD'y^2 = 70 \\ fD'x^2 = 72 & fD'xD'y = 37\end{array}$$



$$\begin{aligned}
r &= \frac{\Sigma fD' xD' y - (\Sigma fd' x. \Sigma fd' y) / n}{\sqrt{\left(\Sigma fDx^2 - \frac{(\Sigma fdx)^2}{n}\right)} \sqrt{\left(\Sigma fD' y^2 - \frac{(\Sigma fdy)^2}{n}\right)}} \\
&= \frac{37 - 42 \times 12 / 60}{\sqrt{\left(72 - \frac{(42)^2}{60}\right)} \sqrt{\left(70 - \frac{(12)^2}{60}\right)}} \\
&= \frac{37 - 8.4}{\sqrt{\left(\frac{213}{5} \times \frac{338}{5}\right)}} \\
&= \frac{28.6}{53.66} = .533
\end{aligned}$$

### Probable Error of the Coefficient of Correlation

The exact value, of various types of errors will be understood when we study sampling methods. We can however, get a working idea of the Probable Error of the correlation coefficient at this stage also.

Suppose out of a student population of 100,00 we select 100 students at random and compute the correlation coefficient between heights and weights, Suppose this coefficient is  $r = .8$  and suppose further that we select 200 more, random. Samples of 100 students each from this population, if we calculate correlation coefficients in respect of each of these, samples, there will be certain limits within which these coefficient will probably lie.

“Probable error of the coefficient of correlation is the amount which if added to and subtracted from the mean correlation coefficient gives, amounts within which the chances are even that a coefficient of correlation from the series selected at random will fall.”

The formula for P. E. of Karl Pearson’s correlation—coefficient is

$$P.E., \text{ or } r = 0.6745 \frac{1 - r^2}{\sqrt{n}}.$$

Where  $r$  is the coefficient of, correlation and  $n$  is the number of pairs of observations.

If in a particular sample of 100 students, the correlation between heights and weights is 0.8 the probable error will be

$$\begin{aligned}
0.6745 \times \frac{1 - r^2}{\sqrt{n}} &= 0.6745 \frac{1 - 8^2}{\sqrt{(100)}} \\
&= 0.6745 \frac{1 - 64}{10} \\
&= 0.6745 \frac{(0.36)}{10} = .24282/20
\end{aligned}$$

The coefficient of correlation in this case will, therefore, lie between and can be written as  
 $r = 0.8 \pm 0.024282$ .

If more samples were taken and their correlation coefficient computed, they will probably lie within  $8 \pm 0.024282$ , i.e. between .715718 and .824282.

### Interpretation of Coefficient of Correlation

Coefficient of correlation for a given 'pair' of variables lies between + 1 (perfect positive correlation) and —1 perfect negative correlation.

It may be asked whether the coefficient of correlation between two variables is "significant" or not. In order to answer this the following points may be kept in mind, they are based on the use of probable error of the coefficient.

1. If  $r$  is less than its probable error, it is not at all significant, i.e. there is no evidence of correlation.
2. If  $r$  is more than six times its probable error, correlation is significant.
3. If the probable error is small and  $r$  is 5 or more, correlation decided exists and  $r$  is significant.
4. Even if probable error is small but  $r$  is less than 3, correlation should be considered are not marked and  $r$  should not be considered as significant.

You can now take a few illustrations and decide for yourself whether  $r$  is significant or not.

### Coefficient of Correlation by Rank Differences

Method of rank differences, is used to calculate correlation coefficient in, those cases where it is possible to arrange the various items of a series in a serial order, but the quantitative measurement of values is impossible or difficult or unpractical. Many attributes are incapable of direct measurement for example; intelligence, honesty beauty character etc. But the observer may be in a position to arrange items in a serial order and to assign ranks to different items. This method can be used also in such, places where dependable measurements cannot be obtained because of lack of finances or absence of investigation. This method involves easier mathematical calculations and it may be preferred for the reason also.

The next two examples will clarify this method.

**Example VI :** Calculate coefficient of rank correlation from data of Example II above.

**Solution:**

I $x$	II Ranks	III $y$	IV Ranks	V Difference or Ranks ( $d$ )	VI $d^2$
78	5	125	5	0	0
89	7	137	7	0	0
97	8	156	8	0	0
69	4	112	3	1	1
59	2	107	1	1	2
79	6	136	6	0	0
68	3	123	4	— 1	2
57	1	108	2	— 1	1
				$\Sigma d = 0$	$\Sigma d^2 = 4$

Values of  $x$  variable appear in column I and  $y$  variable in column III.

Our first task is to assign to  $x$  and  $y$  values. (In some problems, not values but ranks are given. There we do not have to assign the ranks).

Start with  $x$  values. We can assign rank 1 either to the smallest or the largest value. We have assigned rank 1 to the smallest value 57, The next large value is 59 which has been given rank 2. Next large value is 60 which is given rank 3. The process is continued till we reach the largest value 97 of the series which has been given rank 8. The ranks of  $x$  values appear in column II.

The process is repeated in  $y$  series, starting with the smallest "value 107 rank (7) and going up the largest value 156 (rank 8),  $y$  rank appear in column IV above.

Next we find the differences between rank of  $x$  and  $y$  written in column V as (d). It makes no difference whether  $x$  ranks are subtracted from  $y$  ranks or vice-versa;  $y$  ranks from  $x$  ranks. The sum of these differences will be zero.

In column VI are given the squares, of these rank difference  $d^2 \Sigma d^2 = 4$ .

Coefficient of rank correlation or

$$e = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} : [\Sigma d^2 = 4, n = 8]$$

$$e = 1 - \frac{6 \times 4}{8(64 - 1)} = 1 - \frac{24}{504}$$

$$= \frac{480}{504} = .95$$

Rank correlation coefficient is also called Spearman's correlation coefficient. The numerical value of rank correlation coefficient need not be the same as that of Karl Pearson's coefficient for the same data.

**Example V.** Calculate the coefficient of rank correlation for the following data:

$x$	48	33	40	9	16	16	65	24	16	57
$y$	13	13	24	6	15	4	20	9	6	19

(PU 1964)

I X	II Ranks	III V	IV Ranks	V Difference of ranks (d)	VI d <sup>2</sup>
48	8	14	5.5		
33	6	13	5.5	2.5	6.25
40	7	24	10	.5	.25
16	1	6	2.5	-3.0	9.00
16	3	15	7	-1.5	2.25
65	3	4	2	-4.0	16.00
24	10	20	9	2.0	4.00
16	5	9	4	1.0	1.00
57	3	6	2.5	1.0	1.00
	9	19	8	5	.25
				1.0	1.00
$\Sigma d = 0 \quad \Sigma d^2 = 41.10$					

A problem arises here in the assigning, of ranks because some values both in  $x$  and  $y$  series appear more than once: This is how the problem is resolved.

Let us start with  $x$  values. The 'smallest value 9 is given rank 1. The next higher value 16 occurs three times. If these three values of 16 had differed from one another, they would have been given the ranks 2, 3 and 4. Now that they don't differ, average of the ranks 2,3 and 4 i.e  $\frac{2+3+4}{3}$  or 3 rank is assigned to all the three 16's. The next higher value 24 is given rank 5 Because ranks, 2, 3, 4 have been exhausted on 16's. Therefore, whenever ties occur, the items are given the average of the ranks, they would have received if they had slightly differed.

In  $y$  series, rank 1 is given to 4, the smallest value. After 4 comes 6, two times, if the two times had differed slightly, the rank assigned would have been 2, and 3, Average of 2 and 3 is 2.5. Therefore both's receive the rank 2.5 is given rank 4, 13 occurs two times therefore both 13's receive the ranks  $\frac{5+6}{2} = 5.5$  and so on.

You would have noticed that  $x$  rank appear in column II and  $y$  ranks in column IV. Rank differences are given in column V, squares of the ranks :  $d^2$  in column VI.

$$\begin{aligned}\text{Here } n &= 10 \\ \Sigma d^2 &= 41\end{aligned}$$

Due to common ranks, coefficient of correlation has to be modified and the following, is used:

$$e = 1 - \frac{6\Sigma d^2 + \Sigma \frac{1}{12}[(m^3 - m)]}{n(n^2 - 1)} \quad \dots (3.11)$$

Here  $m$  stands for the number or times that a value has been repeated. In our example above three values, 16, 6 and 13 have occurred repeatedly, therefore  $1/12 (m^3 - m)$  will be added three times. Since 16 repeats 3 times, 6 two times and 13. two times, the value of  $m$  will be successively 3,2 and 2.

$$\begin{aligned}e &= 1 - \frac{6\Sigma d^2 + 1/12 \Sigma [(m^3 - m)]}{n(n^2 - 1)} \\ &= 1 - \frac{6[41 + 1/12(3^3 - 3) + 1/12(2^3 - 2) + 1/12(2^3 - 2)]}{10(10^2 - 1)} \\ &= 1 - \frac{6\left[41 + 2 + \frac{1}{2} + \frac{1}{2}\right]}{10 \times 99} \\ &= 1 - \frac{262}{990} = \frac{728}{990} \\ &= + 73\end{aligned}$$

A lot of practice is needed to master the techniques of calculating Karl Pearson's, and Spearman's coefficient of correlation.

The independent variable the other as dependent variable.' After estimating the relationship between two variables one can predict the most likely values of dependent variable on the basis of given values of independent variable.

### **Regression Lines**

A regression line is used to show the functional relationship between the variables and  $y$ . It is a line of "average relationship" as it 'shows the relationship between  $x$  and on the average.

If we have a problem of estimating  $y$  variable for some given values of  $x$  variable the requirement will be to construct a mathematical : equation of the form ( $y = a + bx$ ). Such that  $y$  is dependent variable and  $x$  is independent variable. Such an equation known as regression equation of  $y$  on  $x$  can be used to estimate the most likely value of  $y$ . for some given value of  $x$ . Similarly, the problem which involves, the estimate of most probable or mean value of  $x$  for some given value of  $y$  will require the construction of regression equation  $x$  and  $y$ . This shows that from a set of observation of  $X$  and  $Y$  we can form two regression equation.

This can further be illustrated with the help of the given 'example' Suppose, we want to study the relationship between the demand for wheat, and the, price of wheat in a certain market for a specified period. If our purpose is to estimate the demand for wheat corresponding to some arbitrarily selected price level's then the regression line of wheat on price will ignore' the variation of demand from one buyer to the other at the same price. Corresponding to this particular price, we may notice that some buyer is demanding larger quantity than some other corresponding' each given price level. We may provide different demand levels. But the regression line of demand for wheat and price of wheat will give an estimate, of demand for wheat most likely to be bought at price, i.e. demand on the average. It shows what happens on the average to demand total variation in the demand for wheat may be split up into two parts. One part is explained by the regression line while the other part is not explained.'

### **Why there are two regression lines ?**

Although both regression equation involve  $x$  and  $y$ , the two equation cannot be used interchangeably neither can be employed to predict both  $x$  and  $y$ . This is an important fact which the student must bear in mind constantly. The first regression equation  $y = r\sigma_y/\sigma_x \times x$  can be used only when  $y$  is to be predicted from a given  $x$  (when  $y$  is the dependent variable). The second regression

equation  $x = r \frac{\sigma_x}{\sigma_y} \times y$  can be used only when  $x$  is to be predicted from a known  $y$  (when  $x$  is the dependent variable).

In summary, there are two regression equations in a correlation table, the one through means of the columns and the other through the means of the rows. This is always true unless, the correlation is perfect.

When  $r = 1.00$   $\bar{y} = r \frac{\sigma_y}{\sigma_x} \times \bar{x}$  becomes  $\bar{y} = \frac{\sigma_y}{\sigma_x} \bar{x}$  or  $\bar{y} \sigma_n = \bar{x} \sigma_y$ . Moreover when  $r = 1.00$ ,

$\bar{x} = r \frac{\sigma_n}{\sigma_y} y = y$  becomes  $\bar{x} = r \frac{\sigma_x}{\sigma_y} \times y$  or  $\bar{x} \sigma_y = \sigma_x y$ . In short when the correlation is perfect, the

two regression equation are identical and two regression lines coincide.

## Regression Coefficients

The least square regression equation, of  $y$  on  $x$  is written as

$$y = a_0 + b_0 x.$$

where  $a$ , and  $b$ , are two constants and can be obtained from the normal equations.

## Regression Coefficients

The term regression coefficient is the name of the slope of regression line. Since there are two regression lines there will be two regression coefficients. The slope of regression line, of  $y$  on  $x$  is represented by the regression coefficient of  $y$  on  $x$  and has been denoted by the symbol  $b$ . Another convenient symbol is  $b_{yx}$  and it measures the change in  $y$  corresponding to unit change in  $x$ . When deviations are taken from the means then by

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{\sum xy}{n \sigma x^2} = \frac{r \sigma y}{\sigma x}$$

Similarly the regression coefficient of  $x$  on  $y$  is the slope of the regression line of slope of the regression line of  $x$  on  $y$ .

(2) Regression Equation of  $x$  on  $y$ :

Regression 'Equation of  $x$  on  $y$  is  $x - \bar{x} = b_{xy} (y - \bar{y})$

$$x - b_{xy} \bar{y}$$

$$x = \left( \frac{\sum xy}{\sum y^2} \right) y$$

The alternative method is based on the assumption that deviations  $x$  and  $y$  are taken from the means of  $X$  and  $y$  respectively

Regression Equation of  $x$  on  $y$

$$x - \bar{x} = r \frac{\sigma x}{\sigma y} (y - \bar{y})$$

and Regression equation of  $y$  on  $x$  is

$$y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x}) = r$$

from these we get

$$a_1 = \frac{(\sum y) (\sum y^2) (\sum x) (\sum xy)}{N \sum y^2 - (\sum y)^2}$$

$$\bar{X} = b_{xy}$$

$$b_1 = \frac{N \sum xy - (\sum x) \sum y}{N (\sum y^2) - (\sum y)^2}$$

## Alternative Method :

Regression equations of  $y$  on  $x$  and of  $x$  on  $y$  can also be written in an alternative way.

(1) Regression equation of  $y$  on  $x$  is

$$y - \bar{y} = b_0 (x - \bar{x})$$

$$\text{or } Y = b_0 x$$

$$Y - \bar{y} = y$$

$$X - \bar{x} = x$$

$$b_0 = \frac{\sum xy}{\sum x^2}$$

$$a_0 = \bar{Y} = b_0 \bar{X}$$

$$y = b_0 x \text{ becomes}$$

$$y = \left( \frac{\sum xy}{\sum x^2} \right) x$$

$$\sum Y a = a_0 N + b_0 \sum X$$

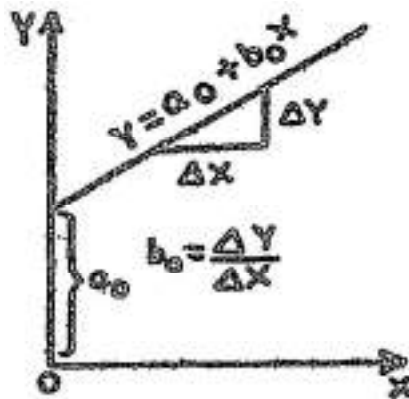
$$\sum XY = a_0 \sum X + b_0 \sum X^2$$

from these we have

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2}$$

$$\text{or } a_0 = -b_0 \bar{Y}$$

$$b_0 = \frac{N(\sum XY) - (\sum X)(\sum Y)}{N(\sum X^2) - (\sum X)^2}$$



$a_0$  is Y intercept and  $b_0$  is the slope coefficient.

### Regression Equation of X on Y

The least sq regression equation of X on Y is written as :

$$X = a_1 + b_1 Y$$

where  $a_1$  and  $b_1$  are constants and can be determined By the following normal" equations.

$$\sum X = a_1 N + b_1 \sum Y$$

$$\sum XY = a_1 \sum Y + b_1 \sum Y^2$$

it is denoted, by  $b_1$  and another convenient symbol is  $b_{xy}$ . As noted above

$$b_{xy} = \frac{\sum xy}{\sum Y^2}$$

**Example 1**

X	2	3	4	5	6
Y	3	5	8	7	12

Suppose the following data are observed

We shall use these data to illustrate the procedure of working calculation. We want to form the line  $y_c = a_0 + b_0 X$  by the method of least squares. Following table shows the calculations of various quantities needed for the estimation of  $a_0$  and  $b_0$ .

X	Y	z	y	xy	x <sup>2</sup>	y <sup>2</sup>
2	3	-2	-4	8	4	16
3	5	-1	-2	2	1	4
4	8	0	1	0	0	1
5	7	+1	0	0	1	0
6	12	+2	5	10	4	25
-	-	-	-	-	-	-
20	35	$\hat{\Sigma}x = 0$	$\hat{\Sigma}y = 0$	20	$\Sigma x^2 = 10$	$\Sigma y^2 = 46$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{20}{5} = 4$$

$$\bar{Y} = \frac{\Sigma y}{N} = \frac{35}{5} = 7$$

$$b_0 = \frac{\Sigma xy}{\Sigma x^2} = \frac{20}{10} = 2$$

$$a_0 = \bar{Y} - b \bar{X} = 7 - 2(4) = -1$$

$$Y_c = 2X - 1$$

if our aim is to form regression line

$X_c = b_1 + b_1 Y$  then we have to estimate  $a_1$  and  $b_1$ .

$$\Sigma Y^2 = 46$$

$$b_1 = \frac{\Sigma xy}{\Sigma y^2} \dots \frac{20}{46} \dots \frac{10}{23}$$

$$a_1 = \bar{X} - b_1 \bar{Y} = 4 - \frac{10}{23} \times 7 = \frac{92 - 70}{23} = \frac{22}{23}$$

$$X_c = \frac{22}{23} + \frac{10}{23} y$$

**Important Remarks**

(1) Both the regression lines pass through  $\bar{X}$  and  $\bar{Y}$ . In example above.

$$Y = 2X - 1 \dots \dots \dots \text{I}$$

$$X = \frac{22}{23} + \frac{10}{23} Y \dots \dots \dots \text{II}$$



$$\text{From II} \quad X = \frac{22}{23} + \frac{10}{23} (2X + 1)$$

$$X = \frac{22}{23} + \frac{20}{23} - \frac{10}{23}$$

$$X = \frac{12}{23} + \frac{20}{23}$$

$$23X = 12 + 20X$$

$$3X = 12$$

$$X = 4$$

Putting  $X = 4$

we get

$$Y = 8 - 1 = 7$$

Hence the point (4,7) Satisfies both the regression equations, Thus  $\bar{X} = 4$  and  $\bar{Y} = 7$ .

(2) As is clear from above when two regression equations are given, then solving them simultaneously for X and Y will give mean value of X and mean value of Y.

(3) Making predictions is an important aspect of regression analysis. When we are to predict Y given X, then the most likely value of Y is to be predicted by using regression equations in which Y is the dependent variable and X is independent variable. Similarly when the problem is to predict X given Y, then regression equation of X on Y is to be used.

Suppose we are to find most likely value of Y when X is 12. For this, we shall use regression equation  $Y_o = 2x (12) - 1$ . From this

$$Y_o = 2 (12) - 1 = 23$$

(4). If two regression equations are given then we can find out correlation coefficient by calculating the Geometric mean of two regression coefficients.

In the above example

$$Y_o = 2x - 1 \dots\dots\dots\text{I}$$

$$\text{and} \quad X_o = \frac{22}{23} + \frac{10}{23} Y \dots\dots\dots\text{II}$$

$$\text{From (I) } b_o (=b_{yx}) = 2$$

$$\text{From (II) } b_l (=b_{xy}) = \frac{10}{23}$$

$$r^2 = 2 \times \frac{10}{23}$$

$$r = \sqrt{\frac{20}{23}} = .93$$

### Properties of Regression Coefficient

(1) The geometric mean between regression coefficient is coefficient of correlation Proof

$$\begin{aligned}
b_0 &= b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \\
b_1 &= b_{xy} = \frac{\Sigma xy}{\Sigma y^2} \\
b_{yx} b_{xy} &= \frac{\Sigma xy}{\Sigma x^2} \times \frac{\Sigma xy}{\Sigma y^2} \\
&= \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} \\
&= r^2 \\
&= r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}
\end{aligned}$$

This proof can also be given as,

$$\begin{aligned}
b_{yx} b_{xy} &= \frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_y}{\sigma_y} \\
&= r^2 \\
r &= \pm \sqrt{b_{yx} b_{xy}}
\end{aligned}$$

### Important Remarks

- (a) If  $b_{xy}$  is positive then  $b_{yx}$  will also be positive.
- (b) If  $b_{yx}$  is negative, then  $b_{xy}$  will be negative.
- (c) Both regression coefficient, must have the same sign.  
If  $b_{yx}$  and  $b_{xy}$  are both positive, then  $r$  will be positive and when  $b_{yx}$  and  $b_{xy}$  are negative then  $r$  will also be negative.
- (d)  $b_{xy}, b_{yx} \leq 1$

2. If one regression coefficient is greater than unity then other regression coefficient must be less than unity.

**Proof:** Let  $b_{xy} > 1$ , then we are to show that  $b_{yx} < 1$ .

$$\text{If } b_{yx} > 1 \text{ then } \frac{1}{b_{xy}} < 1$$

$$\text{Now } b_{yx} b_{xy} \leq 1$$

$$\text{or } b_{xy} \leq \frac{1}{b_{yx}}$$

$$\text{From (1) } \frac{1}{b_{yx}} < 1$$

$$\therefore b_{xy} < 1$$

2. Arithmetic mean of  $byx$  and  $bxy$  is equal to or greater than coefficient of correlation.

**Proof :** We are to prove that

$$\frac{byx + bxy}{2} \geq r$$

$$\text{if } \frac{byx + bxy}{2} \geq r \text{ then}$$

$$byx + bxy \geq 2r$$

$$\text{or } byx + bxy \leq 21 \pm \sqrt{byx.bxy}$$

$$\text{or } (\sqrt{byx + bxy}) \pm 2 \sqrt{byx.bxy} \geq 0$$

$$\text{or } (\sqrt{byx + bxy})^2 \leq 0 \text{ which is always, true.}$$

$$\text{Hence } \frac{byx + bxy}{2} \geq r$$

$$\text{or } byx + bxy \geq 2r$$

$$\text{or } \frac{r\sigma_y}{\sigma_x} + \frac{r\sigma_x}{\sigma_y} \geq 2$$

$$\text{or } \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2$$

$$\text{or } \frac{\sigma_y^2 + \sigma_x^2}{\sigma_x \sigma_y} \geq 2$$

$$\text{or } \sigma_y^2 + \sigma_x^2 \geq 2 \sigma_x \sigma_y$$

$$\text{or } \sigma_y^2 + \sigma_x^2 - 2 \sigma_x \sigma_y \geq 0$$

$$\text{or } (\sigma_y - \sigma_x)^2 \geq 0 \text{ which}$$

shows that Regression coefficients, are independent of origin but not of scale.

**Example (II)**

**Using the following data**

	X series	Y series
Mean value	12	10
Standard deviation	4	3

Coefficient of correlation X and Y is 0.8

(a) Form two regression lines.

(b) Predict most likely value of Y when X is = 11 and most likely value of X when Y is = 13.

**Sol.**

(a) Let the least square regression line of Y on X be

$$y_c = a + bx$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{r\sigma_y}{\sigma_x} = \frac{0.8 \times 3}{4} = 0.6$$

$$\begin{aligned}
 a &= Y - ax \\
 &= 10 - 0.6 (12) \\
 &= 2.8
 \end{aligned}$$

Hence  $y_c = a + bx$   
 $y_c = 2.8 + 0.6 X$

Let the least square regression line of X on Y be

$$\begin{aligned}
 X_c &= c + dY \\
 \frac{\sum xy}{\sum y^2} &= r \frac{\sum x}{\sum y} = \frac{0.8 \times 4}{3} = \frac{16}{15} = 1.07 \\
 c &= \bar{X} - d\bar{Y} \\
 &= 12 - 1.07 (10) \\
 &= 1.3 \\
 xc &= 1.3 + 1.7 Y
 \end{aligned}$$

(b) Most likely value of Y when X is 11

For this we need, regression line of Y on X which is

$$Y_c = 2.8 + 0.6 X$$

Putting  $X = 11$  we get

$$\begin{aligned}
 Y_c &= 2.8 + 0.6 (11) \\
 &= 2.8 + 6.6 \\
 &= 9.4
 \end{aligned}$$

Most likely value of X when Y is 13.

For this prediction, we need regression equation of X on Y which is

$$\begin{aligned}
 X_c &= 1.3 + 1.07y \\
 \text{when } Y &\text{ is } 13 \\
 &= 1.3 + 1.07 \times (13) \\
 &= 15.21
 \end{aligned}$$

### Example (iii)

Two regression equations are given as

$$\begin{aligned}
 x - 4y &= -13 \\
 9Y - X &= 53 \\
 \text{and } ax &= 12
 \end{aligned}$$

- (a) Find mean value of X and Y
- (b) Coefficient of Correlation
- (c) Standard deviation of Y.

$$\begin{aligned}
 x - 4y &= -13 \dots\dots\dots \text{I} \\
 -x + 9y &= 53 \dots\dots\dots \text{II} \\
 \text{Adding I and II} \\
 5y &= 40 \\
 y &= 8 \\
 \bar{y} &= 8
 \end{aligned}$$

Similarly  $\bar{x} = 19$

(b) Coefficient of Correlation

Assume that  $X - 4y = -13$  is regression line of Y on X and  $aY - X = 53$  is regression line of X on Y.

From Y on X equation

$$X - 4y = -13$$

$$-4y = -X - 13$$

$$4Y = X + 13$$

$$Y = \frac{1}{4}x + \frac{13}{4}$$

$$\text{Hence } b_{yx} = \frac{1}{4}$$

From X on Y equation

$$9Y - X = 53$$

$$-X = -9Y + 53$$

$$X = 9Y - 53$$

$$\text{Hence } b_{xy} = 9$$

$$\text{Now } b_{xy} b_{yx} = \frac{1}{4} \times 9 > 1$$

This shows our assumption is wrong. Therefore we take

$$x - 4y = -13 \text{ as X on Y}$$

$$\text{and } 9Y - X = 53 \text{ as Y on X}$$

$$b_{xy} = 4 \text{ and } b_{yx} = \frac{1}{9}$$

$$r = \sqrt{4} \times \frac{1}{9} = \frac{2}{3}$$

$$r = 667$$

(c) Standard deviation of Y.

$$\text{Now } b_{xy} = 4$$

$$r \frac{\sigma_x}{\sigma_y} = 4$$

$$r \sigma_x = 4\sigma_y$$

$$\therefore \sigma_y = r \frac{\sigma_x}{4}$$

$$= \frac{2}{3} \times \frac{12}{4}$$

$$= 2$$

$$\sigma_y = 2$$

### Limitations of the Theory of Linear Correlation

(i) Correlation analysis suffers from serious limitations as a technique for the study of economic relationships

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x^2} \sqrt{\sum y^2}}$$

The above formulae for  $r$  is applicable only when the relationship between the two variables is linear. However, two variables may be strongly connected with a non linear relationship.

The students should also note it well that zero correlation and statistical independence of two variables ( $x$  and  $y$ ) are two different things and they should not treat, them as one and the same, thing. Zero correlation implies zero. covariance of  $x$  and  $y$  so that

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = 0$$

Statistical independence of  $x$  and  $y$ . implies that the probability of  $x_i$  and  $y_i$  occurring simultaneously is the simple product of the individual probabilities.

$$P(x \text{ and } y) = P(x) \cdot P(y).$$

(ii) The ‘second, limitation of the theory is that although the correlation coefficient is a measure of the co variability of variables,’ it does not necessarily imply any functional relationship, between the variables concerned. Correlation theory does not establish, and/or prove any causal relationship between the variables, It seeks to discover if a co variation exists, but it does not suggest that variation in, say,  $y$  are ‘caused’ by variation in  $x$ , or vice versa knowledge of the value of  $r$ , alone, will not enable us’ to predict the value of  $y$  from  $x$  ..... A high correlation between variables  $y$  and  $x$  may describe any one of the following situations:

- (1) Variation in  $x$  is the cause of variation in  $y$ .
- (2) Variation in  $y$  is the cause of variation in  $x$ .
- (3)  $y$  and  $x$  are jointly dependent.
- (4) There is ‘another common factor (2); that affects  $x$  and  $y$  in such way as to show close relation between them.
- (5) The correlation between  $x$  and  $y$  may be due to chance.
- (6) Qualitative phenomena cannot be computed.
- (7) When the number of observations is large, the calculation, of correlation coefficient becomes laborious and time consuming.

In summary the linear correlation coefficient measures the degree to which the points, cluster around a straight line, but it does not give the equation for the line, that it does not assign numerical values to the parameters of the function which is represented by this lines. These parameters are elasticity (or components of elasticity’s) and the knowledge of their numerical value is of particular interest both to entrepreneurs and policy makers.

### SUGGESTED READEINGS

1. Gupta S.P. : Statistical Methods.
2. Draper, N. and H. Smith : Applied Regression Analysis, John Wiley : New York, 1966.
3. Stevenson, W.J. : *Business Statistics* Concept and Applications, Harper Row; New York, 1978.

\*\*\*\*\*

## **LESSON-4**

### **FITTING OF REGRESSION EQUATION AND STANDARD ERROR OF ESTIMATE**

Dear students,

In the preceding lesson, correlation coefficients and limitation of correlation coefficients were discussed. The present lesson deals with basic statistical methods for studying relationship between two or more variables. In economics we are seldom able to give exact prediction of the value of a variable from a knowledge of the value of other variable. Rather, if a relationship Between two variables  $X$  and  $Y$  be established, the relationship can tell us the value of  $y$  which on the average 'can be expected to' be associated with a given value of  $X$ . Relationships of this kind are 'also referred to as stochastic relationship in contrast with exact relationships.

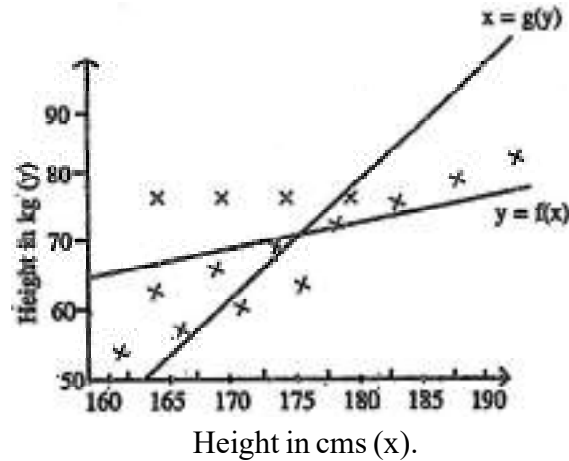
If a stochastic relationship between two or more variables can be expressed by a mathematical equation, so that on the basis of this equation we can estimate the average value of  $y$  associated with, the given  $X$ s, the method of analysis is known as regression analysis. In regression analysis we are concerned with statistical, not functional relationships among variables.' Although regression analysis deals with the dependence of one variable on other variable, it does not necessarily imply causation. In the words of Kendatt and Stuart : "A statistical relationship, however strong and however suggestive, can never establish causal connection : Our ideas of causation must come from outside statistics, ultimately from some theory or other."

In regression analysis, the variable whose average value is being estimated' is called the dependent variable. The variables on which the estimate is to be based are referred to as the independent or explanatory variables. When a regression relationship contains only one independent variable, it is referred to as a two-variable or simple regression. When more-variables are being used to explain, the behaviour of the dependent variable  $y$ , the analysis is known as multiple regression. Where the main objective of regression analysis is to investigate the nature of relationship between two variables  $x$  and  $y$ , the correlation coefficient measures the strength or degree of linear association between two variables. For example, we may be interested in finding the correlation coefficient between smoking and lung cancer. But in regression analysis, we are not interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of fixed value of other variables.

There are some fundamental differences hi the two techniques of regression and correlation which are worth noting. In, regression analysis there is an asymmetry in the way "the dependent and explanatory variables are treated. The dependent or explained variable is assumed to be statistical, random or stochastic, that is, to have a probability distribution. The explanatory, variables on the other hand, are assumed to have fixed values. In correlation, both variables are, assumed to be random, whereas in regression analysis dependent variable is stochastic but the explanatory 'variables are fixed.

## Descriptive Measures of Regression for ungrouped data.

Suppose we wish to investigate the relation between the height and weight of adult males for some given population. If we plot the pair  $[x,y] = [\text{height, weight}]$ , a diagram like figure 1 'will result. Such a diagram, is conventionally called a scatter diagram.



Note that for any given height there is a range of observed weights and vice-versa. This variation will be partially due to measurement error but primarily due to variation - between individuals. Thus no unique relationship between actual height and weight can be expected. But it can be observed that average observed weight for a given observed height increases as height increases. The locus of average observed weight for a given observed Height (as height varies) is called a regression curve of weight on height. Let us denote it by  $y = f(x)$ . There also exists a regression curve of height on 'weight similarly defined which we can denote by  $x = g(y)$ . A pair of random variables such as (height, weight) follows some sort of bivariate probability distribution when we are concerned with the dependence of a random variable  $y$  on quantity  $X$ , which is a variable but not a random variable an equation that relates  $y$  to  $x$  is usually called a regression equation.

### Linear Regression

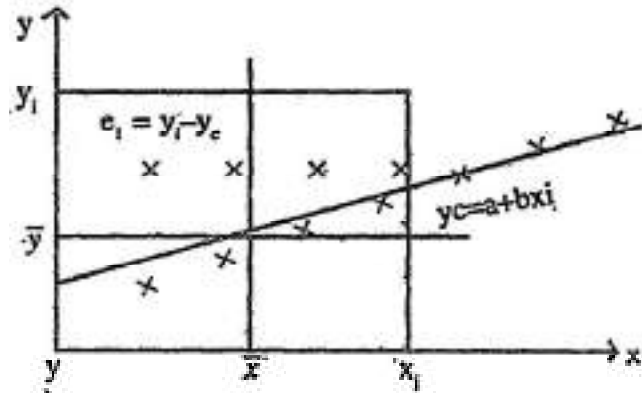
The simplest, and' most commonly used relationship between, two variables  $x$  and  $y$  is that of a straight line. We may write, the linear, first order model as.

$$Y_c = a + bx + \varepsilon \quad \text{..... (4.1)}$$

That is, for a given  $x$ , a corresponding observation  $y_c$  consist of the value  $a + bx$  plus an amount  $\varepsilon$ , the increment, by which an individual  $y$  may fall of the regression line. Equation (4.1) is the model of what we believe  $a$ ,  $b$  are called the parameters of the model  $a$  being the intercept of the straight line on the  $y$  axis and  $b$  its slope. The constant  $b$  is known as the regression coefficient of  $y$  on  $x$ .

Clearly we want to determine the values of  $a$  and  $b$  in such a way that the fitted line is as close as possible to the  $N$  plot points, i.e. we want to minimize the overall discrepancy between the plot points and the line, by figure 4.2, consider the typical pair, of observations, denoted by  $(x_i, y_i)$ . If the line fitted to the pointy has, intercept  $a$  and slope  $b$ , then the value of  $y$  computed from, the relationship which  $x$  is  $Y_c = a + bx_i$  and the deviation of the observed value  $y_i$  from the computed value  $y_c$  is measured by  $e_i = Y_c - Y_c$ .





The deviation  $e_i$  is called a residual. Eventually, the residuals can be positive or negative as the actual point lies above or below the fitted line. Since the positive negative residuals tend to cancel, out, the summation of these i.e.  $\sum e_i$  cannot be used as a measure of overall discrepancy. However, if these are squared and summed, then it is possible to make them as small as possible.

$$\sum e_i^2 = f(a, b)$$

The principle of least square is that  $a$  and  $b$  are to, be chosen so that  $\sum e_i^2$  is a minimum i.e. the sum of squares of the vertical deviations of the observed points from the line is to be a minimum.” This procedure is known as fitting a curve by the method of least squares. An important advantage of the method is that while being computationally simple, it also yields estimators with certain desirable statistical properties like unbiasedness, efficiency and consistency.

To determine the values of  $a$  and  $b$  that satisfy the above requirements, the partial derivatives of the sum with respect to  $a$  and  $b$  should both be zero

$$\begin{aligned} e_i &= Y_i - Y_c \\ e_i^2 &= (Y_i - Y_c)^2 \\ \sum e_i^2 &= \sum (Y_i - Y_c)^2 \\ &= \sum (Y_i - a - bx_i)^2 \end{aligned}$$

Differentiating  $\sum e_i^2$  with respect to ( $\therefore Y_i = a + bx_i$ )  $a$  and  $b$ , we have

$$\frac{\partial \sum e_i^2}{\partial a} = -2 \sum (y_i - \hat{a} - \hat{b}x_i)$$

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial b} &= -2 \sum x (y_i - \hat{a} - \hat{b}x_i) \\ &= -2 \sum (xy - ax - bx^2) \end{aligned}$$

For  $\sum e_i^2$  to be minimum  $\frac{\partial \sum e_i^2}{\partial a}$  and  $\frac{\partial \sum e_i^2}{\partial b}$  must both equal to zero, which they will do when

$$\begin{aligned} &\sum (y - \hat{a} - \hat{b}x) = 0 \\ \text{and} &\sum (xy - ax - bx^2) = 0 \\ \text{i.e. when} &\left. \begin{aligned} \sum y &= na + b\sum x \\ \sum xy &= a\sum x + b\sum x^2 \end{aligned} \right\} \end{aligned}$$

These two equations are known as the normal equations for determining a and b. If we determine the numerical values of a and b such that these equations hold, the least square equation  $Y_c = a + bx$  will satisfy two algebraic properties. First the deviation of observations about the regression line sum to zero, *i.e.*  $\sum e = 0$ ; secondly the sum of squared deviations from the line, *i.e.*  $\sum e^2$  is minimum.

Solving the normal equations for a and b, we have from the first" equation:

$$a = \bar{y} - b\bar{x}$$

and from the second

$$\begin{aligned}\sum xy &= a\sum y + b\sum x^2 \\ &= (\bar{y} - b\bar{x}) \sum x + b\sum x^2 \\ &= \bar{y} \sum x + b\bar{x} \sum x^2 + b\sum x^2\end{aligned}$$

$$\sum xy - n\bar{x}\bar{y} = n - nb + b\sum x^2$$

$$\left[ \begin{array}{l} \because \bar{x} = \frac{\sum x}{n} \\ \because \sum x = n\bar{x} \end{array} \right]$$

$$\sum xy - n\bar{x}\bar{y} = b [\sum x^2 - n\bar{x}^2]$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

The expression for b can be further simplified by putting  $x = x - \bar{x}$  and  $y = y - \bar{y}$  *i.e.* x and y are the deviations from their respective means. Now

$$X = x + \bar{x}, \text{ and } Y = y + \bar{y}$$

$$\sum xy = \sum (x + \bar{x}) (y + \bar{y})$$

$$= \sum xy + \bar{x} \sum y + \bar{y} \sum x + n\bar{x}\bar{y}$$

$$= \sum xy + n\bar{x}\bar{y} \quad (\because \sum x = 0 \text{ and } \sum y = 0)$$

$$= \sum xy = n\bar{x}\bar{y} = \sum xy$$

and similarly  $\sum x^2 = \sum (x + \bar{x})^2$

$$= \sum x^2 + n\bar{x}^2 + 2\bar{x} \sum x$$

$$= \sum x^2 + n\bar{x}^2$$

$$\sum x^2 = \sum x^2 - n\bar{x}^2$$

$$\text{Hence } b = \frac{\sum xy}{\sum x^2}$$

The normal equations can also be solved with the help of Cramer's rule.

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy$$

The solution to these equations is easily written as Follows :

$$\Sigma y \quad \Sigma x$$

$$a = \frac{\Sigma xy - \Sigma x^2}{n \quad \Sigma x} = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n \Sigma x^2 - (\Sigma x)^2}$$

$$\Sigma y \quad \Sigma x^2$$

$$n \quad \Sigma y.$$

$$b = \frac{\Sigma x \quad \Sigma xy}{n \quad \Sigma x} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

Regression equation is a measure of the average relationship between  $x$  and  $y$  such that for a given  $x$ ,  $y_c$  is the value of  $y$  which we would on the average expect to be associated with that  $x$ . The regression coefficient  $b$  measures the change in  $y$  which occurs on average per unit change in  $x$ .  $Y_c$  is of course, expressed in the same units as  $y$ . It will be noted that when  $x = \bar{x}$ ,  $y_c = \bar{y}$  so that the regression line passes through the means of  $X_s$  and  $Y_s$ .

**An Example:** Data on the annual sales of a company in lakhs of rupees over the past eleven years is shown in the table below. Determine a suitable straight line regression model,  $y = a + bx + \varepsilon$  for the data in the table:

Year	Annual sales in Lakhs of Rupees
1988	1
1989	5
1990	4
1991	7
1992	10
1993	8
1994	9
1995	13
1996	14
1997	13
1998	18

**Solution :** The independent variable in this problem is the year whereas the response variable is the annual sales. We see that to estimate the parameter  $b$  we require the four summations  $\Sigma x_i$ ,  $\Sigma y_i$ ,  $\Sigma x_i^2$ , and  $\Sigma x_i y_i$ .

Thus, calculations can be organized as shown below where the totals of four columns yield the four desired summations :

$X_i$	$Y_i$	$X_i^2$	$X_i Y_i$	$Y_i^2$
1	1	1	1	1
2	5	4	10	25
3	4	9	12	16
4	7	16	28	49
5	10	26	50	100
6	8	36	48	64
7	9	49	63	81
8	13	64	104	169
9	14	81	126	196
10	13	100	130	169
11	18	121	198	324
$\Sigma x_i = 66$	$\Sigma y_i = 102$	$\Sigma x_i^2 = 506$	$\Sigma x_i y_i = 770$	$\Sigma y_i^2 = 1194$

we find feat

$$\begin{aligned}
 n &= 11 \\
 \Sigma x_i &= 66 \quad \bar{x} = 6 \\
 \Sigma y_i &= 102 \quad \bar{y} = 9.2727 \\
 \Sigma x_i^2 &= 506 ; \Sigma x_i y_i = 770 \\
 b &= \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} \\
 &= \frac{(11 \times 770) - (66 \times 102)}{(11 \times 506) - (66)^2} \\
 &= \frac{8470 - 6732}{5566 - 4356} = \frac{1738}{1210} = 1.4363 \\
 b &= 1.44 \\
 a &= \bar{Y} - b \bar{X} \\
 &= 9.27 - 1.44 \times 6 \\
 &= 9.27 - 8.64 = 0.63
 \end{aligned}$$

The fitted equation is thus

Regression equation of y on x

$$\begin{aligned}
 y - \bar{y} &= r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \\
 y &= 0.63 + 1.44x
 \end{aligned}$$

Now that the model is completely specified we can obtain the predicted values  $y_i$  and the errors or residuals  $y_i - \hat{y}$  corresponding to the eleven observations. These are shown in table below.

$X_i$	$Y_i$	$\hat{y}_i$	$E = Y_i - \hat{y}_i$	$e_i^2$
1	1	2.07	-1.07	1.15
2	5	3.51	1.49	2.20
3	4	4.95	-0.95	0.90
4	7	6.39	0.61	0.37
5	10	7.83	2.17	4.71
6	8	9.27	-1.27	1.61
7	9	10.71	-1.71	2.92
8	13	12.15	0.85	0.72
9.	14	13.59	0.41	0.16
10	13	15.03	-2.03	4.12
11	18	16.47	1.53	2.34

**The standard Error of Estimate.** It has been shown above feat with the help of the regression equation it is possible for us to estimate the value of y for any given value of x. Thus when x is 8, the sales is estimated to be 12.15 lakhs rupees. Now misestimated value of y is less than its observed value (Rs. 13 lakhs).

This means that the regression line is not a perfect fit and the entire points on the scatter diagram do not follow on this line. In other words, it may be said that regression, line will not enable us to make estimates equal to the observed value, of the sales. It may thus be said that the estimates will be in error. This error is due to the fact that variations in Y may not be due exclusively to variations in X. There are other forces as well which influence the size of y.

In order to know as to how far the regression equation has been able to explain the variation in y, it is necessary to measure the scatter of fee point around the regression line. If all the points on the scatter diagram fall, on the regression line, it means that regression equation anables us to make absolutely correct estimates of the values of y. In other words we can say that the variations in y are fully explained by variations in x and there is no error in the estimates. The scatter of the points from the regression line is called the standard error of estimating y. It is obtained commonly by the formula:

$$S_y = \sqrt{\frac{\Sigma(y - y_c)^2}{N - 2}} = \sqrt{\frac{\Sigma e^2}{n - k}}$$

Where  $S_y$  standard error of estimate  
 $y$  = the observed values of y  
 $y_c$  = the estimated values of y  
 $N - 2$  = degrees of freedom

(Since two parameters  $a$  and  $b$  have been estimated from  $n$  i.e. 11 observations)

$$S_y = \sqrt{\frac{21.25}{9}} = 1.53$$

It will be observed that the, method of computing  $S_y$  is similar, to that of calculating  $\sigma$  or with, the only difference that whereas in calculating  $\sigma$  deviations are measures from the mean, in the case of  $S_y$  the y are measured from the regression, line, (the estimated y) for the purpose of computing the value of standard estimate of y, the formula can further be simplified so that we, avoid the need, for calculating individual  $e_i$ 's the deviations of the observed values from the regression line.

$$e_i = (y_i - \hat{a} - \hat{b}x_i)$$

$$\Sigma e_i^2 = \Sigma (y_i - \hat{a} - \hat{b}x_i)^2$$

$$\therefore \hat{y} = \bar{y} - \hat{b} \bar{X}$$

By substituting for  $\alpha$  we obtain

$$= \Sigma (y_i - (\bar{y} - \hat{b}\bar{x}) - \hat{b}x_i)^2$$

$$= \Sigma [(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})]^2$$

$$= \Sigma [(y_i - \bar{y})^2 + \hat{b}^2 (x_i - \bar{x})^2 - 2\hat{b}(y_i - \bar{y})(x_i - \bar{x})]$$

$$= \Sigma [(y_i - \bar{y})^2 + \hat{b}^2 (x_i - \bar{x})^2 - 2\hat{b}(x_i - \bar{x})^2]$$

$$\text{Since } \hat{b} = \frac{(x_1 - \bar{x})(y_1 - \bar{y})}{(x_1 - \bar{x})^2}$$

$$= \Sigma [(y_i - \bar{y})^2 + \hat{b}^2 (x_i - \bar{x})^2]$$

$$= \Sigma y_i^2 \frac{(\Sigma y)^2}{N} - \hat{b}^2 \Sigma x_i^2 + \hat{b}^2 \frac{(\Sigma x)^2}{N}$$

$$= \Sigma y_i^2 \frac{(\Sigma y)^2}{N} - \hat{b}^2 \left[ \Sigma x_i^2 \frac{(\Sigma x)^2}{N} \right]$$

$$S_y = \frac{\sqrt{\Sigma y_i^2 \frac{(\Sigma y)^2}{N} - \hat{b}^2 \left[ \Sigma x_i^2 \frac{(\Sigma x)^2}{N} \right]}}{n - 2}$$

$$S_y = \frac{\sqrt{1194 \frac{(102)^2}{11} - (1.4363)^2 \left( 506 - \frac{(66)^2}{11} \right)}}{9}$$

$$= \sqrt{\frac{21.25}{9}} = 1.53$$

which is much easier, to compute that taking value of  $e_i$  by subtracting estimated;  $y^s$  from observed  $y_s$  and then squaring and summing.

The various computations outlined in the case of straight line regression equation will now be illustrated to compute the standard error of intercept and slope. Most commonly the intercept and slope. Most, commonly the estimators of  $\text{var}(\hat{a})$  and  $\text{var}(\hat{b})$ , are denoted by  $S^2\hat{a}$  and  $S^2\hat{b}$  respectively. The formulas for calculating  $\text{var}(\hat{a})$  and  $\text{var}(\hat{b})$  are :

$$S^2\hat{a} = \left[ \frac{1}{n} + \frac{x^2}{\left[ \sum x_i^2 - \frac{(\sum x_i)^2}{N} \right]} \right] \sigma^2 = \frac{\sum x_i^2 / n}{\sum (x_i - \bar{x})^2} \sigma^2$$

$$S^2\hat{b} = \frac{\sigma^2}{\left[ \sum x_i^2 - \frac{(\sum x_i)^2}{N} \right]} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\text{Where } \sigma^2 = \frac{\sum e_i^2}{n-2}$$

Standard Error of the intercept  $\hat{a}$  .

$$S^2\hat{a} = \frac{506}{110} \sigma^2 = \frac{46}{110} \sigma^2 = .418 \times 2.36$$

$$= 0.98$$

$$\text{estimate of standard error } (\hat{a}) = \sqrt{\text{esv}(\hat{a})}$$

$$= 0.99$$

Then 95% confidence limit is for  $\hat{a}$  are

$$a_0 \pm \frac{t(9, 0.975) s[\sum e_i^2]^{1/2}}{[n \sum (x_i - \bar{x})^2]^{1/2}}$$

$$= 0.63 \pm (2.262) (0.98)$$

$$= 0.63 \pm 2.2167, \text{ that is } -1.5867 \text{ and } 2.8467$$

Standard error of the slopes  $\hat{b}$  .

$$S^2(\hat{b}) = \frac{\sigma^2}{\sum x_1^2 - \frac{(\sum x)^2}{\wedge 1}} = \frac{\sigma^2}{\sum x_1^2}$$

$$= \frac{2.36}{110} = 0.0215$$

$$\text{estimate of standard error } (\hat{b}) = \sqrt{\text{est}(S^2\hat{b})}$$

$$= 0.1465.$$

Suppose  $\alpha = 0.05$ , so that  $\left(n - 2, 1 - \frac{\alpha}{2}\right)$   
 $= (9, 0.975) = 2.262$  from the table of the distribution.

Then 95% confidence limits for  $\hat{b}$  are

$$b \pm \frac{+ (9, 0.975) S}{\left[ \sum x_1^2 - \frac{(\sum x_i)^2}{N} \right]^{1/2}}$$

$$= 1.44 \pm (2.262) (0.1465)$$

$$= 1.44 \pm 0.3314 \text{ that is } 1.7714 \text{ and } 1.1086$$

### **SUGGESTED READINGS**

1. Draper, N.R. and N. Smith, Applied Regression Analysis, John Wiley : New York, 1966.

\*\*\*\*\*



## LESSON-5 & 6

### THE GENERAL LINEAR REGRESSION MODEL : MATRIX FORMULATION & SOLUTION

**Dear Student,**

In the simple linear regression model, the objectives of the analysis were to determine the degree of relationship between two variables and to predict the behaviour of the dependent variable on the basis of an independent variable. It is generally seen that the dependent variable is 'related' not only to one independent variable but to a number of independent variables all operating at the same time. For example the quantitative demanded for a given, commodity (y) depends on its price ( $x_1$ ) and on consumer's income ( $X_2$ ) and price of related goods ( $X_3$ ). In the present lesson we shall extend the simple linear regression model to relationships with two explanatory variables and later on we shall develop some practical rules for the derivation of the normal equations for models including any number of variables. In section I we shall examine the model with two explanatory variables.

**Relations between three variables:** We shall illustrate the three variable model with an example from the theory of demand. It is well known that quantity demanded for a given commodity is a function of its price and consumers income  $x_1$  and  $x_2$  respectively.

$$y = f(x_1, x_2)$$

We assume that there is a linear relationship, between y,  $x_1$  and  $x_2$ .

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad (i = 1, 2, 3, \dots, n)$$

The, above relationship, is an exact relationship showing that the n variations, in the quantity demanded are fully explained by changes in price, income and income. However, if the gathered information is plotted on a diagram of it. It will be observed that, some will lie on it, but others will lie above or (0) below it and all of them will of it lie on a plane. This scatter is due to error (s) account by introducing a random variable  $\mu$ , in the function, which thus becomes stochastic.

A sample on n households would give related observations.

$$Y_i = B_0 + B_1 x_{1i} + B_2 x_{2i} + \mu_i$$

**systematic component                      random component**

On a priori grounds, the coefficient  $\hat{\beta}_1$ , have a negative sign, showing an inverse relationship between quantity, demanded and price, while  $\hat{\beta}_2$  is expected to have, a positive sign as quantity demanded and income are positively correlated.

Where  $x_{1i} = 1$  for all i; that is y can be regarded as a linear function of the X's, the sample values of the first x variable always being a set of units. Our basic objective is to give the student a firm grasp and understanding, of the basic concepts required in the analysis of a relationship between three variables, so that extension to the general case is facilitated.

Let the regression equation of y on  $x_1$  and  $x_2$  is of the form

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$Y_i = \hat{B}_0 + \hat{B}_1 x_{1i} + \hat{B}_2 x_{2i}$$

Where  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  are estimates of the true parameters  $\beta_0, \beta_1$  and  $\beta_2$  of the demand relationship. As before, the estimates will be obtained by minimizing the sum of squared residuals.

$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$$

A necessary, condition for this expression to assume a minimum ‘value’ is that its partial derivatives with respect to  $\hat{\beta}_0, \hat{\beta}_1$  and  $\hat{\beta}_2$  be equal to zero :

$$\frac{\partial \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{\partial \hat{\beta}_1} = 0$$

$$\frac{\partial \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{\partial \hat{\beta}_1} = 0$$

$$\frac{\partial \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{\partial \hat{\beta}_2} = 0$$

Performing the partial differentiations we get the following system of three normal equations, in the three unknown parameters  $\hat{\beta}_0, \hat{\beta}_1$  and  $\hat{\beta}_2$ .

$$\begin{aligned} \sum y_i &= n \hat{\beta}_0 + \hat{\beta}_1 \sum x_{1i} + \hat{\beta}_2 \sum x_{2i} \\ \sum x_{1i} y_i &= \hat{\beta}_0 \sum x_{1i} + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \\ \sum x_{2i} y_i &= \hat{\beta}_0 \sum x_{2i} + \hat{\beta}_1 \sum x_{1i} x_{2i} + \hat{\beta}_2 \sum x_{2i}^2 \end{aligned} \quad \dots (5.2)$$

This system can be solved by using, determinants or by the use of normal equations. The above normal equations are expressed in terms of deviations from means. In this case the first equation disappears as  $\sum x_{1i}, \sum x_{2i}$  and  $\sum y_i$  are not equal to zero (because the sum of deviation from the inspected mean is equal to zero):  $\sum (x_{1i} - \bar{x}) = \sum x_{1i} = 0$ , the remaining two equations are reduced to the form:

$$\begin{cases} \sum x_{1i} y_i = \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \\ \sum x_{2i} y_i = \hat{\beta}_1 \sum x_{1i} x_{2i} + \hat{\beta}_2 \sum x_{2i}^2 \end{cases} \quad \dots 5.3$$

The coefficient of correlation between y and  $x_1$  from mean is given by

$$r_{yx1} = \frac{\sum y x_1}{N \sigma_y \sigma_x}$$

or  $N \sigma_y \sigma_{x_1} r_{yx_1} = \sum y x_1$

Similarly  $N \sigma_y \sigma_{x_2} r_{yx_2} = \sum y x_2$

and  $N \sigma_{x_1} \sigma_{x_2} r_{x_1 x_2} = \sum x_1 x_2$

The standard deviation (from mean) of  $x_1$  series

$$\sigma_{x_1} = \sqrt{\frac{\sum x_1^2}{N}} \text{ or } N\sigma_{x_1}^2 = \sum x_1^2$$

$$\text{and } N\sigma_{x_2}^2 = \sum x_2^2$$

Substituting the computed values

$$N\sigma_y \sigma_{x_1} r_{yx1} = \hat{\beta}_1 N\sigma_{x_1}^2 + \hat{\beta}_2 N\sigma_{x_1} \sigma_{x_2} r_{x1x2} \quad \text{..... 5.4}$$

$$N\sigma_y \sigma_{x_2} r_{yx2} = \hat{\beta}_1 N\sigma_{x_1} \sigma_{x_2} r_{x1x2} + \hat{\beta}_2 N\sigma_{x_2}^2 \quad \text{..... 5.5}$$

Cancelling  $N\sigma_{x_1}$  from equation (5.4) and  $N\sigma_{x_2}$  from equation (5.5) we get.

$$\sigma_y r_{yx1} = \hat{\beta}_1 \sigma_{x_1} + \hat{\beta}_2 \sigma_{x_2} r_{x1x2} \quad \text{..... 5.6}$$

$$\sigma_y r_{yx2} = \hat{\beta}_1 \sigma_{x_1} r_{x1x2} + \hat{\beta}_2 \sigma_{x_2} \quad \text{..... 5.7}$$

Dividing equation (5.7) by  $r_{x1x2}$  and subtracting it from 5.6, we get

$$\left( \sigma_y r_{yx1} - \sigma_y \frac{r_{yx2}}{r_{x1x2}} \right) = \hat{\beta}_2 \left( \sigma_{x_2} r_{x1x2} \frac{\sigma_{x_2}}{r_{x1x2}} \right)$$

$$\sigma_y \left( r_{yx1} - \frac{r_{yx2}}{r_{x1x2}} \right) = \hat{\beta}_2 \left( \sigma_{x_2} r_{x1x2} \frac{\sigma_{x_2}}{r_{x1x2}} \right)$$

$$\sigma_y \left( r_{yx1} - \frac{r_{yx2}}{r_{x1x2}} \right) = \hat{\beta}_2 \sigma_{x_2} \left( r_{x1x2} - \frac{1}{r_{x1x2}} \right)$$

$$\hat{\beta}_2 = \frac{\sigma_y}{\sigma_{x_2}} \left( \frac{\frac{r_{yx1} r_{x1x2} - r_{yx2}}{r_{x1x2}}}{\frac{r_{x1x2}^2 - 1}{r_{x1x2}}} \right)$$

$$= \frac{\sigma_y}{\sigma_{x_2}} \left( \frac{r_{yx1} r_{x1x2} - r_{yx2}}{\sigma_{x1x2}} \times \frac{r_{x1x2}}{r_{x1x2}^2 - 1} \right)$$

$$= \frac{\sigma_y}{\sigma_{x_2}} \left( \frac{r_{yx1} - r_{yx1} r_{x1x2}}{1 - r_{x1x2}^2} \right)$$

$$\hat{\beta}_1 = \frac{\sigma_y}{\sigma_{x_1}} \left( \frac{r_{yx1} - r_{yx2} r_{x1x2}}{1 - r_{x1x2}^2} \right)$$

or by solving 1.4 and 1.5 algebraically we get

$$\hat{\beta}_1 = \frac{\sum yx_1 \sum x_2^2 - \sum yx_2 \sum x_1 x_2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

Which may be reduced to the following expression in forms of zero coefficient.

$$\hat{\beta}_1 = \left( \frac{r_{yx1} - r_{yx2} r_{x1x2}}{1 - r_{x1}^2} \right) \frac{\sigma_y}{\sigma_{x1}}$$

$$\hat{\beta}_2 = \left( \frac{r_{yx2} - r_{yx1} r_{x1x2}}{1 - r_{x1}^2} \right) \frac{\sigma_y}{\sigma_{x2}}$$

### Elements of Matrix Algebra

It is excessively tedious and complicated to build up a general case of K variables in a stepwise fashion. Fortunately, by the use of matrix algebra compact and powerful way of the treating, the general case be obtained by the use of matrix algebra. A matrix may be defined as on system of m n numbers arranged in the form of an ordered set of m row s, each row consisting of an ordered set of n numbers (or m, n numbers arranged in the form m rows and n columns). This matrix is called m×n matrix (first, letter of  $m \times n$  always denotes rows and second denotes columns).

or A matrix is defined as a rectangular array of elements arranged in rows or column as in

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2n} \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{in} \\ a_{m1} & a_{m2} & \dots & a_{mj} & \dots & a_{mn} \end{bmatrix} \quad m \times n$$

If it has  $mn$  elements arranged in  $m$  rows, and  $n$  columns, it is said to be of order  $m$  by  $n$ , which is often written as  $m \times n$ .

The elements in the  $i$ th row and  $j$ th column is represented by  $a_{ij}$ . The matrix may be indicated more concisely by

$$A = [a_{ij}]$$

A matric of order I'n contains only a single row of elements and is commonly referred as a row vector, for example,

$$b = [b_1 b_2 \dots b_n]$$

While a number of order  $mx_1$  is a column vector,

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}$$

To economize space, comlumn vectors, may be written in horizontal position but enclosed in braces,

$$c = [c_1 c_2 \dots c_m]$$

### Set of Rules and Definitions

- (1) Two matrices A and B are said to be equal when they are of the same order and  $a_{ij} = b_{ij}$  for all  $i, j$ : that is matrices are equal element by element.

- (2) If A and B are of the same order, then we define  $A + B$  to be a new matrix  $C$  of the same order in which

$$c_{ij} = a_{ij} + b_{ij} \text{ for all } i, j.$$

$$A = \begin{bmatrix} 2 & 3 \\ -4 & 5 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -3 & 1 \\ 4 & 1 \end{bmatrix}$$

then

$$C = A + B = \begin{bmatrix} -1 & 4 \\ 0 & 6 \end{bmatrix}$$

- (3) If  $r$  is a scalar, then we define scalar multiplication such that  $rA = [r a_{ij}]$  that is each element of  $A$  is multiplied by  $r$ . For example if

$$A = \begin{bmatrix} 2 & 3 \\ -4 & 5 \end{bmatrix} \quad \text{and} \quad r = -2$$

then

$$rA = \begin{bmatrix} -4 & -6 \\ +8 & -10 \end{bmatrix}$$

It follows from the rules for addition and scalar multiplication that  $A - B = [a_{ij} - b_{ij}]$

**Multiplication of Matrices:** If  $A = [a_{ij}]$  and  $B = [b_{ij}]$  are two matrices, then their product  $AB$  is defined to be a matrix of order  $m \times p$  (if  $A$  is of the order  $m \times n$  and  $B$  is of the order  $n \times p$ ). Whose  $ij$ th element is

$$C_{ij} = \sum_{k=1}^n a_{ik} b_{kj} \text{ then}$$

that is, the  $ij$ th element in the product matrix is by multiplying the elements of the  $i$ th row of the first matrix by the corresponding elements of the  $j$ th column of the second matrix and summing over  $n$  terms.

**Example**

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} 2 \times 3 \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} 3 \times 2$$

Then  $AB$  is  $2 \times 2$  matrix

$$AB = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}$$

While  $BA$  is a  $3 \times 3$  matrix

$$BA = \begin{bmatrix} b_{11}a_{11} + b_{12}a_{21} & b_{11}a_{12} + b_{12}a_{22} & b_{11}a_{13} + b_{12}a_{23} \\ b_{21}a_{11} + b_{22}a_{21} & b_{21}a_{12} + b_{22}a_{22} & b_{21}a_{13} + b_{22}a_{23} \\ b_{31}a_{11} + b_{32}a_{21} & b_{31}a_{12} + b_{32}a_{22} & b_{31}a_{13} + b_{32}a_{23} \end{bmatrix}$$

I. It must be noticed that the commutative law of addition holds. A and B must, of course, be of the same order, and the result follows directly from the definition of the addition of matrices.

$$\begin{bmatrix} 2 & 3 \\ -5 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ -2 & 6 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -2 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 3 \\ -5 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ -7 & 6 \end{bmatrix}$$

II.  $AB \neq BA$  except for rather special square Matrices : *i.e.*, the commutative law of multiplication does not, in general hold. If the matrices are of order  $m \times n$  and  $n \times m$ , then both products will exist, but they will be of different order and hence cannot be equal. If both are square matrices of the same order then both products will exist and will be of the same order but not necessarily equal as the following examples show.

#### Examples

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} B = \begin{bmatrix} 3 & 0 \\ -2 & 2 \end{bmatrix}$$

$$AB = \begin{bmatrix} 6+1 & 0+2 \\ 3+1 & 0+2 \end{bmatrix} = \begin{bmatrix} 7 & 2 \\ 4 & 2 \end{bmatrix}$$

$$BA = \begin{bmatrix} 6+0 & 3+0 \\ 2+2 & 1+2 \end{bmatrix} = \begin{bmatrix} 6 & 3 \\ 4 & 3 \end{bmatrix}$$

Whereas if  $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$   $B = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$

$$AB = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = BA.$$

III.  $(A+B) + C = A + (B+C)$  : that is the associative law of addition holds. Since addition of matrices is simply achieved by the addition of corresponding elements and since it does not matter in which order elements are added together, the associative law holds.

IV. Matrix Multiplication is associative, *i.e.*  $AB(C) = A(BC)$ .

We can first form  $AB$  and then post-multiply by  $C$ . or multiply  $A$  and  $BC$ , of course  $BC$  will have to be found first *e.g.* if  $A$  is  $m \times n$  and if  $C$  is  $p \times q$  then conformability requires that. The distributive law holds for matrix multiplication *i.e.*

$$A(B+C) = AB + AC \text{ (pre-multiply by } A)$$

$$(B+C)A = BA + CA \text{ (post-multiply by } A)$$

VI.  $\lambda(A+B) = \lambda A + \lambda B$  and  $(\lambda + \mu)A = \lambda A + \mu A$ : that is the distributive law of scalar multiplication.

These are the important results for the manipulation of matrices and the students should learn them by working out numerical examples.

A unit of Identity matrix is “defined by

$$I_n = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

a scalar matrix has a common scalar element in the principal diagonal and  $\mu$  nos everywhere, else and is defined by

$$\lambda I = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

A diagonal matrix is defined in which scalar elements are not necessarily equal in the principal, diagonal and zeros in off diagonal positions, i.e.,

$$A = [a_{ij}] \quad i, j = 1, 2, \dots, n$$

$$a_{ij} = 0, \quad i \neq j.$$

That is

$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ 0 & a_{22} & \dots & 0 & 0 \\ 0 & 0 & \dots & a_{nn} & \dots \end{bmatrix} \quad \text{diag } [a_{11} \ a_{22} \ \dots \ a_{nn}]$$

A scalar matrix is thus a special form of a diagonal matrix.

Transposition, the transpose of A is defined to be the matrix obtained from A by inter changing, rows and columns that is, the first row of A becomes the first column of transpose, the second row of A becomes the second column of the transpose, and in general" the  $ij$ th element in the transpose is the  $ji$ th element of the original matrix.

$$A = [a_{ij}] \quad m \times n \text{ or } A' = [a_{ji}] \quad n \times m$$

For example if

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}_{2 \times 3} \quad \text{Then } A' = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}_{3 \times 2}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}_{2 \times 3} \quad \text{Then } A' = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}_{3 \times 2}$$

$$A' A = \begin{bmatrix} a_{11} \times a_{11} - a_{12} \times a_{12} + a_{13} \times a_{13} & a_{11} \times a_{21} - a_{12} \times a_{22} + a_{13} \times a_{23} \\ a_{21} \times a_{11} - a_{22} \times a_{12} + a_{23} \times a_{13} & a_{21} \times a_{21} - a_{22} \times a_{22} + a_{23} \times a_{23} \end{bmatrix}$$

and  $A' A$

$$= \begin{bmatrix} a_{11}^2 & a_{21}^2 & a_{11} a_{12} + a_{21} a_{22} & a_{11} a_{13} + a_{21} a_{23} \\ a_{12} a_{11} + a_{22} a_{21} & a_{12}^2 & a_{22}^2 & a_{12} a_{11} + a_{22} a_{21} \\ a_{13} a_{11} + a_{23} a_{21} & a_{13} a_{12} + a_{23} a_{22} & a_{13}^2 & a_{23}^2 \end{bmatrix}_{3 \times 3}$$

If  $x$  is a column vector of  $n$  elements  $x_i$  is then a row vector of  $m$  elements and

$$x' x = \sum_{i=1}^n x_i^2$$

and

$$XX' = \begin{bmatrix} x_1^2 & x_1 x_2 & \dots & x_1 x_n \\ x_2 x_1 & x_2^2 & \dots & x_2 x_n \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \\ 1 & & & 1 \\ x_n x_1 & x_n x_2 & & x_n^2 \end{bmatrix}$$

Some Important properties of transpose of matrix.

- (1)  $(a')' = A$ .
- (2)  $(A^k)' = (A')$  where  $k$  is a positive integer.
- (3)  $(A + B)' = A' + B'$
- (4)  $(AB)' = B' A'$
- (5)  $(ABC)' = C' B' A'$

**Example:** If  $A = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix}$  and  $B = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 1 \end{bmatrix}$

Prove that (i)  $(A+B)' = A' + B'$

(ii)  $(AB)' = B' A'$

$$(1) A + B = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 0 & 7 & 6 \\ 6 & 2 & 1 \end{bmatrix}$$

$$\therefore (A + B)' = \begin{bmatrix} 4 & 0 & 6 \\ 2 & 7 & 2 \\ 1 & 6 & 1 \end{bmatrix} \dots (1)$$



$$\text{Also } A' = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix} \text{ and } B' = \begin{bmatrix} 1 & 0 & 5 \\ 2 & 3 & 0 \\ 0 & 4 & 1 \end{bmatrix}$$

$$\therefore A' + B' = \begin{bmatrix} 4 & 0 & 6 \\ 2 & 7 & 2 \\ 1 & 6 & 1 \end{bmatrix} \quad \text{..... (2)}$$

from (1) and (2), we get

$$(A+B)' = A' + B'$$

$$\begin{aligned} (AB) &= \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 \times 1 + 0 \times 0 + 1 \times 5 & 3 \times 2 + 0 \times 3 + 1 \times 0 & 3 \times 0 + 0 \times 4 + 1 \times 1 \\ 0 \times 1 + 4 \times 0 + 2 \times 5 & 0 \times 2 + 4 \times 3 + 2 \times 0 & 0 \times 0 + 4 \times 4 + 2 \times 1 \\ 1 \times 1 + 2 \times 0 + 0 \times 5 & 1 \times 2 + 2 \times 3 + 0 \times 0 & 1 \times 0 + 2 \times 4 + 0 \times 1 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 6 & 1 \\ 10 & 12 & 18 \\ 1 & 8 & 8 \end{bmatrix} \end{aligned}$$

$$\therefore (AB)' = \begin{bmatrix} 8 & 10 & 1 \\ 6 & 12 & 8 \\ 1 & 18 & 8 \end{bmatrix} \quad \text{..... (I)}$$

$$\begin{aligned} B'A' &= \begin{bmatrix} 1 & 0 & 5 \\ 2 & 3 & 0 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 10 & 1 \\ 6 & 12 & 8 \\ 1 & 18 & 1 \end{bmatrix} \quad \text{..... (II)} \end{aligned}$$

From (1) and (2), we get

$$(AB)' = B'A'$$

### The General Linear Regression Model

The discussion so far was limited to the regression models containing one or two independent/explanatory variables. Now we shall generalize the model assuming that it contains K explanatory variables. The general linear equation will be of the form.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots \beta_{ik} X_{ki} + U_i (i = 1 \dots n) \quad \text{..... (2.1)}$$

We need some assumption about the error term  $u$ .

**Assumption I :**  $\mu_i$  is a random real variable and has  $\sim$  normal distribution.

**Assumption II :** The mean value of  $\mu$  for each  $x_i$  is zero or  $E(\mu_i) = 0$ .

**Assumption III :** The variance of the disturbance term is constant in each period :

$$E(\mu_i^2) = \sigma\mu^2 (\sigma\mu^2 \text{ is a constant})$$

**Assumption on IV :** The covariance of  $\mu_i$  and  $\mu_j$  is equal to zero.

$$E(\mu_i \mu_j) = 0 \text{ for } i \neq j.$$

**Assumption V:** Every disturbance term, is independent of the explanatory variables:

$$E(x_{2i} u_i) = X_2 E(u_i) = 0$$

$$E(x_{3i} u_i) = X_3 E(u_i) = 0$$

There are  $k$  parameters to be estimated ( $k = k + 1$ ) clearly the system of normal equations will consist of  $k$  equation  $s$ , in which  $B_0, B_1, B_2 \dots$  and  $B_k$  are the unknown parameters, and the known term  $s$  will be the sum of squares and the sum of the products of all the variables in the structural equation.

### 1. Model with one explanatory variable

$$\text{Structural form } Y = \beta_0 + \beta_1 X_1 + \mu$$

$$\text{estimated from } Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + e$$

$$\text{Normal equations} \quad \left\{ \begin{array}{l} \Sigma Y = n \hat{\beta}_0 + \hat{\beta}_1 \Sigma X \\ \Sigma X_1 Y = \hat{\beta}_0 \Sigma X_1 + \hat{\beta}_1 \Sigma X_1^2 \end{array} \right\}$$

### 2. Model with two explanatory variables

$$\text{structural form } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

$$\text{estimated form } y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$$

$$\text{Normal equations } \Sigma y = n \hat{\beta}_0 + \hat{\beta}_1 \Sigma X_1 + \hat{\beta}_2 \Sigma X_2$$

$$\Sigma Y X_2 + \hat{\beta}_0 \Sigma X_2 + \hat{\beta}_1 \Sigma X_1 X_2 + \hat{\beta}_2 \Sigma X_2^2$$

$$X_2 = \hat{\beta}_0 \Sigma X_2 + \hat{\beta}_1 \Sigma X_1 X_2 + \hat{\beta}_2 \Sigma X_2^2$$

### 3. Model with k explanatory variable structural form

$$Y_2 = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots \beta_k X_{ki} + \mu_i$$

$$(i = 1.2 \dots \dots \dots n)$$

Since Subscript represents the  $i$ th observation, we shall have  $n$  number of equations with  $n$  number of observations on each variable :

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \beta_3 X_{31} + \dots B_k X_{k1} + \mu_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \beta_3 X_{32} + \dots \beta_k X_{k2} + \mu_2$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \dots \beta_k X_{kn} + \mu_n$$

These equations are put in matrix form  $y = \Sigma b + u$

where:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{k1} \\ X_{12} & X_{22} & \dots & X_{k2} \\ \cdot & & & \\ \cdot & & & \\ X_{1n} & X_{2n} & X_{kn} & X_{kn} \end{bmatrix}$$

$n \times (k+1)$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} \quad \text{and } U = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ u_n \end{bmatrix}$$

Taking (1.1) and assumptions together, we now apply the least square principle to estimate the parameters of 1.1.

$$\hat{\beta} = \{\hat{\beta}_0 \hat{\beta}_1 \dots \hat{\beta}_k\}$$

denote a column vector of estimate of  $\beta$ . Then we write

$$y = X\hat{\beta} + e \quad \dots \dots \dots 2.2$$

Where denotes the column vector of  $n$  residuals  $(y - X\hat{\beta})$ . You should carefully notice the basic difference between (2.1) and (2.2). In the former the unknown coefficients  $\beta$  and the unknown disturbances  $u$  appear, while in the latter, we have some set of estimates  $\hat{\beta}$  and the corresponding set of residual  $e$ . From 1.2 the sum of squared residual, is

$$\sum_{i=1}^n e_i^2 = e'e$$

We have to minimize :  $\sum_{i=1}^n e_i^2$

$$\begin{aligned} \Sigma e_i^2 &= \Sigma e_i^2 = e_1^2 + e_2^2 + e_3^2 \dots e_n^2 \\ &= [e_1 \ e_2 \ e_3 \dots e_n] \end{aligned}$$

IX

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_{n \times 1} \end{bmatrix} \quad \text{..... 2.3}$$

$$\begin{aligned} \Sigma e_1^2 &= e'e \\ &= (y - X\hat{\beta})(y - X\hat{\beta}) \\ &= (y - X\hat{\beta}')(y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y' - \hat{\beta}'Y'X' + X'X\hat{\beta}' \\ &= Y'Y - 2\hat{\beta}'X'Y' - \hat{\beta}'X'X'\hat{\beta} \end{aligned} \quad \text{..... 2.4}$$

Which follows from noting that  $\hat{\beta}'X'Y$  is a scalar and thus equal to its transpose  $Y'X'\hat{\beta}$ .

To find the value of  $\beta$  which minimizes the sum of squared residuals we differentiate (2-4)

$$\frac{\partial}{\partial \hat{\beta}} (e'e) = 2X'Y + 2XX'\hat{\beta}$$

equating to zero gives

$$\begin{aligned} XX'\hat{\beta} &= 2X'Y \\ \hat{\beta} &= (X'X)^{-1} X'Y \end{aligned} \quad \text{..... 2.5}$$

This is the fundamental result for the least squares estimators. As an illustration of this result consider the two variable case. Here

$$X'X = \begin{bmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{bmatrix} \text{ and } X'Y = \begin{bmatrix} \Sigma X \\ \Sigma XY \end{bmatrix}$$

So that writing (2.5) in the attentive form

$$(X'X)\hat{\beta} = X'Y$$

and substituting gives.

$$\begin{aligned} \Sigma Y &= a\hat{\beta}_0\hat{\beta}_1\Sigma X \\ \Sigma XY &= \hat{\beta}_0\Sigma X + \hat{\beta}_1\Sigma X^2 \end{aligned}$$

which are two normal equations already derived. For the three variable case (2.5) gives

$$\Sigma Y = b\hat{\beta}_0 + \hat{\beta}_1\Sigma X_2 + \hat{\beta}_2\Sigma X_3 \quad \text{.... 2.6}$$

$$\Sigma X_2Y = \hat{\beta}_0\Sigma X_2 + \hat{\beta}_1\Sigma X_2^2 + \hat{\beta}_2\Sigma X_2X_3 \quad \text{.... 2.7}$$

$$\Sigma X_3Y = \hat{\beta}_0\Sigma X_3 + \hat{\beta}_1\Sigma X_2X_3 + \hat{\beta}_2\Sigma X_3^2 \quad \text{..... 2.8}$$

and it is clear from the symmetry how these equations can be built up for higher order cases.

Solving (2.7) and (2.8) for  $\hat{\beta}_1$  and  $\hat{\beta}_2$ .

$$\hat{\beta}_1 = \frac{\Sigma x_1 y \Sigma x_2^2 - \Sigma x_1 x_2 \Sigma x_2 y}{\Sigma x_2^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} \quad \dots 2.9$$

$$\hat{\beta}_2 = \frac{\Sigma x_2 y \Sigma x_3^2 - \Sigma x_2 x_3 \Sigma x_3 y}{\Sigma x_3^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} \quad \dots 2.10$$

### Goodness of fit ( $R^2$ )

$$R^2 = 1 - \frac{\Sigma ei^2}{\Sigma yi^2} \quad (\text{In the case of one explanatory variable}) \quad \dots (2.11)$$

In the present model of two explanatory variables

$$\begin{aligned} \Sigma e_i^2 &= \Sigma (y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2 \\ &= \Sigma ei (y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}) \\ &= \Sigma eiy_i - \hat{\beta}_1 \Sigma e_i x_{1i} - \hat{\beta}_2 \Sigma e_i x_{2i} \\ &= \Sigma e_i y_i \quad [\because \Sigma e_i x_{1i} = \Sigma e_i x_{2i} = 0] \\ &= \Sigma y_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) \\ \Sigma e_i^2 &= \Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i \\ \Sigma y_i^2 &= \hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i + \Sigma e_i^2 \\ \Sigma y_i^2 &= \hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i + \Sigma e_i^2 \end{aligned}$$

Total sum of squares or total	Explain sum of squares (explained Variations)	Residual sum of square (unexplained Variation)
----------------------------------	--	---

Dividing both sides by  $\Sigma y_i^2$

$$\begin{aligned} \therefore &= \frac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i + \Sigma e_i^2}{\Sigma y_i^2} + \frac{\Sigma e_i^2}{\Sigma y_i^2} \\ &\left( = \frac{\text{Sum of Squares explained by } X_1 \text{ and } X_2}{\text{Total sum of squares}} \right) \end{aligned}$$

For estimation of the standard errors of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  we need an estimate of  $\sigma u^2$ .

$$R^2 = 1 - \frac{\Sigma ei^2}{\Sigma yi^2} \quad \text{or} \quad \Sigma ei^2 = \Sigma yi^2 (1 - R^2)$$

$$\text{Var} (\hat{\beta}_1) = \frac{\sigma u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i} \Sigma x_{2i})^2} \quad \dots (2.14)$$

$$\text{Var } (\hat{\beta}_2) = \frac{\sigma^2 \sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i} \sum x_{2i})^2} \cdot \quad \dots (2.15)$$

These variations can be expressed in terms, of simple correlation coefficients.

$$\text{Var } (\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2 \frac{(\sum x_{1i} \sum x_{2i})^2}{\sum x_{2i}^2}} = \frac{\sigma^2}{\sum x_{1i}^2 \left[ \frac{(\sum x_{1i} \sum x_{2i})^2}{\sum x_{2i}^2 - \sum x_{2i}^2} \right]}$$

$$\text{i.e var } \hat{\beta}_1 = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{12}^2)} \quad \dots (2.16)$$

$$\text{Similarly var } \hat{\beta}_2 = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{12}^2)} \quad \dots (2.17)$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - k} \quad (\text{k being the} \quad \dots (2.18)$$

total number of parameters to be estimated. In the two explanatory variables model.

$$k = 3 \text{ and } \hat{\sigma}^2 = \frac{\sum e_i^2}{n - 3}$$

**Example: 3.1** Applications using matrix algebra with one dependent and two independent variables.

Y	:	49	40	41	46	52	59	53	61	55	60
X <sub>1</sub>	:	35	35	38	40	40	42	44	46	50	50
X <sub>2</sub>	:	53	35	50	64	70	68	59	73	59	71

**Worsheet for the model :  $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + 4$**

y	x <sub>1</sub>	x <sub>2</sub>	y <sub>1</sub>	x	x	x <sub>2</sub>	y <sub>1</sub> <sup>2</sup>	x <sub>1</sub> <sup>2</sup>	x <sub>2</sub> <sup>2</sup>	x <sub>1</sub> y <sub>1</sub>	x <sub>2</sub> y <sub>2</sub>	x <sub>1</sub> x <sub>2</sub>
			(y - $\bar{y}$ )		(x - $\bar{x}$ )		(x - $\bar{x}_2$ )					
n												
1	49	35	53	-3	-7	-9	9	49	81	21	27	63
2	40	35	53	-12	-7	-9	144	49	81	84	108	63
3	41	38	50	-11	-4	-12	121	16	144	44	132	48
4	46	40	64	-6	-2	2	36	4	4	12	-12	-4
5	52	40	70	0	-2	8	0	4	68	0	0	-16
6	59	42	59	7	0	6	49	0	36	0	42	0
7	53	44	68	1	2	-3	1	4	9	2	-3	-6
8	61	46	73	9	4	11	81	16	121	36	99	44

9	55	50	59	3	8	-3	9	64	9	24	-9	-24
10	60	50	71	12	8	9	144	64	81	96	108	72
n = 10	$\Sigma y$	$\Sigma x_1$	$\Sigma x_2$	$\Sigma y_i = 0$	$\Sigma xi = 0$	$\Sigma x_2 = 0$	$\Sigma y_1^2$	$\Sigma x_1^2$	$\Sigma_2^2$	= 594	= 270	= 630
	520	420	620									
	$\bar{y} = 52$ ,	$\bar{x}_1 = 42$ ,	$\bar{x}_2 = 62$									
										$\Sigma x_1 y_1 = 319$	$\Sigma x_2 y_1 = 492$	$\Sigma x_1 y_2 = 240$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u \dots\dots\dots 3.1$$

In the matrix notations

$$\hat{\beta} = (X' X)^{-1} = X' Y$$

Where (when we use die quantities in deviation form).

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \quad x = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix} \text{ so that}$$

$$x' x = \begin{bmatrix} \Sigma x_1^2 & x_1 & x_2 \\ \Sigma x_1 x_2 & \Sigma x_2^3 \end{bmatrix} \text{ and } X' Y = \begin{bmatrix} \Sigma x_1 y \\ \Sigma x_2 y \end{bmatrix}$$

Substituting the relevant quantities, we have

$$X' X = \begin{vmatrix} 270 & 240 \\ 240 & 630 \end{vmatrix} \text{ and } X' Y = \begin{vmatrix} 319 \\ 492 \end{vmatrix}$$

$$X' X = 270 \times 630 - (240 \times 240) \\ = 112500.$$

$$(X' X)^{-1} = \frac{1}{112500} \begin{bmatrix} 630 & -240 \\ -240 & 270 \end{bmatrix}$$

$$= \begin{bmatrix} 0056 & -0021 \\ -0021 & 0024 \end{bmatrix} X' Y = \begin{bmatrix} 319 \\ 492 \end{bmatrix}$$

$$\hat{\beta}_1 = 0.7532$$

$$\hat{\beta}_2 = 0.5109$$

The value of a is determined from the relation:

$$\begin{aligned} a &= \bar{y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 && \dots\dots\dots (3.2) \\ &= 52 - (.7532 \times 42) - (.5109 \times 62) \\ &= 52 \times 31.63 - 31.6578 \\ &= - 11.29 \end{aligned}$$

$$y = 11.09 + 7532 X_1 + 5109 X_2 \quad \dots\dots\dots (3.3)$$

Goodness of Fit  $R^2$

$$\begin{aligned}
 R^2 &= \frac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i}{\Sigma y_i^2} \\
 &= \frac{7532(319) + 5109 \times 492}{594} \\
 &= \frac{240 \cdot 2708 + 251 \cdot 3628}{594} = \frac{491 \cdot 6336}{594} \\
 &= 82 \quad \text{..... (3.4)}
 \end{aligned}$$

The variables  $X_1$  and  $X_2$  explain 82 percent of the total variations in  $Y$ .

$$\Sigma e_i^2 = (1 - R^2) \Sigma y_i^2 = (1 - 82) 594 = 106.92$$

### SUGGESIED READINGS

1. Croxton, F.E. and Cowden D.J.: Applied General Statistics, 2nd ed. New York Prentice Hall, 1946.
2. Freund, J.E. and Williams F.J. : Modern Business Statistics; London, Pitman and Sons, 1959.

\*\*\*\*\*



## LESSON 7 & 8

### MULTIPLE AND PARTIAL CORRELATION

**Dear Student,**

In correlation and regression, the main purpose of the analysis was 'to determine the degree of linear relationship between two variables and to predict the behaviour of the dependent variable on the basis of  $n$  independent variable. But often, it is necessary to find correlation between three or more variates. For example, the stature of men is influenced by those of all their ancestors, and the yield of grain is affected by the amount of irrigation and fertilizers used. Whenever we are, interested in the combined influence of a group variates upon a variate not included in the group, our study is that of multiple regression and multiple correlation. Therefore there arises a need for the study of multiple correlation. Multiple correlation may be defined as *a* statistical tool designed to measure the 'degree of relationship "existing among three or more variables. For example; we may be asked to find the relationship, between the yield of wheat and amount of irrigation, fertilizers, seeds, insecticides and the spacing of the plants.

In a multivariate population, the different variates, may be mutually correlated and the correlation will, in general, be influenced, by the other variates of the population. To study the relationship between any two variables, there are two methods, firstly, we may consider only those members of the observed data in which, the other members have 'specified values. Secondly, we may eliminate mathematically the effect of other variables on the two variables under study. The first method has the disadvantage that it limits the size of the data and also the result of this, will be applicable to only those data in which the other variables have assigned values. In the second method it may not be possible to eliminate the entire effect of other variables but we can easily eliminate the linear effects. The correlation between two variables when the linear effect of the other variables in them has been eliminated from both is called partial correlation.

**Distribution of three variables:** The theory of multiple and partial correlation was developed by Kari Pearson (1896) for three variables and then generalized by G. Udny Yule (1897). For the sake of simplicity, we shall study the distribution of three variables only though the arguments will apply to the case of  $n$  variables also. Let these variables, be measured from their respected means and the quantities so obtained be denoted by  $x_1, x_2$  and  $x_3$ .

In multiple regression with two independent variables, there are three constants involved in the equation. Three normal equations are required to compute the values of three constants. The regression equation of  $X_1$ , on  $X_2$  and  $X_3$  is of the form

$$X_1 = a + b_{12.3} X_2 + b_{13.2} X_3 \dots\dots\dots 1.$$

Where the constants  $a$  and  $b$ 's are such as to give on the average the 'best', estimate of  $x_1$  corresponding to any assigned values of  $x_2$  and  $x_3$ . Thus we are to find  $a$  and  $b$ ' such that

$$\begin{aligned} \mu = \Sigma(x_1 - x_1)^2 &= \Sigma(x_1 - a - b_{12.3} x_2 - b_{13.2} x_3)^2 \\ &= \Sigma x_{123}^2 \end{aligned}$$

$$\text{where } x_{123} = x_1 - a - b_{12.3} x_2 - b_{13.2} x_3 \dots\dots\dots (6.1)$$

is a minimum said the summation is taken place over all sets of values of  $x_2$  and  $x_3$ .

The three normal equations for determining  $a$  &  $b$ 's are

$$\begin{aligned}\Sigma(x_1 - a - b_{12.3} x_2 - b_{13.2} x_3) &= 0 \\ \Sigma x_2(x_1 - a - b_{12.3} x_2 - b_{13.2} x_3) &= 0 \\ \Sigma x_3(x_1 - a - b_{12.3} x_2 - b_{13.2} x_3) &= 0 \\ | \Sigma x_{1.23} &= 0 | \\ | \Sigma x_2 x_{1.23} &= 0 | \\ | \Sigma x_3 x_{1.23} &= 0 | \quad \dots\dots (6.2)\end{aligned}$$

The first of these equations gives  $a = 0$  and the last two equation's may be written in the form.

$$\Sigma x_1 x_2 - b_{12.3} \Sigma x_1^2 - b_{13.3} \Sigma x_2 x_3 = 0 \quad \dots\dots (6.3)$$

$$\Sigma x_1 x_3 - b_{12.3} \Sigma x_2 x_3 - b_{13.3} \Sigma x_3^2 = 0 \quad \dots\dots (6.4)$$

The coefficient of correlation between  $X_1$  and  $X_2$  (from mean) is given by:

$$\begin{aligned}r_{12} &= \frac{\Sigma x_1 x_2}{N \sigma_1 \sigma_2} \\ \text{or } N \sigma_1 \sigma_2 r_{12} &= \Sigma x_1 x_2 \\ r_{13} &= \frac{\Sigma x_1 x_3}{N \sigma_1 \sigma_3} \\ N \sigma_1 \sigma_3 r_{13} &= \Sigma x_1 x_3 \\ r_{23} &= \frac{\Sigma x_2 x_3}{N \sigma_2 \sigma_3}\end{aligned}$$

Similarly

$$N \sigma_{2.3} r_{23} = \Sigma x_2 x_3$$

The standard deviation (from mean) of  $X_1$  series

$$\sigma_1 = \sqrt{\frac{\Sigma x_1^2}{N}} \quad \text{or } N \sigma_1^2 = \Sigma x_1^2$$

Similarly  $N \sigma_1^2 = \Sigma x_1^2$  and  $N \sigma_3^2 = \Sigma x_3^2$

Substituting above values in the equations (6.3) and (6.4) we have

$$N r_{12} \sigma_1 \sigma_2 = N b_{12.3} \sigma_2^2 + N b_{13.2} r_{23} \sigma_2 \sigma_3 \quad \dots\dots (6.5)$$

$$N r_{13} \sigma_1 \sigma_3 = N b_{12.3} \sigma_{23} + r_{23} + b_{13.2} \sigma_3$$

Since  $N \sigma_2$  is common in equation 5 and  $N \sigma_3$  in equation \dots\dots (6.6)

$$r_{12} \sigma_1 = b_{12.3} \sigma_2 + b_{13.2} r_{23} \sigma_3 \quad \dots\dots (6.7)$$

$$r_{13} \sigma_1 = b_{12.3} \sigma_2 + r_{23} b_{13.2} \sigma_3 \quad \dots\dots (6.8)$$

Dividing the equation (6.8) by  $r_{23}$  and subtracting the result from equation (6.7) we get:

$$r_{12} \sigma_1 = b_{12.3} \sigma_2 + b_{13.2} \sigma_3 r_{23}$$

$$\frac{r_{13}}{r_{23}} \sigma_1 = b_{12.3} \sigma_2 + b_{13.2} \frac{\sigma_3}{r_{23}}$$

- - - -  
 .....

$$\sigma_1 r_{12} - \sigma_1 \frac{r_{13}}{r_{23}} = b_{12.3} \sigma_3 r_{23} - b_{13.2} \frac{\sigma_3}{r_{23}}$$

$$\text{or } \sigma_1 \left( r_{12} \frac{r_{13}}{r_{23}} \right) = b_{12.3} \sigma_3 \left( r_{23} \frac{1}{r_{23}} \right)$$

$$\text{or } \sigma_1 \left( \frac{r_{12} r_{23} - r_{13}}{r_{23}} \right) = b_{13.2} \sigma_3 \left( \frac{r_{23}^2 - 1}{r_{23}} \right)$$

$$\text{or } \sigma_1 \frac{\sigma_1}{\sigma_3} \left( \frac{r_{12} r_{23} - r_{13}}{r_{23}} \right) \left( \frac{r_{23}}{r_{23}^2 - 1} \right)$$

$$\text{or } b_{13.3} = \frac{\sigma_1}{\sigma_3} \left( \frac{r_{12} r_{23} - r_{13}}{r_{23}^2 - 1} \right)$$

Substituting the value of  $b_{13.2}$  in equation (7) we get:

$$\text{or } b_{12.3} = \frac{\sigma_1}{\sigma_2} \left( \frac{r_{12} r_{23} - r_{12}}{r_{23}^2 - 1} \right)$$

or  $b_{13.3}$  can be write as

$$= \frac{\sigma_1}{\sigma_3} \left( \frac{r_{12} r_{12} - r_{23}}{1 - r_{23}^2} \right)$$

and

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \left( \frac{r_{12} r_{12} - r_{23}}{1 - r_{23}^2} \right)$$

$$b_{12.3} = \begin{vmatrix} r_{12} & \sigma_1 & r_{13} & \sigma_3 \\ r_{12} & \sigma_1 & & \sigma_3 \end{vmatrix} + \begin{vmatrix} \sigma_2 & r_{23} & \sigma_3 \\ r_{23} & \sigma_2 & \sigma_3 \end{vmatrix}$$

$$= \frac{-\frac{\sigma_1}{\sigma_2} \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}}$$

..... (6.10)

$$= \frac{\sigma_1}{\sigma_2} \frac{\Delta 12}{\Delta 11}$$

$$b_{12.3} = \frac{-\frac{\sigma_3}{\sigma_2} \begin{vmatrix} 1 & r_{12} \\ r_{13} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2} \frac{\Delta 13}{\Delta 11}$$

Where  $\Delta_{ij}$  is the co-factor of the element in the  $i$ th row and  $j$ th column in the determinant.

$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

1 and  $r_{33} = 1$  and  $r_{13} = r_{31}$ ;  $r_{12} = r_{21}$  and  $r_{23} = r_{32}$

Hence on substituting the values of  $b_{12.3}$  and  $b_{13.2}$ , the equation to the regression plane of  $x_1$  and  $x_2$  and  $x_3$  is

$$X_1 = \left[ -\frac{\sigma_1}{\sigma_2} \frac{\Delta 12}{\Delta 11} \right] X_2 + \left[ -\frac{\sigma_1}{\sigma_3} \frac{\Delta 13}{\Delta 11} \right] x_3 \quad \dots\dots\dots (6.11)$$

**Example 6.1:** Find the least square regression equation of  $X_3$  on  $X_1$  and  $X_2$  from the following data :

$$\bar{X} = 6.8 \quad \bar{X}_2 = 7.0 \quad \bar{X}_3 = 74$$

$$\sigma_1 = 1.0 \quad \sigma_2 = 0.8 \quad \sigma_3 = 9.0$$

$$r_{12} = 0.6 \quad r_{13} = 0.7 \quad r_{23} = 0.65$$

The regression equation, of  $X_3$  on  $X_1$  and  $X_2$  is of the form:

$$X_3 \text{ and } a_{3.12} + b_{3.12} X_1 + b_{32.1} X_2$$

Computation of regression on coefficients:

$$b_{31.2} = \frac{\sigma_3}{\sigma_1} \left( \frac{r_{13} - r_{23} r_{12}}{1 - r_{12}^2} \right)$$

Substituting the values,

$$b_{31.2} = \frac{9.0}{1.0} \left( \frac{0.7 - 0.65 \times 0.6}{1 - (0.6)^2} \right)$$

$$b_{31.2} = 9 \left( \frac{0.7 - 0.390}{1 - 0.36} \right) = 4.36$$

$$\text{A d } b_{31.2} = \frac{\sigma_3}{\sigma_2} \left( \frac{r_{23} - r_{13} r_{12}}{1 - r_{12}^2} \right)$$

$$= \frac{9.0}{0.8} \left( \frac{0.65 - 0.7 \times 0.6}{1 - (0.6)^2} \right)$$

$$= 4.04$$

The value of  $a_{3.12}$  may be computed from the following equation:

$$a_{3.12} + \bar{X}_3 - b_{31.2} \bar{X}_1 - b_{32.1} \bar{X}_2 \\ = 74 - 4.36 \times 6.8 - 4.08 \times 7.0 = 16.7m$$

Thus the required equation is

$$X_3 - 16.07 + 4.36X_1 + 4.04 X_2 \text{ Ans.}$$

**Example 2 :** Given the following: Determine the regression equation of  $X_1$  on  $X_2$  and  $X_3$ .

$$\bar{X} = 6.8 \quad \sigma_1 = 1.0 \quad r_{12} = 0.6$$

$$\bar{X}_2 = 7.0 \quad \sigma_2 = 0.8 \quad r_{13} = 0.7$$

$$\bar{X}_3 = 7.4 \quad \sigma_3 = 0.9 \quad r_{13} = 0.8$$

**Sol.** Regression equation of  $X_1$  on  $X_2$  and  $X_3$  in determinant form is

$$\frac{\Delta_{11}}{\sigma_1} x_1 + \frac{\Delta_{12}}{\sigma_2} x_2 + \frac{\Delta_{13}}{\sigma_3} x_3 = 0$$

The values of cofactors  $\Delta_{11}$ ,  $\Delta_{12}$  and  $\Delta_{13}$  are determined from the following relationship

$$\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$$

Since  $r_{11} = 1$ ,  $r_{22} = 1$ ; and  $r_{33} = 1$  and  $r_{12} = r_{21}$  and substituting the values  $r_{13} = r_{31}$  and  $r_{23} = r_{32}$ .

$$\Delta = \begin{vmatrix} 1 & 0.6 & 0.7 \\ 0.6 & 1 & 0.8 \\ 0.7 & 0.8 & 1 \end{vmatrix}$$

$$\Delta_{11} = \begin{vmatrix} 1 & 0.8 \\ 0.8 & 1 \end{vmatrix} = (1 - (0.8)^2) = 1 - 0.64 = 0.36.$$

$$\Delta_{12} = \begin{vmatrix} 0.6 & 0.8 \\ 0.7 & 1 \end{vmatrix} = [0.6 - (0.8 \times 0.7)] - 0.6 - 0.56 = 0.04$$

$$\Delta_{13} = \begin{vmatrix} 0.6 & 1 \\ 0.7 & 0.8 \end{vmatrix} = [0.48 - 0.70] = - 0.22$$

Substituting the values in above regression equation.

$$\frac{0.36}{1} x_1 + \frac{0.04}{0.8} x_2 + \frac{-0.22}{0.9} x_3 = 0$$

$$0.36 x_1 = - 0.05 x_2 + 0.24 x_3$$

$$x_1 = \frac{0.05}{.36} x_2 + \frac{.24}{.36} x_3$$

$$x_1 = 0.138x_2 + 0.66x_3$$

The value of 'a' is computed from the following equation :

$$a_{1.23} + \bar{X}_1 - b_{12.3} \bar{X}_2 - b_{13.2} \bar{X}_3$$

Putting the value of  $\bar{X}_1$ ,  $\bar{X}_2$ ,  $\bar{X}_3$ ,  $b_{12.3}$  and  $b_{13.2}$

$$\begin{aligned} a_{1.23} &= 6.8 + 0.138 \cdot 17 - .66 \times 7.4 \\ &= 6.8 + 0.966 - 4.88 \\ &= 2.88 \end{aligned}$$

Thus the required regression equation of  $X_1$  on  $X_2$  and  $X_3$  becomes  $X_2 = 2.88 - 0.138 X_2 + 0.66 X_3$ .

**Variance of a Residual :** We shall now obtain formula for the variance  $\sigma_{1.23}^2$  of the residual  $x_{1.23}$  (the deviation of the observed Values  $x_1$  from its computed value of the regression plane), in terms of  $\sigma_1^2$  and the correlation coefficients.

$$\begin{aligned} \text{We have } N\sigma_{1.23}^2 &= \sum x_{1.23}^2 = \sum x_1 x_{1.23} \\ &= \sum x_1 (x_1 - b_{12.3} x_2 - b_{13.2} x_3) \\ &= N\sigma_1^2 - Nb_{12.3} \sigma_1 \sigma_2 r_{12} - Nb_{13.2} \sigma_1 \sigma_3 r_{13} \\ \text{Or } N\sigma_1^2 - N\sigma_{1.23}^2 &= Nb_{12.3} \sigma_1 \sigma_2 r_{12} + Nb_{13.2} \sigma_1 \sigma_3 r_{13} \end{aligned}$$

Dividing both sides by  $\sigma_1$

or

$$\begin{aligned} \sigma_1 - \frac{\sigma_{1.23}^2}{\sigma_1} &= b_{12.3} \sigma_2 r_{12} + b_{13.2} \sigma_3 r_{13} \\ \sigma_1 \left( 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2} \right) &= b_{12.3} \sigma_2 r_{12} + b_{13.2} \sigma_3 r_{13} \end{aligned}$$

eliminating  $b_{12.3}$  and  $b_{13.2}$  between this equation and equations.

$$\begin{aligned} r_{12} \sigma_1 &= b_{12.3} \sigma_2 + b_{13.2} r_{13} \sigma_3 \\ r_{13} \sigma_1 &= b_{12.3} \sigma_2 + b_{13.2} \sigma_3 \end{aligned}$$

We have

$$\begin{aligned} &\left| 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2} \quad r_{12} \quad r_{13} \right| \\ &\left| r_{12} \quad 1 \quad r_{13} \right| = 0 \\ &\left| r_{13} \quad r_{23} \quad 1 \right| \\ &\Delta - \frac{\sigma_{1.23}^2}{\sigma_1^2} \Delta_{11} = 0 \quad \sigma_{1.23}^2 = \sigma_1^2 \frac{\Delta}{\Delta_{11}} \end{aligned}$$

Thus the variance of residual of order 2 is expressible in variance of zero order and correlation of zero order.

**6. Multiple Correlation Coefficient:** Consider the regression equation for  $x_1$  on  $x_2$  and  $x_3$  viz.,

Next the correlation between  $x_1$  (observed value of the variable) and  $X_1$  (expected value of the variable) is given by

$$R_{12.3} = \frac{\Sigma x_1 X_1}{\sqrt{(\Sigma x_1^2)(\Sigma X_1^2)}}$$

$$\begin{aligned} \text{We have } \Sigma x X_1 &= \Sigma(x_1 (x_1 - x_{1.23})) = \Sigma x_1^2 - \Sigma x_1 x_{1.23} = \Sigma x_1^2 - \Sigma x_{1.23}^2 \\ &= N\sigma_1^2 - N\sigma_{1.23}^2 = N(\sigma_1^2 - \sigma_{1.23}^2) \end{aligned}$$

$$\begin{aligned} \text{Also } \Sigma x_1^2 &= \Sigma(x_1 x_{12.3})^2 \\ &= \Sigma x_1^2 = 2 \Sigma x_1 x_{1.23} + \Sigma x_{1.23}^2 \\ &= \Sigma x_1^2 = 2 \Sigma x_{1.23}^2 + \Sigma x_{1.23}^2 \\ &= \Sigma x_1^2 = \Sigma x_{1.23}^2 \\ &= N\sigma_1^2 - N\sigma_{1.23}^2 \end{aligned}$$

$$\begin{aligned} 6.1 \quad R_{12.3} &= \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma \sqrt{\sigma_1^2} \sqrt{\sigma_1^2 - \sigma_{1.23}^2}} \\ &= \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2 - \sigma_{1.23}^2}} = \left(1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}\right)^{1/2} \\ &= \sqrt{1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}} \end{aligned}$$

Where  $\sigma_1$  stands for the standard deviation of a dependent variable  $x_1$   $\sigma_{1.23}$  stand for the standard error of estimated of  $x_1$  on  $x_2$  and  $x_3$ .

It is computed by the following formula :

$$\sigma_{1.23} = \frac{\Sigma(X_1 - X_{1est})^2}{N}$$

Where

$X_{1est}$  stands for the estimated value of  $X_1$  as compared with the aid of regression equation of  $X_1$  on  $X_2$  and  $X_3$ .

It may also be computed by the following formulae

$$\begin{aligned} \sigma_{1.23} &= \sigma_1 \sqrt{(1 - r_{12}^2)} \sqrt{(1 - r_{13.2}^2)} \\ \text{Or } s_{1.23}^2 &= \sigma_1^2 (1 - r_{12}^2) (1 - r_{13.2}^2) \end{aligned}$$

**6.2** For the two independent variables, the coefficient of multiple correction may also be determined with the help of zero order coefficient of correlation.

$$R_{12.3} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

Where  $r_{12}$ ,  $r_{13}$  and  $r_{23}$  stand for zero-order coefficient of correlation

$$R_{1.24} = \sqrt{\frac{r_{12}^2 + r_{14}^2 - 2r_{12}r_{14}r_{24}}{1 - r_{24}^2}}$$

**6.3** Coefficient of Multiple correlation may also be computed with the help of total variation and explained variation.

$$R_{1.23} = \sqrt{\frac{\Sigma X_c^2}{\Sigma x_1^2}}$$

Where  $x_{1.23}^2$  stands for explained variations in the values of variable  $X_1$  which has been explained by two (independent variables  $X_2$  and  $X_3$ ). It may be completed by the following formula:

$$x_{1.23}^2 = b_{12.3} \Sigma x_1 x_2 + b_{13.2} \Sigma x_1 x_3$$

**Example :** Calculate the multiple correlation coefficient of  $X_1$  and  $X_2$  and  $X_3$  from the following data:

$X_1$	:	5	7	8	10	12	16
$X_2$	:	10	18	6	5	4	3
$X_3$	:	21	21	15	17	20	12

**Solution:**

$X_1$	$X_2$	$X_3$	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_1 - \bar{X}_2$	$\bar{X}_2 - \bar{X}_3$	$\bar{X}_1 - \bar{X}_3$	$x_1$	$x_2$	$x_3$	$x_{12}$	$x_{22}$	$x_{32}$	$x_1 x_2$	$x_2 x_3$	$x_1 x_3$
5	10	21	-5	+4	+3	25	16	9	-20	12	-15						
7	8	21	-3	+2	+3	9	4	9	-6	6	-9						
8	6	15	-2	0	-3	4	0	9	0	0	+6						
10	5	17	0	-1	+1	0	1	1	0	+1	0						
12	4	20	+2	-2	+2	4	4	4	-4	-4	4						
18	3	14	+8	-3	-4	64	9	16	-24	+12	-32						
$\Sigma X_1$	$\Sigma X_2$	$\Sigma X_3$	$\Sigma x_1$	$\Sigma x_2$	$\Sigma x_3$	$\Sigma x_1^2$	$\Sigma x_2^2$	$\Sigma x_3^2$	$\Sigma x_1 x_2$	$\Sigma x_2 x_3$	$\Sigma x_1 x_3$						
=60	=36	=108	=0	=0	=106	=34											
$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$															
=10	=6	=18															

$$\Sigma x_2 x_3 = 27$$

$$\Sigma X_1 X_3 = -46$$

For computing the value of explained variations. We need the value of  $b_{12.3}$  and  $b_{13.2}$  and these may be computed, from regression equation of  $X_1$  on  $X_2$  and  $X_3$  which may be written in the deviation form as given below:

$$x_1 = b_{12.3} x_2 + b_{13.2} x_3$$

Normal equations:

$$\Sigma x_1 x_2 + b_{12.3} \Sigma x_2^2 + b_{13.2} \Sigma x_2 x_3 \dots\dots (1)$$

$$\Sigma x_1 x_3 + b_{12.3} \Sigma x_2 x_3 + b_{13.2} \Sigma x_3^2 \dots\dots (2)$$



Substituting the values in above equations:

$$-54 = 34 b_{12.3} + 27 b_{13.2} \quad \dots (3)$$

$$-46 = 27 b_{12.3} + 48 b_{13.2} \quad \dots (4)$$

Multiplying equation 3rd by 16 and 4th by 9 we get:

$$-864 = 544 b_{12.3} + 432 b_{13.2}$$

$$-414 = 243 b_{12.3} + 432 b_{13.2}$$

$$\begin{array}{r} - \quad - \quad - \\ -450 = 301 b_{13.2} \end{array}$$

$$b_{13.2} = - \frac{450}{301} = - 1.4950.$$

$$b_{12.3} = 0.4010.$$

**Calculation of Explained variation:**

$$\begin{aligned} \Sigma x_c^2 1.23 &= b_{12.3} \Sigma x_1 x_2 + b_{13.2} \Sigma x_1 x_3 \\ &= - 1.4950 (-54) + 0.4010 (-46) \\ &= 80.73 + 18.4473 \end{aligned}$$

$$\begin{aligned} R^2 &= \frac{\Sigma X_c^2 1.23}{\Sigma x_1^2} = \sqrt{\frac{62.2826}{106}} \\ &= 0.7665 \text{ Ans.} \end{aligned}$$

### Partial Correlation Coefficient

A partial correlation coefficient measures the relationship between any two variables, when all other variables connected with those two are kept constant. For example, let us assume we want to measure the correlation between the number of cold drinks ( $X_1$ ) consumed during summers in Shimla and the number of tourists ( $X_2$ ) coming to Shimla. It is obvious feat both these variables are strongly, influenced by weather conditions, which, we may designate by  $X_3$ . On a priori, grounds we expect  $X_1$ , and  $X_2$  to be positively correlated when a large number of tourist arrive in the summer resort (Shimla), one should expect a high consumption of cold drinks and vice-versa. The computation of the simple correlation coefficient between  $X_1$  and  $X_2$  may not reveal the true relationship connecting; these two variables, because of the influence of third variable  $X_3$ . In other words the above positive, relationship Between number of tourists and number; of cold drinks consumed is expected to hold if weather conditions can be assumed constant If weather changes, the relationship between the consumption of cold drinks and number of tourists arrived may be distorted to such an extent as to appear negative. Thus if the weather is cold (due to Unexpected rain), the number of tourists will be less, but because of the chilly weather, they will prefer to consume hot drinks (tea or coffee) rather than cold drinks. If we overlook weather and look only at  $X_1$  and  $X_2$ . We will observe a negative correlation between, these two variables which is explained by the fact cold drinks and as well number of visitors is affected by unexpected weather. In order to measure the true relationship between  $X_1$  and  $X_2$  we must find some way of accounting for changes in  $X_3$ . This is achieved with the partial correlation coefficient between  $X_1$  and  $X_2$  when  $X_3$  is kept constant. The partial correlation coefficient is determined in terms of the simple correlation coefficient among the various variables involved in multiple relationship. In the above mentioned example there are three simple Correlation Coefficients.

$r_{12}$  = Correlation Coefficient between  $X_1$  and  $X_2$ .

$R_{13}$  = Correlation Coefficient between  $X_1$  and  $X_3$ .

$X_{23}$  = Correlation Coefficient between  $X_2$  and  $X_3$ .

There are two partial correlation coefficients.

$r_{12.3}$  = partial correlation coefficient between  $X_1$  and  $X_2$  when  $X_3$  is kept constant

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

and

$r_{13.2}$  = partial correlation coefficient between  $X_1$  and  $X_3$  when  $X_2$  is kept constant.

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

We may postulate the following functional relationship between the number of tourist in Shimla ( $X_1$ ) and She consumption of cold drinks ( $Y$ ) :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + 4$$

$Y$  = consumption of cold drinks

$X_1$  = number of tourists, in Shimla (summer)

$X_2$  = weather conditions measured by an index of rainfall or temperature

The partial correlation coefficient measures the correlation between any two variables, when all the other variables are held constant, that is when we assume, that there is no other, factor influencing, the relationship. Partial correlation coefficient for the model including two explanatory variables is

$$r_{yx_1, x_2} = \frac{r_{yx1} - r_{yx2} r_{x1x2}}{\sqrt{1 - r_{yx2}^2} \sqrt{1 - r_{x1x2}^2}}$$

The partial correlation, on coefficient between  $y$  and  $x_2$  when  $x_1$  is kept constant is obtained from this expression by interchanging the position of the subscripts 1 and 2.

$$r_{yx_2, x_1} = \frac{r_{yx1} - r_{yx2} r_{x1x2}}{\sqrt{1 - r_{yx1}^2} \sqrt{1 - r_{x1x2}^2}}$$

**Proof:** The following relationship between regression coefficients and simple correlation coefficients has been established in lesson.

$$r^2_{yx} = \hat{b}_1 \frac{\sum x^2}{\sum y^2} \quad \text{or} \quad \hat{b}_1 = r_{yx} = \left( \frac{\sqrt{\sum y^2}}{\sqrt{\sum x^2}} \right)$$

In the above mentioned case

$$\hat{a}_1 = r_{yx_2} \left( \frac{\sqrt{\sum y^2}}{\sqrt{\sum x_2^2}} \right) \quad \text{and} \quad \hat{c}_1 = r_{x1x2} = \left( \frac{\sqrt{\sum x_1^2}}{\sqrt{\sum x_2^2}} \right)$$

From the correlation coefficient of the two regressions, we get

$$r^2_{yx_2} = 1 - \frac{\Sigma e_2^2}{\Sigma y_1^2} = 1 - \frac{\Sigma x_2^{*2}}{\Sigma y_1^2}$$

and  $r^2_{x_1x_2} = 1 - \frac{\Sigma e_2^2}{\Sigma x_1^2} = 1 - \frac{\Sigma x_1^{*2}}{\Sigma x_1^2}$

Therefore

$$\Sigma y^{*2} = \Sigma y^2 (1 - r^2_{yx_2})$$

$$\text{and } \Sigma x_1^{*2} = \Sigma x_1^2 (1 - r^2_{x_1x_2})$$

Substitute the terms with asterisks in the formula of partial correlation coefficient

$$\begin{aligned} r_{y_1 x_2} &= \frac{\Sigma (y - \hat{a}_1 x_2)(x_1 - \hat{c}_1 x_2)}{\sqrt{\Sigma y^2 (1 - r^2_{yx_2})} \sqrt{\Sigma x_1^2 (1 - r^2_{x_1x_2})}} \\ &= \frac{\Sigma (yx_1 - \hat{a}_1 x_1 x_2 - \hat{c}_1 y x_2 + \hat{a}_1 \hat{c}_1 x_2^2)}{\sqrt{\Sigma y^2 \Sigma x_1^2} \sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}} \end{aligned}$$

The rationalization of these formulae may be propounded as follows. In order to measure the pure correlation between y(cold drinks) and  $x_1$  (numbers of tourists), the influence of the third variable  $X_2$  (weather conditions) has to be eliminated from both Y and  $X_1$ . This can be done by regressing Y on  $X_2$  and  $X_1$  on  $X_2$ .

$$Y = a_0 + a_1 X_2 + u_2$$

$$X_1 = c_0 + C_1 X_2 + u_2$$

Where  $u_1$  and  $u_2$  and error term satisfying the usual assumption of zero mean and constant variance. From the application of least squares, fee following, estimates are obtained.

$$\hat{a}_1 = \frac{\Sigma y x_2}{\Sigma X_2^2} \quad r^2_{yx_2} = 1 - \frac{\Sigma e^2}{\Sigma y^2}$$

$$\hat{c}_1 = \frac{\Sigma x_1 x_2}{\Sigma x_2^2} \quad r^2_{y_1x_2} = 1 - \frac{\Sigma e^2}{\Sigma x_1^2}$$

The unexplained variance in each regression is

$$e_1 = Y_1 - \hat{y} = Y_1 - \hat{a}_1 x_1 = Y^*$$

and  $e^2 = X_1 - \hat{X}_1 = X_1 - \hat{c}_1 x_2 = X_1^*$

These are variations' in Y and in  $X_1$ , respectively, left unexplained after removing the influence of  $X_2$  (weather conditions).

The partial correlation coefficient between Y and  $X_1$  which is defined as the simple correlation between the above mentioned unexplained parts of the two variables can be proved.

$$r_{y_1 x_2} r_{y^* x_1^*} = \frac{\Sigma Y^* X_1^*}{\sqrt{\Sigma Y^{*2}} \sqrt{\Sigma X_1^{*2}}}$$

$$= \frac{\Sigma y x_1 - \hat{a}_1 x_1 x_2 - \hat{c}_1 \Sigma y x_2 + \hat{a}_1 \hat{c}_1 \Sigma x_2^2}{\sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2} \sqrt{(1-r_{yx2}^2)(1-r^2 x_1 x_2)}}$$

Substituting the values of  $\hat{a}_1$  and  $\hat{c}_1$  For their expression, we get

$$\begin{aligned} & \Sigma r y x_1 - r_{yx2} \left( \frac{\sqrt{\Sigma y^2}}{\sqrt{\Sigma x_2^2}} \right) \Sigma x_1 x_2 - r_{x1x2} \left( \frac{\sqrt{\Sigma x_1^2}}{\sqrt{\Sigma x_2^2}} \right) \Sigma y x_2 \times r_{yx2} \\ & \Sigma r x_1 x_2 \left( \frac{\sqrt{\Sigma y^2 \Sigma x_1^2}}{\sqrt{\Sigma x_2^2}} \right) \Sigma x_2^2 \end{aligned}$$

$$\sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2} \sqrt{(1-r_{yx2}^2)(1-r^2 x_1 x_2)}$$

Multiplying each term of the numerator by appropriate unitary term so that these are transformed into simple correlation coefficients. Multiply the first term

$$\left[ \sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2} \sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2} \right] = 1, \text{ the second term by}$$

$$\left[ \sqrt{\Sigma x_1^2} / \sqrt{\Sigma x_1^2} \right] = 1, \text{ and we get}$$

$$r y x_1 x_2 = \Sigma y x_1 \left( \frac{\sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2}}{\sqrt{\Sigma y^2} \sqrt{\Sigma x_2^2}} \right) - r_{yx2} \Sigma x_1 x_2 \left( \frac{\sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2}}{\sqrt{\Sigma y^2} \sqrt{\Sigma x_2^2}} \right)$$

$$r x_1 x_2 \left( \frac{\Sigma y x_2 \sqrt{\Sigma x_1^2} \sqrt{\Sigma y^2}}{\sqrt{\Sigma y^2} \sqrt{\Sigma x_2^2}} \right) + r y x_2 r x_1 x_2 \sqrt{\Sigma y^2 \Sigma x_1^2}$$

$$\sqrt{\Sigma y^2} \sqrt{\Sigma x_1^2} \sqrt{(1-r_{yx2}^2)(1-r^2 x_1 x_2)}$$

$$\frac{\sqrt{\Sigma y^2 \Sigma x_1^2} (r_{yx1} - r_{yx2} r_{x1x2} - r_{yx2} - r_{x1x2} + r_{yx2} r_{x1x2})}{\sqrt{\Sigma y^2 \Sigma x_1^2} \sqrt{(1-r_{yx2}^2)(1-r^2 x_1 x_2)}}$$

$$= \frac{r_{yx1} - r_{yx2} r_{x1x2}}{\sqrt{1-r_{yx2}^2} \sqrt{1-r_{x1x2}^2}}$$

**Problem:**

If  $r_{12} = +0.80$ ,  $r_{13} = -0.40$ ,  $r_{23} = -0.56$ , find the values of  $r_{12.3}$ ,  $r_{13.2}$ , and  $r_{23.1}$ . we have

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1-r_{23}^2} \sqrt{1-r_{13.2}^2}}$$

$$\begin{aligned}
&= \frac{0.80 - (-0.40)(-0.56)}{\sqrt{\{1 - (0.40)^2\}\{1 - (0.56)^2\}}} \\
&= 0.759
\end{aligned}$$

$$\begin{aligned}
\text{Similarly } r_{13.2} &= \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} \\
&= 0.097
\end{aligned}$$

$$\begin{aligned}
\text{and } r_{23.1} &= \frac{r_{23} - r_{21} r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} \\
&= 0.436
\end{aligned}$$

**Example:** On the basis of observations made on 50 households the linear correlation of coefficients, between  $X_1$  (quantity demanded of tea),  $X_2$  price of tea  $X_3$  (households income) one as follows :

$$r_{12} = 0.75, r_{13} = 0.80, r_{23} = 0.55.$$

Where  $r_{ij}$  is the correlation coefficient between  $X_i$  and  $X_j$

Calculate partial correlation coefficient's of :

- (a) quantity demanded of tea with price of; tea.
- (b) quantity demanded of tea with household's income.

**Solution :** (i) The coefficient of partial correlation between quantity demanded and price when the effect of income is kept constant is given by :

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Putting the values

$$\begin{aligned}
&= \frac{0.75 - 0.80 \cdot 0.55}{\sqrt{\{1 - (0.80)^2\}\{1 - (0.55)^2\}}} \\
&= \frac{0.75 - 0.44}{\sqrt{0.36} \times \sqrt{0.69}} = \frac{0.31}{0.6 \cdot .83} \\
&= \frac{0.31}{0.498} = 0.62
\end{aligned}$$

(ii) The coefficient of partial correlation between quantity demanded and the income of households, when the effect of price is kept constant, is given by

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Substituting the given values for  $r_{12}$ ,  $r_{13}$  and  $r_{23}$  we obtain

$$= \frac{0.80 - 0.75 \cdot 0.55}{\sqrt{1 - (0.75)^2} \sqrt{1 - (0.55)^2}}$$

$$= \frac{0.80 - 0.41}{\sqrt{1 - .56} \sqrt{1 - 30}} \frac{0.39}{66 \times 83}.$$

**Example :** The variable  $X_1$  is thought to be a linear function of  $X_2$  and  $X_3$ . A sample of 12 pairs of readings ( $X_2, X_3$ ) produced the values of  $X_1$  shown in Table 6.1 below :

- Find the least square regression equation of  $X_1$  on  $X_2$  and  $X_3$ .
- Determine the estimated values of  $X_1$  from the given values of  $X_2$  and  $X_3$ .
- Estimate  $X_1$  when  $X_2 = 54$ , and  $X_3 = 9$

**Table 6.1**

$X_1$	64	71	53	67	55	58	77	57	56	51	76	68
$X_2$	57	59	49	62	51	50	55	48	52	42	61	57
$X_3$	8	10	6	11	8	7	10	9	10	6	12	9

**Solution:** The linear regression equation of  $X_1$  on  $X_2$  and  $X_3$  can be written

$$X_1 = b_{12.3} + b_{12.3} X_2 + b_{13.2} X_3$$

The normal equations of least square regression equation are

$$\begin{aligned} \Sigma X_1 &= b_{1.23} N + b_{12.3} \Sigma X_2 + b_{13.2} \Sigma X_3 \\ \Sigma X_1 X_2 &= b_{1.23} \Sigma X_2 + b_{12.3} \Sigma X_2^2 + b_{13.2} \Sigma X_2 X_3 \\ \Sigma X_1 X_3 &= b_{1.23} \Sigma X_3 + b_{12.3} \Sigma X_2 X_3 + b_{13.2} \Sigma X_3^2 \end{aligned}$$

$\Sigma X_1 = 753$	$\Sigma X_1^2 = 48139$	$\Sigma X_1 X_2 = 40830$
$\Sigma X_2 = 643$	$\Sigma X_2^2 = 34843$	$\Sigma X_1 X_3 = 5779$
$\Sigma X_3 = 106$	$\Sigma X_3^2 = 976$	$\Sigma X_2 X_3 = 6796$

Using the values, the normal equation (i) becomes.

$$\begin{aligned} 12 b_{1.23} + 643 b_{12.3} + 106 b_{13.2} &= 753. \\ (2) \quad 643 b_{1.23} + 34843 b_{12.3} + 5779 b_{13.2} &= 40830. \\ 106 b_{1.23} + 5779 b_{12.3} + 976 b_{13.2} &= 6796. \end{aligned}$$

Solving  $b_{12.3} = 3.6512$   
 $b_{13.2} = 0.8546$   
 $b_{1.23} = 1.5063$ , and the required regression equation is

$$(3) \quad X_1 = 3.6512 + 0.8546 X_2 + 1.5063 X_3$$

or

$$X_1 = 3.65 + 0.85 X_2 + 1.50 X_3$$

(b) Using the regression equation (3) we obtain the estimated value of  $X_1$ , denoted by  $X_{1\text{est}}$ , by substituting the corresponding values of  $X_2$  and  $X_3$ . For example, substituting  $X_2 = 53$  and  $X_3 = 8$  in (3) we find  $X_{1\text{est}} = 64.414$ . Similarly the other estimated values of  $X_1$  are obtained and given in the Table 6.1 together with the sample values of  $X_1$ .

**Table 6.2**

$X_{1\text{est.}}$	64.41	69.13	54.56	73.20	59.29	56.92	65.71	58.22	63.15	48.58	73.85	65.92
$X_1$	64	71	53	67	55	58	77	57	56	51	76	68

(c) Putting  $X_2 = 54$  and  $X_3 = 9$  (in) (3) the estimate is  $X_{1\text{est.}} = 63.556$  or about 63 Ans.

**Standard Error of Estimate**

Compute the standard, error of estimate of  $X_1$  on  $X_2$  and  $X_3$  for the data of problem 6.1.

**Solution:** From Table 6.1 of problem 6.1 (b), we have

$$\begin{aligned}
 (1) \quad S_{1.23} &= \sqrt{\frac{\sum (X_1^2 - X_{1\text{est}})^2}{N}} \\
 &= \sqrt{\frac{(64 - 64.41)^2 + (71 - 69.13)^2 + \dots + (68 - 65.92)^2}{12}} \\
 &= 4.6447 = 4.6.
 \end{aligned}$$

The population standard estimate is estimated by

$$\hat{S}_{123} = \sqrt{N / N - 3} S_{1.23} = 5.3 \text{ Ans.}$$

or Alternatively

$$\begin{aligned}
 S_{123} &= 8603 \sqrt{\frac{1 - (0.8196)^2 - (0.7698)^2 - (0.7984)^2 + 2(0.8196)(.7689)(.7984)}{1 - (0.7984)^2}} \\
 &= 4.6
 \end{aligned}$$

The standard error of estimate can be found without use of regression equation in the above mentioned method.

**Example :** If  $R_{12.3} = 1$ , Prove that

(a)  $R_{2.13} = 1$

(b)  $R_{3.12} = 1$

**Solution:**

$$(1) \quad R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$\text{and } R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

(a) In (1) setting  $R = 1$  and scaring both sides,

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{23} = 1 - r_{23}^2. \text{ Then}$$

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{23} = 1 - r_{23}^2 \text{ or}$$

$$\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2} = 1.$$

$$\text{i.e. } R_{2.13}^2 = 1 \text{ or } R_{2.13} = 1.$$

Since the coefficient of multiple correlation is considered non-negative.

(b)  $R_{3.12} = 1$  follows from part (a) by interchanging subscripts 2 and 3 in the result

$$R_{2.31} = 1$$

**Example:** If  $R_{1.23} = 0$ , does it necessarily follow that  $R_{2.13} = 0$  ?

**Solution:**

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$R_{1.23} = 0$  if and only if

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0$$

or  $r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Since  $2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2$

By putting these values

$$\begin{aligned} &= \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} \\ &= \sqrt{\frac{r_{23}^2 + r_{13}^2}{1 - r_{13}^2}} \end{aligned}$$

which is not necessarily zero.

**Partial Correlation:** Compute the coefficient of linear partial correlation (a)  $r_{12.3}$  (b)  $r_{13.2}$  and (c)  $r_{23.1}$  for the data in problem 6.1.

(a) The quantity  $r_{12}$  is the linear correlation coefficient between the variables  $X_1$  and  $X_2$ , ignoring the variable  $X_3$ .

$$\begin{aligned} r_{12} &= \frac{N \Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)}{\sqrt{[N \Sigma X_1^2 - (\Sigma X_1)^2][N \Sigma X_2^2 - (\Sigma X_2)^2]}} \\ &= \frac{(12)(40830) - (753) \times (643)}{\sqrt{[(12)(48131) - (753)^2][(12)(34843) - (643)^2]}} \\ &= 0.8196 \text{ or } 0.82. \end{aligned}$$

Using the above formula, we obtain

$$r_{13} = 0.7698 \text{ or } 0.77$$

$$r_{23} = 0.7984 \text{ or } 0.80$$

Now  $r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{(1 - r_{13}^2)(1 - r_{23}^2)}$



$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{(1 - r_{12}^2)(1 - r_{13}^2)}$$

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{(1 - r_{12}^2)(1 - r_{13}^2)}$$

Putting the values, we find

$$r_{12.3} = 0.5334$$

$$r_{13.2} = 0.3346$$

$$r_{12.3} = 0.4580$$

It follows that the constant  $X_3$  the correlation coefficient between  $X_1$  and  $X_2$  is 0.53. For constant  $X_2$ , the correlation coefficient between  $X_1$  and  $X_3$  is only 0.33. Since these results are based-on a small sample of only 12 observations, they are of course not that reliable as those which would be obtained from a larger samples.

### **Suggested Readings**

1. Croxton, F.E. and Cowden D.J. : Applied General, Statistics, 2nd Ed. New York Prentice Halt, 1946.
2. Mills, F.C. Statistical Methods. 3rd Ed. London, Pitman and Sons, 1955.

\*\*\*\*\*

## LESSON-9

### A PROBABILITY THEORY & CONCEPT OF PROBABILITY DISTRIBUTION & A DENSITY FUNCTION

**Dear Student,**

According to the syllabus, you are required to have an elementary knowledge of probability. In order to understand the normal, probability curve, some understanding of probability is needed. Knowledge of normal curve itself is important to understand the theory of sampling.

#### **Probability**

The idea of probability has great importance in many decision making problems. Most of the people like an HMT watch because the probability that it gives correct time is very high. A person will assign very low probability to the event that an Usha Machine selected from a lot will be defective. Such probabilities help persons in making decisions about the purchase of articles. Manufacturers also benefit from the idea of probability. Hence probability theory help in various decision making situations.

#### **Different Approaches of Probability**

- (i) Classical Approach
- (ii) Relative frequency Approach
- (iii) Subjective Approach

#### **8.1 Classical Approach**

Definition of an experiment can result in  $N$  different, equally likely outcomes and  $NA$  of these outcomes correspond to event  $A$ , then the probability of. event  $A$  is

$$P(A) = \frac{NA}{N}$$

Let  $A$  be event not  $A$ . Then probability of event not  $A$  is

$$P(A) = \frac{N - NA}{N} = 1 - \frac{NA}{N} \quad 1 - P(A)$$

$$\text{Thus } P(A) + P(A) = 1$$

The word experiment is used to describe, any act that can be repeated under given conditions. The experiment may be tossing of one or more coins rolling die or drawing a card from a pack of 52 cards. The coin may land either head or tail. These are the outcomes of the experiment of tossing a coin. Each outcome is called a simple event. A simple event is the outcome of an experiment which cannot be decomposed into a combination of other events.

Suppose an experiment consists of casting a pair of dice, in this case, sample space consists of 36 different elementary events. This is shown in this table.

No. of spots in a die I

Y/X	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Suppose event A is defined as throwing a total of 7 point. In this case, event A is a composite event. The event A consists of 6 simple events, (6,1) (5,2) (4,3) (3, 4), (2,5), (1, 6). If dice are fair, then all the 36 events are. equally, likely. Now what is the probability of occurrence of event A. Hence  $N_A = 6$  and  $N = 36$ .

$$\therefore P(A) = \frac{N_A}{N} = \frac{6}{36} = \frac{1}{6}.$$

Note that probability of Not—A is

$$P(\bar{A}) = 1 - P(A) = 1 - 1/6 = 5/6$$

Note that if a single die is rolled, then there will be six points in the sample space. The probability of each will be  $1/6$ . Since there are there even numbers (2,4 and 6) and three odd numbers (1,3 and 5), hence  $P(\text{even number}) = (\text{odd number}) = 1/2$ .

## 8.2 Difficulties in Classical Approach :

- (i) One difficulty in regard the phrase “equally likely” We know that as long as a deck of playing cards is well shuffled. One is equally likely to draw any of the 52 cards. Similarly as long as a die is fair each side is equally likely to turn up. But it is very difficult to Verify that these conditions apply in the given instance. In reality the assumption that different outcomes are equally likely is not justified. The assumption is based on abstract reasoning. It is not based on experience.
- (ii) One more difficulty arises when number of cases in a trial is infinite. No one can observe infinite number of trials.
- (ii) If we use the classical approach, men it will *not* be possible to assign probabilities to certain events. For example what is the Probability that it will rain during the 48 hours or what is the probability that Ram will become a millionaire within two years? Classical definition breaks down in such cases. Here the events are not the result of an experiment that can be repeated under same conditions.

### 1.3 Relative Frequency Approach :

In certain cases relative frequency is taken as an estimate of true, probability. If an experiment is repeated  $n$  times under uniform conditions and the outcome  $A$  is observed  $m$  times. In this case relative frequency is  $\frac{m}{n}$  and is taken as probability of event  $A$ , i.e.,  $P(A) = \frac{m}{n}$ .

But according to relative frequency approach it is assumed that relative frequency  $\frac{m}{n}$  approach true probability as  $n$  becomes large. Thus probability of an event.  $A$  is Defined, as a number which the relative frequency tends to approach as  $n$  tends to infinite. The main difficulty that arises with this definition is that it is very costly and time consuming.

#### Subjective Approach

According to this approach, probability measures the confidence that a particular individual has in the truth of a particular proposition. On the basis of this definition different persons may assign different probabilities to an event by using the same evidence. The probabilities assigned are not necessarily based on abstract reasoning.

The subjective approach has become important in recent years. A person using this approach can talk about probability of railway strike the next month or the probability that it will rain in the next 48 hours. Thus a person using subjective approach can assign probabilities to many, event that, from die classical view point, do not have probabilities associated with them. Consider another case where a person states that the probability his firm, will fail is 5 percent. Such a statement has no meaning from the point of view of relative frequency approach. But it has meaning and utility according to subjective approach.

As pointed out earlier, the main defect of this approach is that different persons may assign different probabilities to the same event. This is due to personal prejudices.

It may be pointed out that no single definition of probability is completely satisfactory. If one approach is suitable in some cases, the other approach may be more suitable in other cases.

Suppose you toss a coin. What is the chance of the coin falls with the tail upwards? I think you know the answer to this question. The chance is  $\frac{1}{2}$ . Suppose you cast a die. (dice) A die has six faces. What is die chance that the die will fall with 6 upwards. Again you know the answers. The chance is  $\frac{1}{6}$  Well this chance is probability. We can also see that odds are 1:5 in favour of getting 6 or 5 :1 against our getting 6.

Let us put this common sense knowledge in more concrete and mathematical language.

If an event can happen in  $m$  ways and fails to happen in  $n$  ways and each of these ways is equally likely, the probability of the chance of  $i$  s happening is  $P = \frac{m}{m+n}$  and that of its not

happening for failing to happen is  $q = 1 - p = 1 - \frac{m}{m+n} = \frac{n}{m+n}$ . In other words the chances are  $m$  to  $n$  that even will happen or  $n$  to  $m$  that the event will not happen.

Let us take the case of the casting of a die, with faces having 1, 2, 3, 4, 5 and 6 dots inscribed on them. Let us call the occurrence of in a throw as the event. The probability of any of the faces turning up on a throw, is  $\frac{1}{6}$ .

Hence probability of 6 occurring i.e.  $p$  is  $\frac{1}{6}$  and the probability of its not occurring i.e  $q$  is  $1 - \frac{1}{6} = \frac{5}{6}$ .

This is because the sum of the probability of the event happening ( $p$ ) and of its not happening ( $q$ ) must be unity. Hence  $p + q = 1$  or  $p = 1 - q = 1 - \frac{5}{6} = \frac{1}{6}$ .

The probability that an event will happen is obtained by dividing the number of ways favourable to the happening of the event by the total number of ways in which the event can happen and cannot happen.

You draw a card from a pack of cards. What is the probability, it is a queen of hearts ? .....  $\left(\frac{1}{52} = \frac{1}{13} \times \frac{1}{4}\right)$ . What is the probability that it is a queen ? ....  $\left(\frac{4}{52} = \frac{1}{13}\right)$  What is the probability that it is a

black card ?  $\left(\frac{26}{52} = \frac{1}{2}\right)$ . What is the probability that you will get first Prize in a state lottery in which

there are 10,000,000 tickets ?....  $\frac{1}{10,000,000}$  or almost 0. (Assuming you purchased only one ticket).

In order to find the probability of an event happening, you have just to know the number of ways in which the event can happen ( $m$ ) and the total number of ways in which the event can happen and not happen ( $m+n$ ).

### Permutations and Combinations

In order to find the numbers of ways in which an event can happen or cannot happen, a knowledge of the algebraic concepts of permutations and combinations is required.

Take the letters ABC. The number of ways in which these letters can be arranged by taking all three at one time is called the number of permutations. The permutations are 6 in the case and they are ABC, ACB, BAC, BCA, CAB, CBA.

Although there are different arrangement or permutations, they are one combination of selection.. ABC is not a separate combination or group, from ACB although they are different arrangements.

If we want to know the number of ways in which  $n$  things can be arranged and permuted by taking  $r$  things at a time, the number can be found with the help of a formula.

$${}^n P_r = \frac{n!}{(n-r)!}$$

${}^n P_r$  stands for permutations of  $n$  things taking  $r$  at a time,  $n !$  is called  $n$  factorial (or  $!$  is the factorial sign.)

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \times 2 \times 1$$

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

$$2! = 2 \times 1; 1! = 1; 3! = 1$$

If four letters ABCD are to be arranged :3 at a time:

$${}^4P_3 = \frac{4!}{(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{1} = 24$$

If four letters ABCD are to be arranged four at a time;

$${}^4P_4 = \frac{4!}{0!} = \frac{4 \times 3 \times 2 \times 1}{1} = 24 \quad (\because 0! = 1)$$

Combinations or selections are different, from permutation or arrangements. The number of combination of  $n$  things taken  $r$  at a time is given by the formula.

$${}^nC_r = \frac{{}^nP_r}{r!} = \frac{n!}{(n-r)! r!}$$

If three letters are to be selected out of the four letters. ABCD. What shall be the number of these selections?

$${}^4C_3 = \frac{{}^4P_3}{3!} = \frac{4!}{(4-3)! 3!} = \frac{(4 \cdot 3 \cdot 2 \cdot 1)}{1 \cdot 3 \cdot 2 \cdot 1}$$

**Example 1.** In how many ways can 4 persons be selected put of 7 ? (It is a question on. combinations. The arrangement of the 4 selected persons is not important here).

The required number of ways or combinations is

$${}^7C_4 = \frac{{}^7P_4}{4!} = \frac{7!}{(7-4)! 4!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (4 \times 3 \times 2 \times 1)} = 35$$

**Example 2.** In how many ways can 4 persons sit on 7 chairs ?

Here the order or arrangement of sitting is also important. It is a question or permutations.

$${}^7P_4 = \frac{7!}{(7-4)!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = 840$$

After this brief discussion on permutation and combination let us now come back to the consideration and calculation of probability.

**Example 3.** From a bag containing 5 red and 6 black balls three balls are drawn at random. What is the probability they all three balls drawn Would be red ?

$$\text{We know that } p = \frac{m}{m+n}$$

We have to find values of  $m$  and  $m + n$ ,  $m$  standing for the number of ways favourable to the event i.e. the number of ways in which 3 red balls can be drawn and  $m + n$  for the total number of ways in which three balls can be drawn whether they be red or not, now 3 red can be drawn out of 5 red balls in  ${}^5C_3$  ways.

$$\therefore m = {}^5C_3$$

3 balls can be drawn out of 11 ball's in  ${}^{11}C_3$  ways.

$$\therefore m + n = {}^{11}C_3$$

$$\begin{aligned}\therefore p &= \frac{m}{m+n} = \frac{{}^5C_3}{{}^{11}C_3} = \frac{\frac{3!}{\frac{{}^{11}P_3}{3!}}}{\frac{{}^{11}P_3}{3!}} \times \left( \frac{3!}{{}^{11}P_3} \right) \\ &= \frac{5!}{(5-3)!} \times \frac{(11-3)!}{(11)!} \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} \times \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ &= \frac{2}{33}\end{aligned}$$

### Simple and Compound Events

When we discuss the probability of the happening or not happening of a single event, the event is called a simple event. All the examples, taken so far are examples of simple events.

When two or more events, happen together they are called compound events. If we want to find, the probability of a card drawn from a pack of cards being a queen, the event is simple. If we want to find the probability of the first card being a queen and the second card a king the event is a compound event.

Events can either be independent or dependent If we draw a card from the pack and put it back before drawing the next, the second sub-event is not affected by the first draw.” The events are independent. If we do replace the card the probability of the second draw is influenced by the first draw such events are dependent events, the tossing of queen and throwing of die shall always be independent events. Drawing of balls or cards may or may not be independent events, depending on whether replacement were or were not made after the first draws.

**Probability Theorems:** There are two very important theorems of probability which are applicable in the case of compound events. They are (i) Addition Theorem, (ii) Multiplication Theorem.

**Addition Theorem:** If an event can happen in different ways which are mutually exclusive, the probability that it will happen, is the sum of the probabilities of its happening in these different ways. The theorem is simple and self evident. If you throw a die what is the probability “that in the first throw either 6 comes upwards, or 5 comes upwards ? The two events are mutually exclusive. It is a question of either this, or that the probability is the sum of the separate probabilities and is

$$\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

The addition theorem will hold good only if

- (1) events are mutually exclusive
- (2) mutually exclusive events belong to the same set.

The condition (1) is satisfied if events belong to Either r' category. For condition (2) just read this statement. The probability that a card is a queen is  $\frac{4}{52}$ . The probability that 6 comes upwards, in a throw of a dice is  $\frac{1}{6}$ . What is the probability that either a queen is drawn on 5 comes up? Not  $\frac{4}{52} + \frac{1}{6}$ . The addition theorem cannot be applied. The items do not belong to the same set.

**Multiplication Theorem:** If a compound event is made of a number of separate but not mutually exclusive, sub-event the probability of the occurrence of the compound event is the product of the probabilities of each of the sub-event happening. If a dice is thrown 2 times what is the probability that 6 comes up in the first throw and 5 in the second throw? The probability is  $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$  in multiplication theorem we do not consider the probability of either this or that event, but of this as well as that event.

Suppose there are two events 1 or 2,  $p_1$  and  $p_2$  are the probabilities of their happening and  $q_1$  and  $q_2$  the probabilities of their not happening respectively. Let  $m$  be the number of ways in which, the first each can happen and  $n$  in which it cannot happen. Let  $m^1$  be the number of way in which the second event can happen  $n^1$  in which it cannot happen.

$$p_1 = \frac{m}{m+n} \quad q_1 = \frac{n}{m+n}$$

$$p_2 = \frac{m^1}{m^1+n^1} \quad q_2 = \frac{n^1}{m^1+n^1}$$

The number of ways in which the first as well as the second event can happen:

$$\text{or} \quad p_1 p_2 = \frac{mm^1}{(m+n)(m^1+n^1)}$$

The number of ways in which the first event can happen and the second event can be happen.

$$\text{or} \quad p_1 p_2 = \frac{mn^1}{(m+n)(m^1+n^1)}$$

$$\text{Similarly,} \quad q_1 q_2 = \frac{nm^1}{(m+n)(m^1+n^1)}$$



$$q_1 q_2 = \frac{nn^1}{(m+n)(m+n^1)}$$

The following illustration of drawing a card three times from a pack of cards may Help you, decided when the addition theorem is to be used and when the multiplication theorem. The card is replaced after each draw. What is the probability that the card is a queen in the first draw...  $\frac{4}{52} =$

$\frac{1}{13}$ . What is the probability that it is a king in the second draw ? Again  $\frac{4}{52} = \frac{1}{13}$ . That it is an ace in

the third draw ? Again  $\frac{4}{52} = \frac{1}{13}$ . What is the probability that the card is either 5 queen in the first draw or a king in the second draw. The composite probability should be greater than the separate

probabilities. It is  $\frac{1}{13} + \frac{1}{13}$ . What is the probability that the card is queen in the first draw or a king

in the second draw or an ace in the third draw? It is greater still and is  $\frac{1}{13} + \frac{1}{13} + \frac{1}{13} + \frac{3}{13}$  and so on.

$$\text{Probability of ship A arriving safely} = \frac{2}{2+5} = \frac{2}{7}$$

$$\text{Probability of ship B arriving safely} = \frac{3}{3+7} = \frac{3}{10}$$

$$\text{Probability of ship C arriving safely} = \frac{6}{6+11} = \frac{6}{17}$$

The chance that all the three (A as well as B as well as C) arrived safely

$$= \frac{2}{7} \times \frac{3}{10} \times \frac{6}{17} = \frac{18}{595}.$$

What would be the probability that at least one of these ships arrives safely? Find the solution yourself and then compare with the following:

$$\text{Probability of A not arriving safely} = \frac{5}{7}$$

$$\text{Probability of B not arriving safely} = \frac{7}{10}$$

$$\text{Probability of C not arriving safely} = \frac{11}{17}$$

∴ Probability that none of them arrives safely

$$\frac{5}{7} \times \frac{7}{10} \times \frac{11}{17} = \frac{11}{34}$$

∴ Probability that at least one ship of three arrives safely

$$= 1 - \frac{11}{34} = \frac{23}{34}.$$

Can you imagine another way of arriving at the answer  $\frac{2}{3} \times \frac{3}{4}$  ? (A, B, C) all reach : A, B reach; C does, not; A, C reach, B does not; and so on. Add all such mutually exclusive probabilities. The method is however, lengthy and time consuming).

**Example 6.** The probability that A can solve a problem is  $\frac{2}{3}$  the probability that B can solve it is. If both try, what is the probability that the problem is solved.

(H.P. University Feb. 1972)

The probability that A cannot solve the problem :—  $1 - \frac{2}{3} = \frac{1}{3}$ .

The probability that B cannot solve the problem: —  $1 - \frac{4}{5} = \frac{1}{5}$ .

The probability that both A as well as B cannot solve, the problem (i.e. if A and B both try, the problem, will remain unsolved) =  $\frac{1}{3} \times \frac{1}{5} = \frac{1}{15}$ .

∴ the probability that if both A and B try the problem the problem will be solved (i.e. if A and B both try, the problem will not remain unsolved)

$$1 - \frac{1}{15} = \frac{14}{15} \text{ Ans.}$$

Alternatively the problem could be solved like this—

(i) The probability that both A and B solve the problem  $\frac{2}{3} \times \frac{4}{5} = \frac{8}{15}$ .

(ii) The probability that A solves and B does not solve the Problem  $\frac{2}{3} \times \frac{1}{5} = \frac{2}{15}$ .

(iii) Probability that A does not solve and B solve the problem  $\frac{1}{3} \times \frac{4}{5} = \frac{4}{15}$ .

The problem will be solved if either (i) or (ii) or (iii) above happens (mutually exclusive and events).

∴ the probability the problem will be solved =  $\frac{8}{15} + \frac{2}{15} + \frac{4}{15} = \frac{14}{15}$  Ans.

**Example 7.** Two balls are drawn from a bag containing 9 red and 14 white balls. Find the chance that they are —

- (i) both of the same colour.
- (ii) each of a different colour.

**Solution:**

(i) The requirement is met if both the balls are either red or white.

Two red balls can be drawn out of 9 red balls in  ${}^9C_2$  ways. The total number of balls in the bag is  $9 + 14 = 23$ , and two balls can be drawn out of 23 balls in  ${}^{23}C_2$  ways.

∴ the probability of drawing of 2 red balls is

$$= \frac{{}^9C_2}{{}^{23}C_2} = \frac{\frac{9 \times 8}{2 \times 1}}{\frac{23 \times 22}{2 \times 1}} = \frac{9 \times 8}{23 \times 22} = \frac{36}{253}$$

the probability that both the balls are of the same colour (either both red or both white)

$$\frac{36}{253} + \frac{91}{253} = \frac{127}{253}$$

(ii) The requirement is met if one ball is red and the other is white.

One red ball can be drawn out of red balls in  ${}^9C_1$  or 9 ways.

One white ball can be drawn out of the white balls in  ${}^{14}C_1$  or 14 ways.

∴ total number of ways favourable to the drawing of one red and one white ball.

$$= 9 \times 14 = 126.$$

The total number of ways in which two balls can be drawn out of a total of 23.

$${}^{23}C_2 = \frac{23 \times 22}{2 \times 1} = 253$$

∴ the chance that one ball is red and the other white

$$= \frac{126}{253}.$$

**Binomial Expansion:**

The following are some binomial expansion.

$$(p + q)^1 = p + q$$

$$(p + q)^2 = p^2 + 2pq + q^2$$

$$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$$

You may be remembering these expansions from your high school days. But suppose you were to write down the value of  $(p + q)^{12}$ . You may not be able to expand this binomial so easily. We shall make use of a general formula writing such expansions.

$$(p+q)^n = p^{n+n}c_1 p^{n+1} q^1 + {}^nc_2 p^{n-2} q^2 + {}^nc_3 p^{n-3} q^3 + \dots + q^n$$

$$p^n + np^{n-1} q + \frac{n(n-1)}{1 \times 2} p^{n-2} q^2 + \frac{n(n-1)(n-2)}{1 \times 2 \times 3} p^{n-3} q^3 + \dots + q^n$$

(Compare these terms with the terms, of Newton's forward formula for interpolation).

Now we can write the terms of say  $(p+q)^4$  more easily.

$$(p+q)^4 = p^4 + {}^4C_1 p^3 q + {}^4C_2 p^2 q^2 + {}^4C_3 p q^3 + q^4.$$

$$= p^4 + 4p^3 q + 6p^2 q^2 + 4p q^3 + q^4.$$

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

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$

Such binomial expansions are of immense use in finding certain probabilities. Suppose  $P$  stands for the probability of the happening of an event in a single trial,  $q$  for the probability of the event not happening in a single trial. If two trials are made, the probability that the event will happen both the times, one time and zero time (*i.e.* is not happening at all) are the terms of the expansion  $(p+q)^2$ , *i.e.*  $p^2$ ,  $2pq$  and  $q^2$  respectively. If 4 trials are made, the probability that the event will happen all the 4 times, 3 times, 2 times, 1 time and 0 times are the respective terms of the expansion  $(p+q)^4$ , *i.e.*  $p^4$ ,  ${}^4P_3 q$ ,  $6p^2 q^2$ ,  $6p q^3$ , and  $q^4$ , respectively.

**Example 8:** Let us take specific cases now. If tossing of a coin is called a trial and coming tip of a head in a toss of a coin is called success or an event then the probability of event happening

of the probability of success in a single trial or  $p = \frac{1}{2}$ ,  $q$  then is automatically  $\frac{1}{2}$ . If the coin is tossed 4 times, (*i.e.* 4 trials are made) what is the probability that we shall get heads in all the 4 tosses, in 3 tosses, in 2 tosses, in 1 toss, and in no toss at all? These are given simply by the

successive terms in the expansion of the binomial  $(p+q)^4$  or  $\left(\frac{1}{2} + \frac{1}{2}\right)^4$ . The terms are

$$\frac{1}{16}, \frac{4}{16}, \frac{6}{16}, \frac{4}{16}, \frac{1}{16}.$$

**Example 9 :** Suppose a die is cast and coming upwards of 1 or 2 is called success What would be, the value of  $p$ ?

It will be  $\frac{2}{6}$  or  $\frac{1}{3}$   $q$ , therefore will be  $1 - \frac{1}{3} = \frac{2}{3}$ . Suppose the die is cast 4 time. What is the probability that we shall get success all the 4 times. What is the probability that we shall get success all the 4 times, 3 times, 2 times, 1 time and no success at all ? (Remember that success is coming up of 1 or 2 whose  $p = \frac{1}{3}$ ). The probabilities are the respective terms of  $(p+q)^4$  or  $\left(\frac{1}{3} + \frac{2}{3}\right)^4$ . They are:-

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

$$= \left( \frac{1}{81} \right) + \frac{3}{81} + \frac{34}{31} + \frac{32}{81} + \frac{16}{81}.$$

$$\text{Probability of getting 4 successes} = \frac{1}{16}$$

$$\text{Probability of getting 3 successes} = \frac{8}{81}$$

$$\text{Probability of getting 2 successes} = \frac{22}{81}$$

$$\text{Probability of getting 1 successes} = \frac{32}{81}$$

$$\text{Probability of getting 0 successes} = \frac{16}{81}$$

From specific we move back to the general  $p$  is the probability of an “event” happening or of “success” in a single trial  $q$  has the opposite meaning. The trial is repeated  $n$  times, what are the probabilities of the event happening  $n$  times,  $n-1$  times  $n-2$  times.....2 times, 1 time, 0 times ? They are terms of the expansion of the binomial  $(p+q)^n$ . Let us write these terms. (Plus signs here, are not important)

$$(p+q)^n p^n + {}^nC_1 p^{n-1} + {}^nC_2 p^{n-2} q^2 + \dots \dots {}^nC_{n-2} p^2 q^{n-2} + {}^nC_{n-1} p^3 q^{n-1} + q^n$$

From here, we can write down. In  $n$  trials.

The probability of getting  $n$  successes =  $p^n$

The probability of getting  $n-1$  successes =  ${}^nC_1 p^{n-1} q^1$

The probability of getting  $n-2$  successes =  ${}^nC_2 p^{n-2} q^2$

... ..

... ..

The probability of getting  $r$  success =  ${}^nC_{p-r} p^r q^{n-r}$

The probability of getting successes =  ${}^nC_r p^r q^{n-r}$

The probability of getting 2 successes =  ${}^nC_{n-2} p^2 q^{n-2}$

The probability of getting 1 success =  ${}^nC_{n-1} p^1 q^{n-1}$

The probability of getting 0 success =  ${}^nC_n p^{n-n} q^n = q^n$

Do a little brain twisting to understand how the probability of getting  $r$  success which is  ${}^nC_{n-r} p^r q^{n-r}$  or  ${}^nC_r p^r q^{n-r}$  fits into the above scheme. If you find it is difficult to start from the top, start from the bottom.

Probability of getting 1 success is  ${}^nC_{n-1} p^1 q^{n-1}$ ; of 2 successes is  ${}^nC_{n-2} p^2 q^{n-2}$ ; of  $r$  successes it should be  ${}^nC_{n-r} p^r q^{n-r}$ . Since  ${}^nC_{n-r}$  is the same thing as  ${}^nC_r$   $p^r q^{n-r}$ .

Let us now write a general rule. The probability of the happening of an event in one trial being known, the probability that the event will happen exactly  $r$  times in  $n$  trials is  ${}^nC_r p^r q^{n-r}$  whether  $p$  stands for the probability of its happening and  $q$  for the probability of its not happening in a single trial.

The probability that an event will happen at least  $r$  times in  $n$  trials is  $p^n + {}^nC_1 p^{n-1} q^1 + {}^nC_2 p^{n-2} q^2 + \dots + {}^nC_{n-r} p^r q^{n-r}$ .

Because the probability of an event happening exactly  $n$  times is  $p^n$ , exactly  $n-1$  times is  ${}^nC_1 p^{n-1} q^1$ , exactly  $n-2$  times is  ${}^nC_2 p^{n-2} q^2$  ..... and exactly  $r$  times is  ${}^nC_{n-r} p^r q^{n-r}$  and because all of them are mutually exclusive and in any of them the event happens at least  $r$  times, therefore the probability that the event happens at least  $r$  times is simply the sum of these different probabilities.

**Example 10.** Find the chance of getting exactly 5 heads in 6 throws of an unbiased coin.

(H.P. University Feb. 1972)

**Solution:**

Probability of a head or  $p = \frac{1}{2}$

$$q = \frac{1}{2}$$

Number of trials or  $n = 6$

Number of successes desired or  $r = 5$ .

$\therefore$  the chance of getting exactly 5 heads in 6 throws of the coin  $= {}^nC_r p^r q^{n-r}$ .

$$\begin{aligned} &= {}^6C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^1 \\ &= 6 \times \frac{1}{32} \times \frac{1}{2} = \frac{6}{64} = \frac{3}{32} \text{ Ans.} \end{aligned}$$

**Example 11.** Find the chance of getting (i) at least 5 heads, (ii) at least 4 heads, in six throws of unbiased coin.

**Solution**

(i) If  $p$  is the probability of success in a single trial, the probability of getting at least  $r$  successes in  $n$  trials  $= p^n + {}^nC_1 p^{n-1} q^1 + {}^nC_2 p^{n-2} q^2 + \dots + {}^nC_{n-r} p^r q^{n-r}$  probability of getting a head in one throw of a coin or  $\frac{1}{2}$ .

$$\therefore q = \frac{1}{2}$$

Probability of getting at least 5 heads in 6 throws of a coin

$$\begin{aligned} &= \left(\frac{1}{2}\right)^6 + {}^6C_1 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^1 \\ &= \frac{1}{64} + \frac{6}{64} = \frac{7}{64} . \end{aligned}$$

(ii) Probability of getting at least 4 heads in 6 throw of a coin

$$= \left(\frac{1}{2}\right)^6 + {}^6C_1\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^1 + {}^6C_2\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^2$$

$$= \frac{1}{64} + \frac{6}{64} + \frac{15}{64} = \frac{22}{64} = \frac{11}{32}$$

The problem could have been solved as follows also :

(i) Probability of getting 6 heads:  $= \left(\frac{1}{2}\right)^6 = \frac{1}{64}$ .

Probability of getting 5 heads  $= {}^6C_5 p^5 q^1 = \frac{6}{64} \left(\frac{1}{2}\right)^1$

$\therefore$  Probability of getting 5 heads.

$$= \frac{1}{64} + \frac{6}{64} = \frac{7}{64}$$

Similarly, for the (ii) part.

Let us now use the knowledge of probability to study Binomial and Normal Distributions. As we shall see, these distributions are indispensable for analysis and interpretation of data and are the foundation, of sampling methods.

In previous lessons we have dealt with many frequency distributions. Those distributions were based on observations or experiments. In binomial and normal distributions we start on certain assumptions and then try to calculate different frequencies. Such distributions which are not based on actual observations or experiments, but are calculated mathematically on the basis of certain assumptions, are called “Theoretical Frequency Distributions.”

We are studying only two binomial and normal distributions. A third important distribution. Poisson distribution is not being discussed here.

### The Binomial Distribution

Suppose two coins,  $a, b$  are tossed simultaneously. There are 4 possible ways in which they could fall:

Possible outcomes on coins			Probability			
	a	b				
(1)	H	H	$= p^2$	$= p^2$	$= \frac{1}{2}^2$	$= \frac{1}{4}$
(2)	T	H	$= pq$	$= {}^2pq$	$= 2 \times \frac{1}{2} \times \frac{1}{2}$	$= \frac{2}{4}$
(3)	H	T	$= pq$			
(4)	T	T	$= q^2$	$= q^2$	$= \left(\frac{1}{2}\right)^2$	$= \frac{1}{4}$

Here H stands head, T for tail,  $p$  stands for the probability of head falling upwards, and  $q$  stands for the probability of head not falling upwards. The first outcome is that both heads fall upwards. Since this is only one of the 4 possible ways in which the coins could fall, probability of the first outcome is one out of four, or  $\frac{1}{4}$ . Second and third outcomes represent 1 head and 1 tail, the probability of one head and one tail is, therefore two out of 4 or  $\frac{2}{4}$ . Fourth outcome is both tails; the probability of both tails is one out of four or  $\frac{1}{4}$ .

Here the probabilities of 2 heads, 1 head and 1 tail, 2 tails have been written, as  $p^2, 2pq + q^2$ . These terms are simply the expansion of  $(p + q)^2$

In this illustration,  $p = q = \frac{1}{2}$ . That is why the probabilities of various outcomes are

$$\left(\frac{1}{2} + \frac{1}{2}\right)^2 = \frac{1}{4} + \frac{2}{4} + \frac{1}{4}.$$

Similarly if we toss three coins simultaneously, there shall be 8 possible outcomes. The probabilities of 3 heads, 2 heads, 1 head, 0 head (i.e. 3 tails) can be found simply by binomial expansion. But let us write again all the 8 possible outcomes to satisfy ourselves that the binomial, expansion really gives correct results.

Possible Outcomes on coins				Probability
	a	b	c	
(1)	H	H	H	$= p^2 = p^3$
(2)	T	H	T	$= p^2q$
(3)	H	T	H	$= p^2q = 3p^2q$
(4)	T	T	H	$= pq^2$
(5)	T	T	H	$= pq^2$
(6)	T	T	H	$= pq^2 = 3pq^2$
(7)	H	T	T	$= pq^2$
(8)	T	T	T	$= q^3 = q^3$

Probabilities of 3 heads, 2 heads; 1 head and 0 head are given by the successive terms of the binomial expansion  $(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$ . Here  $p = q = \frac{1}{2}$ .

Therefore, probabilities of 3, 2, 1 and 0 heads are respectively

$$\frac{1}{8} \left[ = \left(\frac{1}{2}\right)^3 \right] \frac{3}{8} = \left[ 3 \left(\frac{1}{2}\right)^3 \times \frac{1}{2} \right] \frac{3}{8} = \left[ 3 \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^2 \right] \text{ and } \frac{1}{8} = \left[ \left(\frac{1}{2}\right)^3 \right].$$



We can see easily that these results are correct. Out of the 8 possible outcomes, there is only one way in which three heads fall upwards. The probability is only one out of eight or  $\frac{1}{8}$ . There are three outcomes where coins fall with, two heads upwards. Therefore, the chance of two heads is three out of eight or  $\frac{3}{8}$ . And so on.

By the way expansions of type  $(p+q)^n$  are called binomial because there are two terms  $p$  and  $q$  which are to be expanded and bi means two bi-cycle, bilateral, binoculars. If several terms were involved the expansion would have been called multinomial.

Suppose the experiment of tossing three coins simultaneously is repeated 200 times. How many out of the 200 experiments or tosses or trials can we expect 3 heads, 2 heads, 1 head and, 0 head ? This shall be given by the successive terms of  $200 \left( \frac{1}{2} + \frac{1}{2} \right)^3$ .

$$= 200 \left( \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} \right).$$

$$= 25 + 75 + 75 + 25.$$

The probable frequencies of heads, 2 heads, 1 head and 0 head are 25, 75, 75 and 25 respectively.

One thing should be clear now. If we want to find the probable frequencies of various outcomes in a given number of experiments or trials, we can use the expression.

$$N(p+q)^n$$

where capital N stands for the number of times the experiment was repeated and  $n$  for the number of independent events.

**Example 1:** Three dice are thrown 27 times. If coming upwards of 3 or 4 is considered to be a success, find the expected frequencies of 3 successes, 2 successes and 0 success.

3 or 4 is considered to be success.

$$\therefore p \text{ or probability of success} = \frac{2}{6} = \frac{1}{3}$$

$$q = 1 - \left( \frac{1}{3} \right) = \frac{2}{3}$$

The required expected frequencies are given by  $N(p+q)^n$

$$\text{or} \quad 27 \left( \frac{1}{6} + \frac{2}{3} \right)^3$$

$$= 27 \left[ \left( \frac{1}{3} \right)^3 + 3 \times \left( \frac{1}{3} \right)^2 \times \frac{2}{3} + 3 \times \frac{1}{3} \times \left( \frac{2}{3} \right)^2 + \left( \frac{2}{3} \right)^3 \right]$$

$$= 27 \left( \frac{1}{27} + \frac{6}{27} + \frac{12}{27} + \frac{8}{27} \right)$$

$$= 1 + 6 + 12 + 8$$

Therefore, once out of the 27 trials we can expect all the three successes, 6 times 2 successes, 12 times 1 success, 8 times no-success.

### **Suggested Readings**

1. Emroy, C.W., : Business Research Methods, Richard D. Irwin Inc: Homewood 1976.
2. Plan, D.R. and EB. Oppermann : Business and Economic Statistics, Business "Publications Inc : Piano, 1986.

\*\*\*\*\*

## LESSON-10

### MATHEMATICAL EXPECTATION

**Dear Student,**

You have already studied Probability; and Theoretical distributions (Normal and Binomial distributions). Elementary knowledge of these two topics particularly, that of the Normal distribution will be frequently needed to understand this topic. Besides we should also be conversant with the concept of Mathematical expectation about which you have not been told so far. So before we switch over to the main topic of Estimation we will talk a little about Expectation first.

#### 1. Univariate Probability Distribution

If the variables  $X$  assumes  $X_1, X_2, \dots, X_n$  (denoted by  $X_i$ , for  $i = 1, 2, \dots, N$ ) values with probabilities  $p_1, p_2, \dots, p_n$  (denoted by  $p_i$  for  $i = 1, 2, \dots, N$ ) respectively, corresponding to the  $N$  exhaustive and mutually exclusive cases the  $X$  is called a Chance or Random variable and the set of values  $X_i$  together with their probability  $p$  constitutes what is called. Univariate Probability, distribution of the variable. For example if a fair die is cast and  $X$ .....denotes the number of the die, then Probability Distribution for the variable  $X$  can be given as:

$X$	:	1	2	3	4	5	6	
$P$	:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	..... (i)

Similarly, if in a throw of a pair of fair, dice and  $X$  denotes the sum of the number on two dice the Probability. Distributions will be:

$X$	:	2	3	4	5	6	7	8	9	10	11	12
$P$	:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

(ii)

#### 2. Mathematical Expectation:

Let  $\phi(x)$  be a function of  $X$  such that it takes values  $X_i$  [ $i = 1, 2, \dots, N$ ] i.e.  $\phi(X_1), \phi(X_2), \dots, \phi(X_n)$  when  $X$  takes values  $X_i$  ( $i = 1, 2, \dots, N$ ) i.e.,  $X_1, X_2, \dots, X_n$  with probabilities  $P_1 = 1: 2, \dots, N$  i.e.  $P_1, P_2, \dots, P_n$ , the EXPECTED OR PROBABLE value of  $\phi(x)$  denoted as  $E(\phi\{x\})$  is defined as

$$\begin{aligned} \{ \phi(x) \} &= \sum P_i \phi(X_i) \\ &= P_1 \phi(X_1) + P_2 \phi(X_2) + \dots + P_n \phi(X_n) \end{aligned}$$

Where  $\sum P_1, P_2, \dots, P_n = 1$

(Since  $X_i, i = 1, 2, \dots, N$ )

are mutually exclusive and exhaustive cases. The expected values of  $X$  in cases (i) and (ii) above,

can be given as  $E(X) = \sum p_i X_i = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + \dots + 6 \times \frac{1}{6} = \frac{7}{2}$  and  $E(X) = \sum p_i X_i = 2$

$$\frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \dots + 12 \times \frac{1}{36} = 7 \text{ resp.}$$

### 3. Parameter and Statistic:

Any function of observation in the population is defined as parameter and that in the sample as statistic.

### 4. Estimation of Parameters:

**4.1 Problem of Estimation:** Let the population under investigation have the density function (probability distribution  $f(x)$  :  $\theta_1; \theta_2; \theta_3; \dots \theta_m$ ) where  $x$  is the variable and  $\theta_1, \theta_2, \dots, \theta_m$  are  $m$

parameters of the distribution. For example the density function of Normal dist.  $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$  can be written as  $f(x; \mu, \sigma)$ .

Suppose that a sample of observation  $x_1, \dots, x_2, \dots, x_n$  is available. The estimation problem is to define estimator's for one or more of the parameter  $\theta_1; \theta_2; \theta_3, \dots, \theta_m$  as function's of the sample observations  $x_1, \dots, x_2, \dots, x_3, \dots, x_n$ .

It is customary to represent the as  $\theta_1; \theta_2; \dots, \theta_m$  for the parameter  $\theta_1; \theta_2; \dots, \theta_m$  respectively. For one parameter there can be number of estimators. The main part of the probe of estimation lies in selecting one estimator out of many of the parameter and under consideration in such a manner that its distribution is concentrated around the true value of the parameter  $\theta$  as closely as possible. Let us try to understand this concept clearly. Suppose, the population consists of  $N$  observation  $X_1, X_2, \dots, X_n$ . If we want to draw a sample of size  $n$  from this population to estimate the value; of  $x$  which is unknown, it can be done in  $N$  different ways.

In other words there will be  $N$  different sample of size  $n$  each. Suppose  $x = \sum_{i=1}^n x_i$  is defined as an estimator of Parameter  $x$ . Each one of the  $N$  samples will give one estimate for the estimator  $x$ . Let these estimates be represented by  $x_1, x_2, x_3, \dots, x_{N_{cn}}$  for sample Nos. 1, 2,  $\dots, N_{cn}$ , respectively. This forms a distribution of  $x$  values. These distributions should have concentration round about the true value, of  $x$ .

### 4.2 Estimation can be done in two ways: (i) Point Estimation (ii) Interval Estimation.

**(i) Point Estimation:**— As the name indicates, estimates is point or a single number which is calculated as function of the observation in the sample. For example sample mean is a point estimates of the population mean. In case of point estimates it is not possible to indicate, the amount of confidence that may be placed on it.

**(ii) Interval Estimation:**— A single value 'd' or point estimate is quite unlikely, to concede with the true value of the parameter. It is, therefore, considered more appropriate to obtain a range or values or an interval in which the true value of the parameter may be expected to lie with some definite probability or degree, of confidence. This interval is called "Interval Estimate" or "Confidence Intervals" and the probability or degree of confidence is called "Confidence coefficient". Obviously corresponding to different confidence coefficients, there will be different "Confidence intervals" Higher the value of probability or Confidence Coefficient wider will be the confidence interval there will be one confidence coefficient associated with it.

**4.3 Properties of Good Estimators :—** We have already, mentioned in 4.1 that out of a large number of possible estimators for a parameters. We have to select one those distribution is most closely concentrated round about the true value of the parameter. For presence of the characteristic, an estimator is tested by examining the presence of the following properties in it.

- (i) Unbiasedness
- (ii) Efficiency
- (iii) Consistency
- (iv) Sufficiency

Here, we will discuss (i) and (ii) only.

**(i) Unbiasedness**

An estimator  $\hat{\theta}$  for a parameter  $\theta$  is said to be unbiased if its expected value is equal to  $\theta$  i.e. if  $E(\hat{\theta}) = \theta$ . then  $\hat{\theta}$  is an unbiased estimator of  $\theta$ . In other words, if the value of the estimator is calculated for all possible samples and if on the average estimator assumes, the true value of the parameter, then the estimator is said to be unbiased estimator. For example;

Let  $\bar{X} = \sum_{i=1}^n x_i / n$  be the estimator for the population mean where  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$  are the  $N_{cn}$

estimator for  $N_{cn}$  possible samples of size  $n$  of the estimator  $\bar{x}$ . It can be shows that  $\bar{x}_1 + \bar{x}_2 \dots + \bar{x}_n$ .

$$\frac{N_{cn}}{N_{cn}} = \frac{N_{cn}}{\bar{x}}$$

Hence, we say  $\bar{x}$  i.e. sample mean is an unbiased estimate of population mean  $\bar{x}$ . This can be proved mathematically also.

When  $E(\hat{\theta}) = \theta$  is said to be unbiased estimator,

$$\sum_{i=1}^n (x_i - \bar{x}_2)$$

For example  $\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x}_2)$  is a biase

estimator of  $\sigma = \frac{\sum_{i=1}^n (x_i - \bar{x}_2)}{n}$  as  $E(\hat{\sigma})^2 = \sigma^2$ .

but  $= \frac{\sum_{i=1}^n (x_i - \bar{y})^2}{n-1}$  is an unbiased

estimator of  $\sigma^2$  as  $E(s^2) = \sigma^2$ .

If  $E(\hat{\theta}) > \theta$  estimator is said to be positively biased and if  $E(\hat{\theta}) < \theta$ , estimator is said to be negatively biased.

**Efficiency**— If there are two unbiased estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  of the same parameter i.e.  $E(\hat{\theta}_1) = \theta$  and  $E(\hat{\theta}_2) = \theta$ , the comparison between the two is made on the basis of their variations. The estimator with smaller variance is said to be more efficient than the other with higher variance. Thus if  $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$  then  $\hat{\theta}_1$  is said to be more efficient than  $\hat{\theta}_2$  and vice versa,

Efficiency of  $\hat{\theta}_1$  respect to  $\hat{\theta}_2$  is defined as

$$E = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)} \times 100$$

If there are more than one unbiased estimators of a parameter then the one with the smallest variance, is said “most efficient estimator”. If the criteria of goodness of an estimator be efficiency only then the defined most efficient estimator can be named as the best unbiased estimator.

#### 4.4 Interval Estimate for Population mean

A point estimate of a parameter is not very meaningful without some measure of the possible error in the estimate, An estimate  $\hat{\theta}$  of parameter should be accompanied by some interval about possibly of the form  $(\theta-d)$  and  $(\theta+d)$  and together with some measure of assurance that the true parameter  $\theta$  lies within the interval. “A cost accountant for a publishing company may estimate the cost to be  $80 \pm 5$  Rs. per volume with the implication, the correct cost very probably lies between 75 and 85 Rs. per volume.

Suppose, a samples of  $n$  observation  $(x_1, x_2, \dots, x_n)$  is drawn from a normal population with unknown mean, and known standard deviation  $\sigma$ ,  $\bar{x} = \frac{\sum x}{n}$  is a point estimate of population mean. We wish to determine the upper and lower limits which are rather certain to contain the true parameter value between them.

We know the  $y = \frac{x - \mu}{\sigma / \sqrt{n}}$  will normally distributed with mean 0 and S.D. = 1. Thus

$$\text{Prob. } (-1.96 < y < 1.96) = 0.95$$

$$\text{or } P \left( -1.96 < \frac{x - \mu}{\sigma / \sqrt{n}} < 1.96 \right) = 0.95$$

$$\text{or } P \left[ -1.96 \left( \frac{\sigma}{\sqrt{n}} \right) < (x - \mu) < 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) \right] = 0.95$$

$$\text{or } P \left[ -x - 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) - x \right] = 0.95$$

$$\text{or } P \left\{ x - 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < x + 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) \right\} = 0.95$$

$$\text{or } P \left[ x - 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < x + 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) \right] = 0.96$$

Thus the two limits  $(x-1.96)\sigma/\sqrt{n}$  and  $x + 1.96 \sigma/\sqrt{n}$  have been obtained which we may say with 95% certainty to contain the true parameter value, between them, (iii) has to be clearly understood, We mean that if samples of size  $n$  were repeatedly from the population and if the interval  $(x-1.96) [\sigma/\sqrt{n}]$  to  $x+1.96 [\sigma/\sqrt{n}]$  were computed for each sample, then 95% of those intervals, would be expected to contain the true mean. We therefore, have considerable confidence that the interval  $x-1.96 \sigma/\sqrt{n}$  to  $x+1.96 \sigma/\sqrt{n}$  contains the true mean. The measure of confidence is 0.96.

The interval  $x-1.96 \sigma/\sqrt{n}$  to  $x + 1.96 \sigma/\sqrt{n}$  is called a 95% confidence interval, the probability 0.95 in this case called the confidence coefficient. We can obtain interval with any desired degree of confidence less than one confidence interval with confidence coefficient 0.93 and 0.99 can be obtained by replacing 1.96 by 2.33 and 2.50 respectively.

### Example

Find 95% 98% and 99% confidence interval, for  $\mu$  when an example of four observations (1, 2, 3, 4, 016 and 5,6) has been drawn from a normal population, with unknown mean  $\mu$  and known S.D. = 3.

$$x = \frac{12 + 3.4 + 6.6 + 5.5}{4} = 2.7; \alpha = 3$$

This limits of 95% confidence interval are given by  $x - 1.96 \frac{\sigma}{\sqrt{n}}$  and  $x + 1.96 \frac{\sigma}{\sqrt{n}}$ .

$$= 2.7 - 1.96 \frac{3}{\sqrt{4}} \text{ and } 2.7 + 1.96 \frac{3}{\sqrt{4}}.$$

$$= -0.24 \text{ and } 5.64$$

and hence 95% confidence interval can be written (-0.24, 5.64).

Similarly, find out 98% and 99% confidence intervals by replacing 1.96 by 2.33 and 2.58, yourself.

The method described above cannot ordinarily be used to find interval, estimate of the mean of a normal population because  $\sigma_2$  is not ordinarily known in such a case  $\sigma_2$  is estimated by

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

which is an unbiased estimate of  $\sigma_2$  when the sample is small. Then

$$= \frac{\sum x - \mu}{\sqrt{n}}$$

follows the 't' distribution with  $(n-1)$  degree of freedom.

This is evident from the property of normal distribution that the area covered between the points  $-1.96$  and  $+1.96$  and (in terms of Standard Normal Variate) is 0.96 when the total area under normal curve is considered = 1.

In this case, the 95% confidence interval is given by

$$[z - t_{0.05, n-1} (s/\sqrt{n}) x + t_{0.05, n-1} (s/\sqrt{n})]$$

Similarly to find 98 % and 99% confidence intervals we replace  $t_{0.05, n-1}$  by  $t_{0.02, n-1}$  and  $t_{0.01, n-1}$  values respectively.

**Example :** Deduce that for a random sample of 16 values with mean 41.5 inches and the sum of squares of deviations from the mean 135 (inches drawn from a normal population, 95%) confidence limits for the mean of the population are 39.9 and 43.1 inches.

$$n = 16, x = 41.5 \Sigma (x_i - x)^2 = 135$$

$$\sqrt{\left(\frac{\Sigma (xi - x)^2}{n - 1}\right)} = \sqrt{\left(\frac{135}{15}\right)} = \sqrt{9} = 3.$$

t-value at 5% level of significance for 15 d.f. i.e.  $t_{0.05, 15} = 2.13$  (from t-table).

Confidence limit will be

$$x - t_{0.05, 15} \left(\frac{s}{\sqrt{n}}\right) \text{ and } x + t_{0.05, 15} \left(\frac{s}{\sqrt{n}}\right)$$

$$\text{or } 41.5 - (2.13) \left(\frac{3}{\sqrt{16}}\right) \text{ and } 41.5 + (2.13) \left(\frac{3}{\sqrt{16}}\right)$$

$$\text{or } 39.9 \text{ and } 43.1$$

**Note.** When  $n > 30$ , t values can be replaced by 1.96, 2.33 and 2.58 for 95% 98% and 99% conf coefficients respectively. This is so because for large samples, the sampling distribution of mean resembles the normal distribution.

#### 4.5 Confidence intervals for Proportions:

Let  $p'$  be the proportion in the population. It's point estimate in the sample can be given by the sample proportion  $P = \frac{x}{n}$  where  $x$  is the number of items possessing the attribute under consideration

and  $n$  is the total number of observation in the sample  $\frac{p - p'}{\sqrt{\left(\frac{p - q}{n}\right)}}$  is approximately normally distributed

with mean 0 and S.D. = 1 for sufficiently large  $n$  ( $q' = 1 - p'$ ). For sufficiently large  $n$  the variance

$\frac{p' - q'}{n}$  can be estimated  $\frac{pq}{n}$  hence  $\frac{p - q'}{\sqrt{\{pq\}/\sqrt{n}}}$  can also be taken as standard normal variate i.e.

$\frac{p - q'}{\sqrt{\{pq\}/\sqrt{n}}}$  is normally distributed with mean 0 and S.D. = 1, so 95% confidence intervals for  $d'$

$$\text{can be given by } \left[ p - 1.96 \sqrt{\left(\frac{pq}{n}\right)}, p + 1.96 \sqrt{\left(\frac{pq}{n}\right)} \right].$$

---

$t_{0.05}$  is the table value of t at 5% level of significance of (n-1) of you should show to consult at table:



Similarly 91% and confidence intervals can be obtained by replacing 1.96 by 2.33 and 2.50, respectively.

**Example:** A sample poll of 100 voters chosen at random from all voters in a given district indicated that 55% of them were in favour of a particular candidate. Find 95% and 99% confidence limits for the proportion of voters in favour of that particular candidate if an unlimited number of voters are allowed to cast their votes.

$$p = \frac{x}{n} = \frac{55}{100} = 0.55$$

$$q = 1 - p = 1 - 0.55 = 0.45$$

$$\angle\left(\frac{pq}{n}\right) = \angle\left(\frac{(0.55)(0.45)}{100}\right)$$

So 95% confidence limits will be

$$\left\{ 0.55 - 1.96 \angle\left(\frac{(0.55)(0.45)}{100}\right), 0.55 + 1.96 \angle\left(\frac{(0.55)(0.45)}{100}\right) \right\}$$

i.e. (0.45 and 0.65)

Similarly you can find out 99% confidence limits by replacing 1.96 by 2.58.

\*\*\*\*\*

## LESSON- 11

### STATISTICAL HYPOTHESIS

Dear Student

You have learnt about the problem of estimation in an earlier section. The text problem under the title of statistical inference is that of the Testing of Hypothesis', let us try to understand what is meant by Hypothesis or more precisely Statistical hypothesis.

#### Statistical Hypothesis:

Statistical hypothesis, in the general form is an assertion about the density function or probability distribution of a random variable. Thus the assertion is an example of Statistical hypothesis. For example, let the variable be height, measurement of students in a class of Students. Then the statement that the height measurements of student in the said class follow normal distribution is an example of statistical hypothesis. There is another form also of statistical hypothesis. Let it for example, be presumed or known that the random variable follows, normal distributions. Then the statement that mean  $\mu$  the normal random variable is 12 is also an example of statistical, hypothesis. In practice, it is latter type of hypothesis which have to tackled most of the times. Thus, therefore, amounts to saying, that density function will be assumed to be known and hypothesis will consist of an assertion about values of all or a few of the parameter of the density function. Let the density function of  $x_1$  for example be of the form  $\theta(x) = e^{\theta x}$  then the assertion that value of  $\theta$  is 2, is a statistical hypothesis.

By nature, a statistical hypothesis can be either simple or composite. If a hypothesis specifies the values of all the parameters of a density function completely. It is called "Simple hypothesis". Otherwise it is called. Composite hypothesis", i.e. statistical hypothesis which is not simple is called a "Composite hypothesis."

Let the density function be

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

If the hypothesis  $\mu = 10$  and  $\sigma = 2$  then this is a simple hypothesis. If, however, the hypothesis is  $\mu = 10$  and  $\sigma = 2$  or  $\mu = 10$  and  $\sigma > 2$  or  $\mu = 10$   $\sigma = 2$  then the hypothesis will be composite hypothesis, because the value of  $\alpha$  has not been completely specified.

#### 2. Testing a statistic hypothesis, Null hypothesis, Alternative Hypothesis and Critical region

The procedure for deciding whether to accept or reject the hypothesis is called  $\mu$ . Testing a Statistical hypothesis". As per definition, the statistician has unlimited freedom in developing this procedure i.e. designing the test, naturally, he will be guided by its properties. For example let us take (i) as the density function of the variable X, i.e. let the density function be  $f(x) = \theta e$ . Let it be further supported that  $\theta$  can take only two values, either 2 or 1. The statistician who is working on the problem, either with his intuition, on the basis his experience or on account of some information available on the related factors favours the value  $\theta = 2$ . He therefore assumes that  $\theta$  has the value 2. The assumption is the statistical hypothesis to be tested. Let this hypothesis be denoted as  $H_0$ . This is known as, "Null hypothesis".

According to Prof. R.A. Fisher, null hypothesis is the hypothesis which the tested for possible rejection under the assumption that is true. In other words, testing the null hypothesis involves testing whether the observed values are significantly different from the expected values. The null hypothesis may be rejected or accepted with some specified probability depending upon the outcome of the test applied on it. When testing the null hypothesis, the position of a statistician is on the null, inclined neither to accept it nor to reject it, but allowing full say of observational facts and of the testing procedure adopted, in taking decision either way decision about it.

A possible or acceptable hypothesis alternative to the null in many instance, we formulate a statistical hypothesis for the sole purpose of rejecting or mortifying it. Similarly, if we want to decided whether one procedure is better than another, we formulate the hypothesis that there is no difference between the procedure (i.e. any observed differences are merely due to fluctuation in sampling from the sample population). Such hypothesis are often called null hypothesis and are denoted by  $H_0$ . Any hypothesis which differ from a given hypothesis, is called “Alternative hypothesis” and is denoted by  $H_1$ .

So in the problem under consideration  $\theta = 1$  is our Alternative hypothesis i.e.  $H_1 : \theta = 1$ . Thus our problem is not that of testing  $H_0$ , against  $H_1$  where.

$$\begin{aligned} H_0 &= \theta = 2 \\ \text{and} \quad H_1 &= \theta = 2 \end{aligned}$$

Now, to test  $H_0$ , sample of observations on the random variable  $X$ , will be taken. To avoid the complications, and to make it easier to understand the phenomenon involved in testing, let the sample consist of one observation only. In real life problems, one usually takes several observations but here we are considering only one observation for the reason mentioned above, the philosophy of the testing procedure remains the same. Now, the decision whether to accept the hypothesis or reject it will be made on the basis of the value of the random variable  $X$  obtained. This is quite obvious that rejection of  $H_0$  implies acceptance of  $H_1$  and vice-versa. So the problem then is to determine those values of  $X$  which will correspond to rejection of  $H_0$  and those which will correspond to acceptance of  $H_0$ . It is obvious that if the choice has been made of the values of  $X$  that will correspond to rejection  $H_0$ , then the remaining values will necessarily correspond to acceptance of  $H_0$ .

Aggregate of all such values which correspond to the rejection of  $H_0$  is called the Critical region of the test. All other values of  $X$  which correspond to acceptance of  $H_0$ , constitute what is known as “Acceptance Region”, or Non-critical Region’.

In the problem under consideration let us suppose that the variable  $X$  can take values on the positive half of  $X$ -axis only i.e.  $(0 \leq x < \infty)$ . Every positive out some of  $X$  can be represented by a point on the positive half of the  $X$  axis with its  $x$  coordinate giving the value of the associated random Variable  $X$ . The problem of constructing a test for  $H_0$ , is therefore, the problem of choosing a critical reason on the positive half of the  $X$ -axis. Suppose the statistician arbitrarily chooses, the part of the  $X$ -axis to the right of  $X = 1$  as the critical region. To decide whether this was a wise choice, its consequences will have to be considered fully well.

### 3. Type I and Type II Errors and their sizes :

Since we are basing our decision of accepting or rejecting the hypothesis on the sample observations only and the critical region has been chosen arbitrarily, there is always the likelihood, however small it may be of committing an error in decision making. There, existing only two possibilities with regard to  $H_0$ ; either it is true or it is wrong (i.e.  $H_1$  is true). Similarly, there exist two possibilities with regard to the Decision: either accept  $H_0$  or reject  $H_0$  as a result of testing procedure. If  $H_0$  is really true and the observed value of  $X$  exceeds 1,  $H_0$  will be rejected because it has been agreed to reject  $H_0$  when sample observation falls in the Critical region. Obviously, this is an incorrect decision. If this kind of decision is taken, it will amount to committing an error. This kind of error is called “Type I Error”. On the other hand if  $H_0$  is really wrong (i.e.  $H_1$  is true) and the observed value of  $X$  does not exceed 1,  $H_0$  will be accepted. This is also an incorrect decision because we are accepting  $H_0$  while it is wrong. This kind of error is known as Type II Error. In fact, in all, there are four possibilities, two mentioned above lead to incorrect decision and the remaining two viz. accepting  $H_0$  when it is true and rejecting  $H_0$  when it is not true lead to the correct decision. These four possibilities have been displayed in the table given below:

**Table I**

	$H_0$ is true	$H_0$ is wrong ( $H_1$ is true)
$X > 1$ Reject $H_0$	Type I Error (incorrect decision)	Correct decision
$X < 1$ Accept $H_0$	Correct decision	Type II Error (Incorrect decision)

It is just not enough to know the kind of error that may be committing in decision making but it is also necessary to measure them in some way before one can judge, whether or not the choice of critical region was wise. This can be accomplished by using what is known as the size of an error as the measure of its seriousness. The “size of type I error” is the probability of making a type I error which in turn is the probability that the sample point will fall in the critical region  $H_0$  is true. The size of type II error, similarly is the probability of making a type II error which in turn again is the probability that the sample point, will fall in the non-critical or acceptance region when  $H_0$  is wrong i.e.  $H_1$  is true. Size of type I and type II errors are denoted by  $\alpha$  and  $\beta$  respectively.

$$\alpha = \int_1^{\infty} 2e^{-x} dx + 135 \int_{\text{hence we put } \theta=2 \text{ here}}^{\theta-2} \quad \text{if } H_0 \text{ is true}$$

$$\beta = \int_0^1 e^{-x} dx + 632 \int_{\text{hence we put } \theta=1 \text{ here}}^{\theta-2} \quad \text{if } H_0 \text{ is true}$$

For the problem under consideration, these will be as follows:

Now, in terms of the two types of errors, it is possible to introduce a simple principle to be followed in the process of determination of a good test of hypothesis from amongst many that may exist. Many principles can be suggested. For example, one may think of minimizing the sum of two types of errors or the product of the two errors or any other desirable function of two errors. However, among all possible alternatives, the principle : *“Among all test-procedures possessing the same size type I error choose one for which the size of the type II error is as small as possible”* has been found to be the best. Size of the type II error usually increases if the size of type I error is decreased and hence one cannot think of making type I error as small as desired without paying for an increasingly large type II error. In real life experiments, it is often necessary to adjust the type I error until a satisfactory balance has been reached between the size of two errors.

### **Level of Significance:**

In testing a given hypothesis, the maximum probability with which, one is willing to risk the type I error is called the level of significance of the test. Thus the level of significance is the size of the critical region : i.e., the probability assigned to the critical set. The commonly assigned probabilities are 0.05 and 0.01.

An event E is said to be

- (i) significant if under  $H_0$  :  $1 - P(E) < 0.5$

This normal curve of distribution is the most important theoretical distribution in statistical theory. The probability distributions (we shall see the meaning of probability distributions in the next lesson) of most sample statistics closely resemble the normal distribution. The fundamental importance of the normal distribution in statistics arises from the fact that the measures computed from samples usually tend to be normally distributed, whether or not the original data conforms to a normal distribution. “The normal curve of error stands out in the experience of mankind as one of the broadest generalizations of natural philosophy. It serves as the guiding instrument in researches in the physical and social sciences and in medicine, agriculture and engineering. It is an indispensable tool for the analysis and interpretation of the basic data obtained by observation and experiment.

### **Properties of the Normal Curve**

Drawing the graph of a normal curve. Any symmetrical curve is not necessarily a normal curve.

Although every normal curve is a symmetrical curve.

- (i) The arithmetic average, median and mode in a normal curve coincide. This holds in any bell shaped symmetrical distribution. They lie at the point where the curve has the maximum height. Draw a normal curve and mark the point where Md and Mo lie.
- (ii) The first and third quartiles are equidistant from the median. Similarly, third and seventh deciles, twentieth and, eightieth percentiles, etc. are equidistant from the median. This relationship also holds good, in all symmetrical curves. You know this from your knowledge of skewness.

- (iii) Mean deviation is 7979 or about  $\frac{1}{5}$  th of standard deviation. This relationship is not needed in our, subsequent studies, but it should be given if a question is asked on the characteristics of a normal curve.
- (iv) Semi inter-quartile range or quartile deviation is equal to the probable error and probable error is 6747 or approximately  $\frac{2}{3}$  rd of the standard deviation.
- (v) The normal curve is asymptotic to the x-axis. Understand clearly what is meant by “asymptotic” here. This means that the normal curve continues to approach the base line but never reaches or touches it. It can go to any distance on either side of the mean point, but it will not touch the x-axis.
- (vi) The points of inflection of the curve lie at a distance of one standard deviation on either side of the mean. Ordinate points of inflection are the points where the curvature of the curve changes its direction.
- (vii) This characteristic refers to the relationship of the ordinates of height of the curve to the height of the mean ordinate. The ordinate or the vertical length of the curve at the mean (or Md or  $M\sigma$ ) called the mean ordinate is the highest, ordinate. The height of fee ordinate at a distance of one standard deviation ( $\sigma$ ) from the mean is 60.653% of the height of the mean ordinate. Heights of other ordinates at various distance from the mean are also in fixed relationship with the height of the mean ordinate.  
In your graph of the normal curve measure a distance of one standard deviation from the mean. Draw a vertical line from here to meet the curve. This vertical distance is 60.653% of the ordinate at the mean.
- (viii) Area relationship is the most important relationship in a normal curve. Most of the sampling theory is founded on this area relationship.

The area of the curve-covered between the mean ordinate and an ordinate at  $\sigma$  (standard deviation) distance from the mean always has a fixed relationship with the total areal of the curve. Thus, the area enclosed between mean ordinate and the ordinate at a distance of one  $\sigma$  from the mean is always 34.136% of the total area of the curve. Thus, if you draw two coordinates each at a distance of one  $\sigma$  on either side of the mean the area enclosed between them would be 68.272% approx = 68.267%. This will always be true whether we are drawing one normal curve or another.

The following figure shows the area relationship in a normal curve.  
Area relationship in a normal c

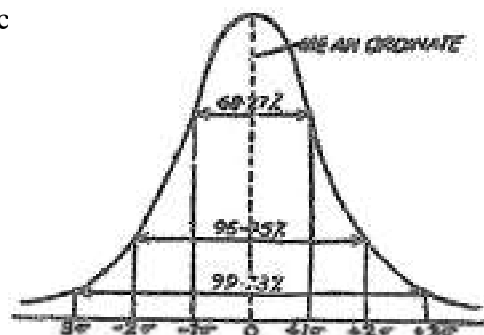


Fig.1

The above figure shows the area enclosed by ordinates at 0;  $\sigma$  and 3  $\sigma$  distances from the mean ordinate. The following table shows the area relationship in a normal curve in more details.

Similarly, the area enclosed between two ordinates at 2 $\sigma$  distances on both, sides of mean is 95.45% of the total of the curve between two ordinates at 3  $\sigma$  distances on both sides of mean is 99.73% of the total area, of the curve.

These relationships are available in the form of printed tables. Five area relationships, however are of special importance in sampling and should be remembered by all.

Distance on both sides of Mean ordinate	Percentage of total area enclosed
$\pm 1.00\sigma$	68/27
$\pm 1.96\sigma$	95.00
$\pm 2.00\sigma$	95.45
$\pm 2.58\sigma$	99.00
$\pm 3.00\sigma$	99.73

Thus, if we draw ordinates at 2.58 $\sigma$  distances on both sides of mean, they are covered between these ordinates will be 99% of the total area, i.e. only 1% of the area shall be left out of these ordinates.

Various tests of significance are constructed on the basis of taking 95% 99%, or 99.73 of the area into account.

#### **Equation of the normal curve :**

The density function of a normal curve is given as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \left( \frac{N - \mu}{\sigma} \right)^2}$$

$$x = X - \mu = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \frac{\sigma^2}{\sigma^2}}$$

Height of ordinate on any given point on  $x$ -axis for a normal curve can be written as follows.

$$y = \frac{Ni}{\sigma\sqrt{2\pi}} e^{-x^2 / 2\sigma^2}$$

Where  $y$  is the ordinate  $e$  is a mathematical constant having a value of 2.71828,  $\sigma$  is the standard deviation, and  $x$  is a given value of the independent variable expressed as deviation from the mean.  $N$  stands for total number of cases.

The maximum ordinate or mean ordinate can be derived from the equation.

$$y_0 = \frac{Ni}{\sigma\sqrt{2\pi}}$$

where  $N$  is the total number of items in the simple,  $i$  is the class interval,  $\pi$  (called pie) is a constant

having a value  $\frac{22}{7} = 3.1416$ .

$$\angle (2\pi) = 2.5066.$$

$$\therefore y_0 = \frac{Ni}{2.5066\sigma}$$

The equation of the normal curve can thus be written as:

$$y = \frac{Ni}{2.5066\sigma} \times 2.71828^{-x^2/2\sigma^2}$$

The theoretical frequencies can thus be found with the help of this curve.

But calculation of theoretical frequencies like this is neither necessary nor advisable.

**Example 1.** Fit a normal curve of the following frequency distribution relating to the height of certain children,

Heights (Inches)	No. of Children
40.5—42.2	1
42.2—44.5	4
44.5—46.5	2
46.5—48.5	18
48.5—50.5	14
50.5—52.5	23
54.5—56.5	10
56.5—58.5	7
58.5—60.5	10
60.5—62.5	3
62.5—64.5	0
64.4—55.5	1
<b>Total</b>	<b>106</b>

in the above frequency distribution, number of frequencies is 106 and class interval 2. The standard deviation is 4.7 (you can calculate S.D. yourself).

Height of the mean ordinate,

$$y_0 = \frac{Ni}{2.5066\sigma} \quad N = 106$$

$$y_0 = \frac{106 \times 2}{2.5066 \times 4.7} \times \frac{212}{11.78102} \quad \sigma = 4.7$$

$$i = 2$$

= 17.993 or approximately 18. Thus the height of the mean ordinate is 18.

The height of the ordinate at one  $\sigma$  distance from the mean would be

$$= \frac{Ni}{2.5066\sigma} 2.71828 - \frac{(4.7)^2}{2(4.7)^2} \quad \frac{(4.7)^2}{(4.7)^2}$$

150



$$= \frac{106x^2}{2.5066 \times 4.7} 2.71828^{-\frac{1}{2}} \quad \text{cancels out}$$

$$= \frac{106x^2}{2.5066 \times 4.7} \frac{1}{(2.71828^{-1/2})}$$

$$\left[ \text{as } e^{-1/2} \text{ can be written as } \left( \frac{1}{e^{1/2}} \right) \right]$$

$$= 18 \times \frac{1}{\sqrt{271823}} = 18 \times \frac{1}{1.6489}$$

$$= 10.9175.$$

Thus the height of the ordinate at one  $\sigma$  distance from the mean on either side would be 10.9175. Similarly, the heights of other ordinates can be calculated and these points can be plotted to obtain, a normal curve.

Alternatively, the height of ordinate at  $\sigma$  distance can be obtained by multiplying the height of mean ordinate by .60653. In this case.

$$18 \times 0.60653 = 10.91754.$$

### Suggested Readings

1. Mason, R.D. : Statistical Techniques in Business and Economics, Richard D. Irwin, Inc: Homewood, 1986
2. Plano, D.R. and E.B. Oppermann : Business and Economic Statistics, Business Publications, Inc Plano 1987.

\*\*\*\*\*

## LESSON-12

### NON-PARAMETRIC TEST : THE SIGN TEST, RANK SUM TEST, THE MANN-WHINEY U TEST, ADVANTAGES AND LIMITATIONS

Dear Student,

Most of the tests which we have considered in earlier lessons have been based on the assumption that parent population is normal. Even if the parent population is not normal, we can often find transformations to reduce it to the normal form. In practice, however, our knowledge, about the parent population may not be sufficient to enable us to find such a transformation and in such case we need tests which do not depend on any assumption about the form of the population. In the last few decades, many new statistical procedures have been developed especially, to take care of the experimental situation in which samples are smalls and the form of population distribution is not normal. We shall consider some of these non parametric or distribution free tests in this lesson.

It is obvious, however that the collection of these results cannot be so comprehensive as in the normal case, but these tests are not only capable of wider application, but also are simpler to apply and do not require complicated sampling theory. Distribution free methods are based on ordered statistics or ordered samples i.e. we suppose that the sample  $x_1, x_2, \dots, x_n$  is ordered so that the observations in it are in ascending order of magnitude. Also while in the parametric case the measure of location and dispersion which are most commonly, used are the mean and standard deviation, respectively which do not depend on order in the sample in the non-parametric case, we prefer to use the median, quartiles, inter-quartile range, etc. for which an ordered sample is desirable.

**Point Estimation and Confidence Intervals.** The population median  $v$  is estimated by the sample median  $\bar{x}$  which however is not an unbiased estimate, though the bias is not serious and tends to zero as  $n$  tends to infinity. Similarly to estimate the population quartiles or deciles we use the corresponding sample quartiles or deciles as estimators.

To construct a confidence interval for  $v$ , we use the fact that the probability of an observation falling to the left or right of  $v$  is one half in either case. The probability that  $x_r$ ,  $r$ th order statistic, exceeds  $v$  is given by

$$P(x_r > v) = \sum_{i=0}^{r-1} (x_{c_i})(1/2)^n \quad \dots (1)$$

$$\text{Similarly } (P \bar{x}_s < v) = \sum_{s=0}^n (x_{c_i})(1/2)^n \quad \dots (2)$$

$$\text{and } (P x_r < v < x_s) = \sum_{i=r}^{s-1} (x_{c_i})(1/2)^n \quad \dots (3)$$

$(x_r, x_s)$  gives the confidence interval for the confidence coefficient given by right hand side, of the third equation. A confidence interval for every confidence coefficient cannot be constructed as R.H.S. of (3) can take only a certain set of values for different values of  $r$  and  $s$ . Though linear interpolation can be done but generally we restrict ourselves to the confidence levels available with simple order statistic. One of the drawbacks of distribution free method is the paucity of confidence

levels for small samples. For moderate samples sizes, we can compute the RHS of (3) either directly or by use of tables of incomplete data functions. For large samples we can use the results that the number of successes will be asymptotically distributed with mean  $(x/2)$  and variance  $(x/4)$

### Test of differences with Correlated Data

#### The Sign Test

One of the simplest test of significance in the non parametric category is the sign test Let us .say that we have parallel set of measurements that are paired off in some way or let  $(x_1, x_2, \dots, x_n)$  and  $(y_1, y_2, \dots, y_n)$  be two random samples of same size  $x$  from two populations and let the sample value be paired by pairing the  $i$ th member of one wife the  $i$ th member of the other and consider the signs of the differences  $(x_i - y_i)$ ,  $(i = 1, 2, \dots, n)$ . If the two populations are continuous and identical the probability of a difference being positive or negative is one half, so we can test the null hypothesis by treating the number of positive signs as a binomial variable with mean  $(n/2)$  and variance  $n/i$ .

In case of ties  $(x_i - y_i)$ , can either ignore them or decide to allot them positive or negative signs by tossing a coin or assign half of them positive sign or other half negative signs. Usually the first choice of ignoring them is preferred and in that case the conditional probability of the sign being positive, given that the difference is non-zero, is one half.

Some of the distribution free or non-parametric, methods have lower power to detect a teal difference as significant. When there is arty choice, a t test is generally more efficient them a sign test and we should prefer a parametric test. But the sign test is much easier to apply and is applicable even when t test is not applicable., viz, when the parent population is not necessarily normal.

#### Example

In Table 12.1, we find, two set of knee-jerk measurements, both from the same men but obtained under, two conditions. In the first case  $(x)$ , the subjects were squeezing a hand dynamometer just before the stimulus struck the knee and in the sex and case  $(y)$  the relaxed kneejerk was obtained in a released sitting posture. The hypothesis to be tested is that they arose from random sampling from the same population. If this hypothesis is true, half the changes from  $x$  to  $y$  should be positive, and half should be negative or we can state null hypothesis that the median change is zero.

Application of the sign test to 1- pairs of knee-jerk data from table 12.

**Table 12.1**

$x$	$y$	Sign of $x-y$
19	14	+
19	19	0
26	30	-
15	7	+
18	13	+
30	20	+
18	17	+
30	29	+
26	18	+
28	21	+
*X = Knee-jerk measurement under tension		
R = measurement under relaxation		

There are 10 pairs of observations; therefore, 10 changes are involved. Since one change is zero and hence cannot be included as either positive, or negative. The hypothesis now calls for 4.5 positive differences, whereas we obtained eight. Is this a significant deviation ?

The null hypothesis is

$$H_0 : P = 1/2$$

$$H_1 : P > 1/2$$

The obvious test to make is based on the binomial distribution for  $P = 5$  and  $N = 9$ . On this basis, 8 or more plus signs could occur by chance 10 times in 512 trials (1 chance in 512 for exactly 9 plus 9 chances for exactly 8).

For a one tail test this deviation is significant with  $p$  equal to approximately 0.02. For a two tailed test we double the probability, which gives a departure, significant at 0.04 level. We would make one tailed test, if the alternate hypothesis at the start were to expect  $x$  values to be higher than  $y$  values.

The assumption involved in making the sign test include mutual independence of the differences. The two parallel set of values may or may not be related. Nothing is assumed regarding the equality of variances. The differences need not even be measured accurately but the direction of each difference should be experimentally established.

### The Sign Rank test of Differences

Let us use an illustration of the sign rank test of differences the same data to which the sign test was applied in Table 12.1. The ten pairs of knee-jerk measurements under tensed and relaxed conditions are repeated for convenience in Table 12.2. Here the numerical differences with algebraic signs, are also listed. As in the sign test, however, we cannot use zero differences, since the differences must be classified according to algebraic sign.

Rank the differences according to size irrespective of their algebraic signs, giving the smallest difference a rank of 1.

**Table 12.2**

X*	Y	X*-Y	Rank of absolute difference	Rank with minority signs
19	14	+5	4.5	-3
19	19	0	-	
26	30	-4	3	
15	7	+8	7.5	
18	13	+5	4.5	
30	20	+10	9	
18	17	+1	1.5	
30	29	+1	1.5	
26	18	+8	7.5	
28	21	+7	6	
*X = knee jerk score under tension X = -3				
Y = score under relaxation				

Two difference of rank 1 are given an average rank of 1.5. The next smallest difference is 4, which is given a rank 3, and so on until all non zero differences ranked.

Now single out all differences whose signs are in minority. If there are fewer negative than positive signs, we select all ranks corresponding to the difference having that sign. There is only one negative difference in Table 12.2. We put this rank with negative sign in the last column. We sum this column to give an statistic T.

The hypothesis test is that the difference are symmetrically distributed about a mean difference of zero. If this is true T would coincide with the mean of much sums of randomly selected ranks  $\bar{T}$  which is also the sum of N successive ranks and is given by the formula

$$\bar{T} = \frac{N(N+1)}{N} \text{ (Mean of sum of ranks)}$$

The obtained T of (-3) (the algebraic sign does not matter in the use of table) is significant at the 0.2 level (a two tail test) when we have a nine differences involved.

For samples larger than 25, a standard deviation and a z ratio can be computed and  $\Sigma$  can be interpreted in terms of the normal distribution. For a sample of size N,

$$\sigma_1 = \frac{\sqrt{N(N+1)(2N+1)}}{24}$$

and  $\Sigma$  is equal to  $(T - \bar{T})/\sigma_1$

### The Run Test

We have to test the null hypothesis that two ordered random samples  $x_1, x_2, \dots, x_n$ , and  $y_1, y_2, \dots, y_n$  come from the same population. The two sets of observations are combined and arranged in order of magnitude as say.

$x_1, y_1, y_2, x_2, x_3, x_4, y_3$  and in this new arrangement we find the total number of runs, where a run is defined as sequence of letters the same kind bounded by letters of the same kind. Thus starts with a run of one x, followed by a run of two y's which is succeeded by a run of three x's and so on.

It is obvious that the two samples are from the same population, x's and y's will be well mixed and r, the number of runs would be large. If the two populations are so widely separated that they do not overlap the number of runs would be only two. There will be a long run of x's at the end if the two populations have the same mean/median, but the first population has a larger variance. Any difference in mean or variance tends to reduce r.

The run test for testing the null hypothesis that the two populations are identical accordingly consists in counting the number of runs in the combined ordered samples and rejecting the null hypothesis if  $s \leq r_0$  where  $r_0$  is number to be determined from the distribution of runs and depends on  $n_1$  and  $n_2$  and the level of significance but not on the form of the distribution of parent population.

The portability of exactly r runs is given by

$$P(r) = \left\{ \begin{array}{l} \frac{2n_1 - 1_{c \ k-1} \times n_2 - 1_{ck-1}}{n_1 + n_{2cn1}} \quad r = 2k \\ \frac{n_1 - 1_{c \ k} \times n_{2-1 \ k-1} + n_{1-1ck} - 1 \times n_{2-1ck}}{n_1 + n_{2cn1}} \quad r = 2k+1 \end{array} \right\}$$

To test the null hypothesis at probability level P, we find  $r_0$  from.

$$\sum_{r=0}^{r_0} p(r) = P$$

as nearly as possible and reject  $H_0$ , if  $r \leq r_0$ .

For large  $n_1, n_2$  the distribution of  $r$  is asymptotically normal with

$$E(r) = \bar{r} = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\text{Var}(r) = \sigma_{r^2} = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 + 1)}$$

and we reject the null hypothesis if  $r < r_1, -1.045 \sigma_n$ . The above approximation can be used if both  $n_1, n_2$  exceed 10.

The run test can also be used for testing the randomness of a give a sample *i.e., to test* whether the sample, observations are independent and identically distributed or in other words to test whether the phenomenon yielding the data is in statistical control.

### The Median Test

The run test discussed above is sensitive to both differences in shape and location of two distributions. If we interested in differences in location only, the median test is to be preferred as it is very little sensitive to differences in shapes.

Let  $x_1, x_2, \dots, x_{n_1}$  and  $y_1, y_2, \dots, y_{n_2}$  be two ordered random samples from the two populations  $f(x)$  and  $f(y)$  respectively and  $z_1, z_2, \dots, z_{n_1+n_2}$  be the combined order, sample of  $z_a$  as its median. Under this null hypothesis, the portability that  $m_1$  or  $r_s$  and  $m_2$  of the  $y$ 's are less than  $z_a$  is given by

$$n_{1cm2} \frac{n_{2cm2}}{n_1 + n_{2ca}} (m_1 - m_2 + 1 = a)$$

where we take  $a$  to be  $(n_1 + n_2) / 2$  if  $n_1, + n_2$  is even and  $(x_1 + x_2 + 1) / 2$  if  $x_1 + x_2$  is odd. If the null hypothesis is true,  $m_1$  has a hypergeometric distribution with mean  $n_{1/2}$  and variance.

**Example :** The Median Test

**Table 12.3**

Application of the median to two samples under conditions A and B.

Samples				
A	B	Contingency table sample		
14	5	10+ 9-		
13	7			
10	6			
12	5			
15	11			
9	8			
9	10			
Md <sub>n</sub> = 9.5				

A common median, has to be calculated of the two samples to carry out median test. The number of cases above and below the common median are to be counted in each sample resulting in a fourfold contingency table. The observations are not paired or correlated and the N may differ in the two-samples. Equal N's would make the test easier to apply. Then chi-square & test can be used with the test statistics s follows, with equal number of observations in each sampler we can conveniently use Table M for a test of significance without computing chi-square.

The median of the 14 observations is 9.5 Values of 10 and above are easily segregated from those of 9 and below, as show in the four-fold table. With such small frequencies, chi-square is not computed.

$$P = \frac{7_{c5} 7_{c5}}{14_{c7}} = \frac{21 \times 21}{3432} = \frac{441}{3432} = 1.12$$

As this probability is greater than 0.05 the level of significance, hence we can accept the null hypothesis, that median is the same for both the populations Median Test with more than two samples.

Suppose that we have three samples, each from its own set of conditions. We want to test the homogeneity of their central values. For example consider the samples in Table 12.4.

**Table 12.4**

Sample			<div>Contingency table</div> <div><div>10+</div><div>9-</div></div>
N	P	K	
2	10	12	
7	7	15	
5	12	9	
6	14	16	
8	9	14	
3	8		
10			
Ni 6	7	5	Mdn = 9.0

The median of all 18 observations is 9.0 we cannot make the point of dichotomy at exactly 9. In such a situation it is made clear that it should be near the median. Let it be point 9.5. We then set up a contingency table as in above table. From these data chi-square is 7.82 with 2df this chi-square is significant near the 0.2 point, We reject the null hypothesis and say that the three medians are not homogeneous.

### Wilcoxon-Mann Whitney U Test

Let two ordered samples combined and ordered in order of magnitude as follows, say

$$x_1, y_1, x_2, x_3, y_2, y_3, \dots$$

For each y we count the number of inversions i.e., the number of x's that precede that y in the sequence. Thus for  $y_1, y_2, y_3$  in the sequence there are 1, 2, 3 inversions. Alternatively for each pair of observations  $x_i, y_i$ , let

$$Z_{ij} = \begin{cases} 1 & \text{if } x < y_j \\ 0 & \text{if } x_i < y_j \end{cases}$$

Then the total number of inversions would be

$$U = \sum_{j=1}^{nz} \sum_{l=1}^{ni} z_{ij}$$

Under the null hypothesis  $Z_{ij}$  is a Berroullian variable with  $p = 1/2$  and

$$E(u) = \frac{n_1 n_2}{2}, \text{ Var } \mu = \frac{n_1 n_2}{2} (n_1 n_2 + 1)$$

also  $n$  is asymptotically normal with, these parameters.

The hypothesis being tested by the Mann-Whitney'  $u$  test takes care of samples of unequal size and the operation through to the finding the sums of the ranks are the same as in the composite rank method When  $N_a$  and  $N_b$  are both as large as 8, a  $\bar{z}$  test can be used and  $\bar{z}$  can be computed by the formula.

$$\bar{z} = \frac{2R_i - N_i(N+1)}{\sqrt{na \, nb \, (N+1)/3}}$$

( $\bar{z}$  value for an obtained sum of ranks of 4 test).

where

$R_i$  = one of the sum of ranks

$N_a$  and  $N_b$  = replication in samples A and B respectively.

$N$  = total number of cases =  $N_a + N_b$ .

$N_i$  = number of cases corresponding to  $R_i$ .

The hypothesis being tested is that one set of measures, as a group, is equal to another. Statistically me  $H_0$  being tested is that the obtained  $u$ , minus the  $u$  to be expected for a particular combination of  $N_a$  and  $N_b$  is zero. The expected  $u$  is equal, to  $na \, nb/2$ ,  $u_i$  is given by the formula.

$$u_i = N_a \, N_b + \frac{N_i(N_i+1)}{2} - R_i$$

(The Mann-Whitney  $U$  statistic).

Deducting the expected  $u$  from the obtained  $u_i$  and multiplying through by  $z$ , the numerator of formula is obtained. The denominator is also 2 times the standard error of  $u$ .

Let as apply formula to a problem with data in Table 12.5.



**Table 12.5**

Measurements		Ranks	
A	B	A	B
14	5	13	1.5
13	7	12	4
10	6	8.5	3
1.2	5	11	1.5
15	11	14	10
9	8	6.5	5
9	10	6.5	8.5
		Σ 71.5	Σ 33.5
		Ra	Rb

In this problem, Na from sample A is 10 and Nb from samples B is 8, All 18 measurements were ranked together.

The sum for sample A ( $R_a$ ) was 123, and that for sample B ( $R_b$ ) was 48. The sum of those two values is 171, which gives us one check, since the total sum of ranks, which is given by  $N(N+1)/2$ , also 171. Using the Small Ri and applying formula.

$$\begin{aligned}\bar{z} &= \frac{2(48) - 8(19)}{\sqrt{(10)(8)(19)/3}} \\ &= \frac{96 - 152}{\sqrt{1520/3}} = -2.49\end{aligned}$$

Assuming a normal distribution for this  $\bar{z}$  the difference between the two set of ranks - and thus between die two set of measurements appears to be significant beyond 0.5 level in a two tailed test. The algebraic sign of  $\bar{z}$  here does not matter.

#### **Advantages and limitations of Non-parametric tests**

When certain assumption of normality are not satisfied, men the desirable properties of the estimators are no longer valid. In such cases the statistical tools to obtain the estimators satisfying the desirable properties are quite complicated and'at times inadequate too. This objective can be achieved pretty easily and quickly too using appropriate non-parametric tests. Non parametric tests are applicable to qualitative data as well as quantitative data. These tests can be applied to small samples and large samples. But the main drawback of these, statistic is that the sensitivity of results gets affected by the choice of initial and terminal observations. One weakness of the Sign test is that it does not use all the available information. If the measurement are on a scale of equal units, on which differences. may be computed for size as well as for direction, the sign test ignores the information provided by the size. Barring small samples, the sign test is only about 60 percent as powerful as at test would be for the same data, where both apply.

#### **Suggested Readings**

1. Siegal, S, and Castellan, N.J. (jr) : Non-Parametric Statistics for Behavioural Sciences, Mc. Graw Hill Book Co. New York, 1988.
2. Kapur, J.N. and Saxena, H.C. : Mathematical Statistics S. Chand & Co. 1969.

\*\*\*\*\*

## LESSON- 13&14

### STANDARD ERROR OF MEAN STUDENT'S 'T' DISTRIBUTION, CHI-SQU ARE TEST

You have already studied Testing of Hypothesis. Hypothesis are set in accordance With the objectives of the problems. For testing, various kinds of hypothesis different tests based on different distributions have been evolved. For example tests based on  $\chi^2$  = distribution have been found suitable and appropriate for solving the problems related with Goodness of fit and Independence of attributes and tests based on Normal distribution for testing if a given sample belongs to a particular population or not and to test the significance of difference between two sample means. There are a number of other tests based on other distributions but we will study only these two here.

#### 1. Test based on $\chi^2$ distribution:

As has been mentioned above that the tests based on  $\chi$  distribution are used mainly to solve the following kinds of problem;

- (i) Goodness of fit
- (ii) Independence of Attributes

First of all we will discuss (1).

#### Testing Goodness of fit Observed Frequencies:

Many a times situation arises when we want to see if an observed frequency distribution follows a particular theoretical distribution or not This is done by examining how well one fits into the other. It is on account of this fact that the test has been named as a test of the goodness of fit. For example, 200 digits may be chosen at random, from a table of random numbers frequencies of the digits may be noted and it may be desired to example 1, here after). Similarly, 12 dice may be considered to have been thrown 4096 times.

[Considering a throw of 4, 5 or 6 as a success it may be desired to test if the dice were unbiased],

As a result, frequency distribution of number of successes obtained is as follows :

Success	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency		7	60	198	430	731	948	847	536	257	11	11	-

(To be referred as example 1, here after) Now if the dice were unbiased then prob. of success =  $p = \frac{1}{2}$  (For each dice) and the frequencies of the distribution can be obtained by the terms in the Binomial expansion.

$$4096 \left( \frac{1}{2} - \frac{1}{4} \right)^{12}$$

Now the problem of testing if the dice were unbiased, reduce to the problem of examining how well do the observed frequencies fit into the expected (hypothetical) frequencies (ii).

The later ones are called expected or hypothetical or frequencies because these have been obtained with the expectation or under the hypothesis that the observed frequencies follows this sort of distribution viz, binomial distribution in the example (ii) given' above.

Now before we proceed for defining  $\chi^2$  test for the problems of the kind mentioned above, it should be clearly understood that this test cannot be applied if (i) the total number of frequencies is not large enough i.e. at least 50 and (ii) the expected frequency of an individual class is very small i.e. less than 5. If these two conditions are fulfilled then  $\chi^2$  with  $n-k$  degrees of freedom is obtained with the help of the formula.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = \frac{\Sigma(O - E)^2}{E} \left( \frac{\text{Simple way of writing}}{\text{writing}} \right)$$

where  $n$  is the effective number of classes

$K$  is the number of constraints

$O$  is the observed frequency of the 7 class.

$h$  is the expected frequency of the 7 class.

In the problems of the kind given above, generally,  $k = 1$  hence the test has  $(n-1)$  degrees of freedom. The calculated value of  $\chi^2$  is compared with the table value\* of  $\chi^2$  for the number of degrees of freedom of the test on specified level of significance.

The most commonly used level of significance by applied statisticians is 0.05. If the calculated value of  $\chi^2$  is greater than the table value of  $\chi^2$ , null hypothesis is rejected, if the calculated value of  $\chi^2$  less than the table value of  $\chi^2$ , null hypothesis is accepted. Null hypothesis is, Fit is good i.e. observed frequencies fit well into the expected frequencies. It is easy to follow that when all observed frequencies are identically equal to the expected frequencies, the calculated value  $\chi^2$  is 0, implying that the fit is the best. Obviously, larger the value of  $\chi^2$  wider the gap between observed and expected frequencies and proper the fit.

Let us now try to understand the whole procedure with the help of the examples given above. Take example first, say the observed frequencies of the digits were :

Digits :	0	1	2	3	4	5	6	7	8	9	Total
Frequencies:	18	19	23	21	16	15	22	20	21	15	200

Under the hypothesis : "digits were equally distributed", expected frequencies for all the

digits should be  $\frac{200}{10} = 20$  each. So we have

										Total	
Ohs. Freq = O :	18	19	23	21	16	15	22	20	21	15	200
Exp. Freq = E :	20	20	20	20	20	20	20	20	20	20	200
(O-E) = :	-1	-1	3	1	-5	5	2	0	1	-5	-5
(O-E) <sup>2</sup> :	4	1	9	1	16	25	4	0	1	25	25
(O-E) <sup>2</sup> :	4	1	9	1	16	25	4	0	1	25	86
---	-	-	-	-	-	-	-	-	-	-	-
E :	20	20	20	20	20	20	20	20	20	20	20

hence  $\chi^2 = \frac{86}{20} = 4.3000$  with  $(10-1) = 9$  d.f.

Table value of  $\chi^2$  for 9 d.f. for 5% L.S (level of significance) = 16.919.

Now, since the calculated value  $\chi^2$  is much lesser than the table value of  $\chi^2 = 0.05$ ,  $9 = 16.919$  the hypothesis is accepted at 5% level of significance *i.e.* digits are equally distributed.

Similarly, in regard-to the example 1 the expected frequencies as follow :

Success	0	1	2	3	4	5	6	7	8	9	10	11	12
Exp. freq.	1	12	66	220	495	792	924	792	495	220	66	22	1

Now, the condition (ii) is not fulfilled as the expected frequencies for success and 12 are less than 5 each and hence the test cannot be applied to this problem as such. To overcome this difficulty, the classes corresponding to successes 0 and 1 and the classes corresponding to successes 11 and 12 may be combined together, to form one class each so that after doing so, no class is left with frequency less than 5 and effective number of classes is not reduced by two. After doing so the data and the solution of the problem will be as follows:

Success	0,1	2	3	4	5	6	7	8	9	10	11,12	Total
Obs. Freq O	7	60	198	430	731	948	847	526	257	71	11	4096
Exp. Freq. E	13	66	220	495	792	924	792	495	220	66	13	4096
O – E	-6	-6	-22	-65	-61	+24	55	41	37	5	-2	
$\frac{(O - E)^2}{E}$	$\frac{37}{13}$	$\frac{36}{66}$	$\frac{484}{220}$	$\frac{4225}{495}$	$\frac{3721}{792}$	$\frac{576}{924}$	$\frac{3025}{792}$	$\frac{1681}{495}$	$\frac{1369}{220}$	$\frac{25}{66}$	$\frac{5}{13}$	33.49

Effective No. of class = 11 (No. of classes after regrouping)

Hence d.f. =  $(11 - 1) = 10$

$\chi^2$  0.05, 10 (table value) = 18.31

$\chi^2$  0.01, 10 (table value) = 23.21

Since calculated value of  $\chi^2$  is greater than the tabulated value even at 1% I.s, the hypothesis is rejected *i.e.*, the data is not consistent with the hypothesis that the dice were unbiased even at 1% level of significance.

### (iii) Independence of Attributes:

We have studied simple Correlation and Regression and we know that the coefficients of Correlation and Regression are meant, for studying the relationship between the two variables. This is possible only if both the variables are of quantitative nature. If even one of the two is of qualitative nature techniques of correlation and digression become non-applicable. Two way table presenting a bi-variate distribution is known as contingency table. In it at least one of the two variables is qualitative Various measures of association have been developed to study the relationship between the two variables in such a case. One of them is based on  $\chi^2$  —distribution which is meant to examine if the two variables or attributes (more appropriate, word to be used for qualitative variables) are

\* you should know how to consult the table.

independent or not. If there exists no mutual relationship of any kind between two attributes A and B then they are said to be independent. If “A” stands for tossing a coin by right hand and “α” for tossing the coin by left hand. “B” i for getting head and “β” for getting tail, then, the two attributes tossing of the coin and the result of the loss will be said to be independent if there is absence of any relationship between them in the sense that the same proportion of heads is observed whether the coin is tossed with right hand or left hand.

Let the following table represent the observed frequencies,

Attribute	B	β	Total
A	a	b	$R_1$
α	c	d	$R_2$
Total	$C_1$	$C_2$	N

Now under the hypothesis that the two attributes are independent i.e. the proportion of heads should be the same whether, the coin is tossed by right hand or left hand, the corresponding expected frequencies, will be given as below :

Expected frequency corresponding to observed, frequency ‘a’ denoted by  $E(a)$  will be given as

$$E(a) = \frac{R_1 \times C_1}{N}$$

Similarly:

$$E(b) = \frac{R_1 \times C_2}{N}$$

$$E(c) = \frac{R_2 \times C_1}{N}$$

$$E(d) = \frac{R_2 \times C_2}{N}$$

Thus the expected frequency corresponding to any cell in the contingency table is obtained, by multiplying the totals, of Row and Column in which the cell falls divided by the total number of frequencies N. Thus in a contingency table when each of the two attributes has two or even more than two classes the expected frequency of the (ij) the cell i.e. the cell corresponding to its *ith* row and *jth* column is obtained as

$$E(i, j) = \frac{R_i \times C_j}{N}$$

where  $R_j$  is the sum of the *ith* row,  
and  $C_j$  is the sum of *jth* column.

Now when we have got observed as well as expected frequencies we can calculate the value of  $\chi^2$  with the help of the same formula viz.

Degrees of freedom of the test are  $(m-1)(n-1)$  where  $m$  is the effective number of rows and  $n$  that of the columns. Rest of the testing procedure remains the same. The null hypothesis (i.e.  $H_0$ ). The two attributes are independent. The two conditions mentioned earlier have to be satisfied here also i.e. (i) the total number of observations should not be less than 50 and (ii) no cell frequency (expected) should be less than 5, In case of a contingency table of order  $m \times n$  ( $m > 2$ ) ( $n > 2$ ) if any expected cell frequency is less than 5, then either two rows or two columns are pooled together, so that the new table has no expected cell frequency less than 5. If there are two or more cells having expected cell frequency less than 5 pooling of both rows and columns may have to be done. Obviously the number of column? or rows or both will, be reduced. The reduced numbers of columns and/or rows are known as the effective number of columns and/or rows. This reduction in the number of columns and/or rows will bring about a reduction 'in the number of degrees of freedom of the test also. This should be done such that the reduction in the number of degrees of freedom is minimum. The case of  $2 \times 2$  table for such a situation will be dealt with later on.

**Example 3.** Discuss on the basis of the following data given in respect of 1000 school boys if the two attributes general ability and mathematical ability are independent or not.

Maths Ability	General ability			Total
	Good	Fair	Poor	
Good	44	22	4	70
Fair	265	257	178	700
Poor	41	41	98	255
Total	350	370	280	1000

Assuming that the general ability and mathematical ability are independent, hypothetical or expected frequencies will be as given below :

	General ability			Total
	Good	Fair	Poor	
$\frac{70 \times 350}{1000} = 24.5$	$\frac{700 \times 370}{1000} = 25.9$	$70 - (24.5 + 25.9) = 19.6$	70	
$\frac{70 \times 350}{1000} = 24.5$	$\frac{700 \times 370}{1000} = 25.9$	$700 - (259 + 245) = 501$	700	
245	259	196		
$350 - (24.5 + 245) = 80.5$	$370 + (25.94 - 259) = 276.84$	$280 - (19.6 + 196) = 64.4$	290	
80.5	75.1	64.4		
350	370	280		1000
$\chi^2 = \frac{(44 - 24.5)^2}{24.5} + \frac{(22 - 25.9)^2}{25.9} + \dots + \frac{(98 - 84.4)^2}{84.4} = 72.1$				

d.f of the test =  $(3-1) (3-1) = 4$ .

Since the calculated value of  $\chi^2$  is much greater than the tabulated of  $\chi^2$ , hypotheses rejected i.e. the two attributes, general ability and mathematical ability are not independent i.e. they are associated.

**Note :** If any of the expected cell frequencies would have been less than 5, two rows or two columns should have been merged together as per discussion given above.

In a  $2 \times 2$  table if the observed frequencies are represented as a, b, c and d as given below :

Observed frequencies in a  $2 \times 2$  table

			Total
	a	b	$R_1$
	c	d	$R_2$
Total	$C_1$	$C_2$	N

then  $\chi^2$  with  $(2-1)(2-1) = 1$  d.f. can be directly obtained by the formula

$$\chi^2 = \frac{(ad - bc)^2 \times N}{R_1 \times R_2 \times C_1 \times C_2}$$

In a  $2 \times 2$  table, if any one of the expected cell frequencies is less than 5 then we have to apply Yate's correction which is as given below ; Calculate the products ad and bc if  $ad > bc$  subtract 1.5 from a and b each and add 0.5 to c and d each and vice-versa.

This will leave  $R_1$ ,  $R_2$ ,  $C_1$ ,  $C_2$  and N unaltered? Now work out the expected frequencies and follow the usual procedure. The D.F. of the test will remain = 1. Directly the value of  $\chi^2$  will be given by the formula

$$\chi^2 = \frac{\left\{ (ad - bc) - \frac{N}{2} \right\}^2 \times N^*}{R_1 R_2 C_1 C_2}$$

with I.d.f.

**Example 4:** Examine if the vaccination has any effect on the possibility of survival in case of human beings on the basis of the data given below.

	Survived	Died	Total
Vaccinated	40	4	40
Non-Vaccinated	50	1	51
Total	90	5	95

Expected frequencies on the assumption that the possibility of survival is independent of vaccination, expected frequencies will be as given below:

---

\*(ad-bc) indicates difference between ad and bc, such that the smaller product is subtracted from the greater i.e. (ad-bc) is always a positive quantity.

	Survived	Died	Total
Vaccinated	$\frac{44 \times 90}{95} = 41.7$	$(44 - 41.7) = 2.3$	44
Non-vaccinated	$90 - 41 = 48.3$	$(5 - 2.3) = 2.7$	51
Total	90	5	95

Since two of the expected frequencies are less than 5 each Yale's correction will have to be applied and accordingly the observed, frequencies will be corrected as:

	Survived	Died
Vaccinated	40.5	3.5
Non-vaccinated	49.5	1.5
	$40 \times 1 = 40$	
	$50 \times 4 = 200$	
	$200 > 40$	

hence 5 added to 40 and 1 each, and 5 subtracted from 50 and 4 each.

Now the  $\chi^2$  can be calculated with the help of the formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

by taking the corrected Observed frequencies. The test will have l.d.f.

Alternatively, the problem can be solved by using the direct formula :

$$\chi^2 = \frac{\left\{ (ad - bc) - \frac{N}{2} \right\}^2 \times N^*}{R_1 R_2 C_1 C_2}$$

$$= \frac{\left\{ (40 \times 1) - (50 \times 4) - \frac{95}{2} \right\}^2 \times N^*}{44 \times 51 \times 90 \times 5} = \frac{\left\{ (200 - 40) - \frac{95}{2} \right\}^2 \times 95}{44 \times 51 \times 90 \times 5}$$

with l.d.f.

**Note :** Calculate  $\chi^2 = \sum \frac{(O - E)^2}{E}$  and examine that it is equal to the value obtained by the formula

$$\chi^2 = \frac{\left\{ (ad - bc) \left( -\frac{N}{2} \right) \right\}^2 \times N^*}{R_1 R_2 C_1 C_2}$$

hint for calculating  $\chi^2 = \sum \frac{(O - E)^2}{E}$



Complete the problem

0	E	O-E	(O-E) <sup>2</sup>	(O-E) <sup>3</sup> /E
40.5	$\frac{44 \times 90}{95}$			
3.5	$\frac{44 \times 5}{95}$			
49.5	$\frac{51 \times 90}{95}$			
1.5	$\frac{51 \times 5}{95}$			

### 13.2 Tests based on Normal distribution

Broadly, there are two kinds of problems which are solved with the help of tests on normal distributions.

- (1) Whether or not a given sample belongs to a specified population with mean  $\mu$ . by testing the significance of difference between population mean  $\mu$  and sample mean  $\bar{X}$ .
- (2) Whether or not means of two samples, are significantly different.

Besides these, there are some tests for populations and percentages also.

#### (i) Testing the significance of the Sample mean $\bar{x}$ .

The distribution of the sample mean  $\bar{x}$  of a simple random sample  $x_1, x_2, \dots, x_n$  approaches to the normal distribution with mean  $\mu$  and variance  $\sigma^2/n$ , (refer to chapter 12, topic “Sampling distribution”.) ( $\mu$  is the population mean and  $\sigma^2$  the population variance and  $n$ , the sample size), as  $n$  becomes increasingly large, even if the parent population from which the sample has been drawn is not normally distributed. So far every population with mean  $\mu$ , and variance  $\sigma^2$  the statistic

$$Z = \frac{|x - \mu|}{\sigma / \sqrt{n}}$$

behaves like a standard normal variate, provided that  $n$  is sufficiently large.

$\sigma^2$  is rarely known. If the sample size is very large, then the sample estimate of variance.

$$S^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

can be taken as an approximation to the population variance without any significant error of approximation and then the statistic  $Z$  takes the form

$$\frac{|x - \mu|}{S / \sqrt{n}}$$

where  $Z$  now also, follows normal distribution with mean 0 and S.D.1.

For testing : null hypothesis is. set  $H_0$  ;  $\bar{x} = \mu$  i.e. sample belongs to the population.  $Z$  value is calculated and compared with 1.96, 2.33 and 2.51 for 5%, 2% and 1% levels of significance, respectively. If the calculated value of  $Z$  exceeds the value (s) specified above, the hypothesis is rejected otherwise it is accepted.

**Example 5.** A sample of 900 members is found to have a mean of 3.4 cms. Could it be reasonably regarded as a sample from a large population whose mean is 3.25 cms and S.D. = 2.61 cms.

Null sample of 900 members may be regarded as a sample from the population with mean = 3.25

\* | | is called Mod, meaning there by that only the absolute value of the figure enclosed is to be considered.

$n = 900$  (large  $\bar{x} = 3.4$  cms.

$\mu = 3.25$  cms and  $\sigma = 2.61$  cms.

$$Z = \frac{|x - \mu|}{\sigma / \sqrt{n}} = \frac{|3.4 - 3.25|}{2.61 / \sqrt{900}} = 1.72$$

$Z$  (calculated)  $1.72 < 1.96$  and hence hypothesis is accepted i.e. sample 900 members may be regarded as a sample from the population with mean 3.5 cms.

**Example 2.** Testing the significance of difference between the means of two samples :

(i) Suppose the two sample of sizes  $n_1$ , and viz.  $x_1, x_2, \dots, x_{n_1}$ , and  $y_1, y_2, \dots, y_{n_2}$  have been drawn from the sample population with standard deviation  $\sigma$ . We want to test whether the difference  $(\bar{x} - \bar{y})$  of their means is significant. Null hypothesis  $H_0: x = y$  i.e. there is no significant difference between the sample means,

$$Z = \frac{|\bar{x} - \bar{y}|}{\sigma \sqrt{\left\{ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right\}}}$$

will be a standard normal variate for sufficiently large  $n_1$  and  $n_2$ .

If  $Z < 1.96$ , we would accept  $H_0: \bar{x} = \bar{y}$  at 5% l.s. If  $1.96 < Z < 2.58$  we would reject  $H_0: \bar{x} = \bar{y}$  at 5% l.s. but will accept at 1% l.s and so on.

(ii) If the two samples have been drawn from two populations having different variance  $\sigma_1^2$  and  $\sigma_2^2$  respectively then our problem may be to examine whether the two populations differ in their means or not, leaving apart the difference in their dispersion.

Let the means and variances of the two populations be  $\mu_1, \sigma_1^2$  and  $\mu_2, \sigma_2^2$  respectively.

Null hypothesis in this case will be  $H_0: \mu_1 = \mu_2$  i.e. two populations do not differ in their means.

$$Z = \frac{|\bar{x} - \bar{y}|}{\sigma \sqrt{\left\{ \left( \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right) \right\}}}$$

In this case will behave like a standard normal variate i.e.

$Z \sim N(0, 1)$ .

Further procedure will remain the same as described above in (i).

(iii) Variances  $\sigma_1^2$  and  $\sigma_2^2$  are known only very rarely in practice and hence are to be replaced by  $S_1^2 = \frac{\Sigma(x - \bar{x})^2}{n_1 - 1}$  and  $S_2^2 = \frac{\Sigma(y - \bar{y})^2}{n - 1}$  estimates of  $\sigma_1^2$  and  $\sigma_2^2$  respectively obtained from sample.

$$\text{Then } Z = \frac{|\bar{x} - \bar{y}|}{\sigma \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \text{ will behave like a standard normal variate provided } n_1, \text{ and } n_2$$

are large enough if parent populations are not normal  $n_1$  and  $n_2$  need not necessarily be large if parent populations are normal.

Further test procedure will remain the same as discussed above in (i).

**Example 6.** The means of the samples of 1000 and 2000 are 67.5 and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of S.D. 2.5 inches.

$H_0 : \bar{x} = \bar{y}$  i.e. two samples have been drawn from the same population  $n_1 = 1000$ ,  $n_2 = 2000$ ;  $\sigma = 2.5$ ,  $\bar{x} = 67.5$ ,  $\bar{y} = 68.0$ .

$$Z = \frac{|\bar{x} - \bar{y}|}{\sigma \sqrt{\left\{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}}} = \frac{|67.5 - 68.0|}{2.5 \sqrt{\left\{\left(\frac{1}{1000} + \frac{1}{2000}\right)\right\}}} = 5.17$$

Now  $Z$  cal. (5.17) > 2.58 and hence hypothesis is rejected even at 1% l.s. i.e. the samples cannot be regarded as drawn from the same populations.

**Example 7 :** A random sample of 1000. men from Shimla shows their mean wage to be Rs. 2.50 per day with a standard deviation of Rs. 1.50. A sample of 1500 men from Chandigarh gives a mean wage of Rs. 2.68 per day with a standard deviation of Rs. 2.0. Discuss the suggestion that the wages vary between Shimla and Chandigarh ?

$H_0 : \bar{x} = \bar{y}$  i.e., the wages do not vary between Shimla and Chandigarh.

$n_1 = 1000, x = 2.50, S_1 = 1.50$

$n_2 = 1500, y = 2.68, S_2 = 2.00$

$$Z \sigma = \frac{|2.50 - 2.58|}{\sqrt{\left\{\frac{(1.50)^2}{1000} + \frac{(2.00)^2}{1500}\right\}}} = 2.57 \text{ and}$$

$$1.90 < 2.57 < 2.58.$$

Thus it may be concluded that at 5%, l.s. we reject the hypothesis i.e. wages vary between Shimla and Chandigarh, but at 1% l.s., we accept the hypothesis i.e., difference between wages of two places is insignificant and existence of variation in wages is not accepted at this level.

**Example 8.** Balls are drawn from a bag containing an equal number of red and white balls, each ball being returned before drawing another. In 2250 drawing 1018 red and 1232 white balls have been drawn. Do you suspect some bias in the part of the drawer ?

**Solution :** Your first job is to ascertain values of  $p$ ,  $q$  and  $n$ . Red and white balls being equal in number, probability of a red ball or  $p = \frac{1}{2}$  ;  $q = \frac{1}{2}$   $n$  the total number of drawing = 2250.

The expected number of red balls in 2250 drawings =  $\frac{1}{2} \times 2250 = 1125$ .

The actual number of red ball = 1018s

The numerical difference between expected and actual frequency =  $1125 - 1018 = 107$ .

The standard deviation of simple sampling =  $\sqrt{npq} = \sqrt{(2250) \times (\frac{1}{2} \times \frac{1}{2})} = 23.7$

The difference 107 is about 4.5 times  $\left(\frac{107}{23.7} \text{ times}\right)$  the  $\sigma$  and there is hardly any probability that arose due is fluctuations of sampling. Almost definitely the drawer is biased against red balls (May be a victim of commission phobia!).

The above solution is sufficient for the examination. But how shall we look at its terms of our normal curve.

The expected number of red balls, 1125 lies at the point of maximum frequency of the normal curve. When sample of 2250 drawings each are taken, actual number of red balls in some sample will be greater than the expected number 1125 and in others less. The actual numbers of red balls in each of the many samples will be spread in a symmetrical normal way around the expected numbers. As usual 99.73% of such actual numbers of red balls will lie within mean  $\pm 3\sigma$ ,  $\sigma$  in this case = 23.7. Therefore 99.73% of sample will be such whose red ball drawing lie within  $1125 \pm 3 \times 23.7$  i.e., within 1054 and 1196.

The number of red ball drawn in Example 8 lies outside these limits. Hence, it is extremely unlikely dial the difference has arisen due to sampling fluctuations.

**Example 9.** A group of scientists reported 1705 sons and 1527 daughters do these figures confirm the hypothesis that the sex ratio is 1 : 1.

**Solution :** Your first job is to write the values of  $p$ ,  $q$  and  $n$ . The hypothesis is that the sex. ratio is 1:1.

$\therefore p$ , the probability of a son =  $\frac{1}{2}$  and  $q = \frac{1}{2}$

$n$ , the number of events =  $1527 + 1705 = 3232$ .

Number of observed male births = 1705.

Number of expected male births.

according to the hypothesis =  $\frac{1}{2} \times 3232 = 1616$ .

The difference between expected and observed number =  $1735 - 1616 = 89$ .

The standard deviation of simple sampling

$$= \sqrt{npq} = \sqrt{\left(3232 \times \frac{1}{2} \times \frac{1}{2}\right)}$$

$$= \sqrt{(808)} = 21.43 \text{ approximately}$$

The observed difference 89 is more than three time (about 3.13 time) this standard deviation, therefore it is unlikely that it arose on account of fluctuations of simple sampling.

Please note again that a categorical statement has been made. It has been stated clearly that probability of the difference having arisen due to sampling fluctuations is extremely small.

The question asked was : “Do these figures-confirm to the hypothesis that the sex ratio is 1 : 1 ? After writing as we have done above, just state in the end that the figures do not confirm to the hypothesis that sex ratio is 1:1.

This problem can also be solved in terms of proportion rather than numbers :

$$\text{The expected male ratio} = \frac{1}{2} = 5.$$

$$\text{The observed male ratio} = \frac{1705}{3232}.$$

The difference between the expected and the observed male ratio.

$$= \frac{1705}{3232} - \frac{1616}{3232} = \frac{8}{3232} = .075$$

The standard deviation of proportion.

$$= \sqrt{pqn} \quad \dots \quad \dots \quad \dots$$

$$\sqrt{\frac{1}{2} \times \frac{1}{2} + \frac{1}{3232}} = 0088.$$

The difference between the observed and the expected proportion is more than three times (3.13 times) the standard deviation of proportion. Therefore, it is improbable that the difference arose on account of sampling fluctuations. The figures as such do not confirm to the hypothesis of 1 :1 sex ratio.

You can reconstruct this problem also in term of a normal curve.

### 13.3 Standard Error :

In sampling problem, standard deviation of simple sampling has to be used again and again. It would be convenient if a shorter name could be used for this quantity. This shorter name is STANDARD ERROR. Do not attach much significance to the word Error here. The use of the word Error is justified here by the fact we usually regard the expected value at the true value and divergences from this excepted value in certain observations and samples are regarded as errors of estimation due to sampling effects.

For our purpose standard error will just mean standard deviation of simple sampling though the term can be used in a slightly wider sense.

From now onwards, we shall use the term Standard Error instead of standard error of simple sampling.

In the solution of problems on sampling we should use  $p$  and  $q$  of the universe rather than of sample. This is what we have done up to now. But sometimes  $p$  and  $q$  of the universe are not known. In such cases  $p$  and  $q$  of the sample are taken as  $p$  and  $q$  of the universe. The assumption is justified if  $n$  is large, say 100 or more, and neither  $p$  nor  $q$  is very small. You will get only those problems where these assumptions hold. Another course open to us when  $p$  and  $q$  of the universe are not known is to take the highest value of  $p \times q$ .

The highest value of  $p \times q = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ . When either  $p$  or  $q$  is found to be rather small we can take  $p = q = \frac{1}{4}$  in order to calculate the standard error.

The question on meaning of standard error and its importance and usefulness in sampling studies is very popular with the examiner. I trust you understand, this question and can answer it. The main use of standard error is to find whether the difference between observed and expected value or between one observed value and another is significant or not. If the difference is more than 3 times the standard error chances are that it could have arisen due to sampling fluctuations.

**Example 10.** 400 children are examined in a town; and 150 are found to be under weights. Assuming conditions of simple, sampling, estimate the percentage of children who are underweight in that town.

Population  $p$  and  $q$  are not known, we shall take sample  $p$  and  $q$  as population  $p$  and  $q$ .

The probability of  $p$ , of getting an underweight children.

$$= \frac{150}{400} = \frac{3}{8} \qquad \therefore q = \frac{5}{8}$$

No. of children examined or  $n = 400$

Standard error of the proportion of under-weight children

$$\sqrt{\left(\frac{pq}{n}\right)} = \sqrt{\left(\frac{3}{8} \times \frac{5}{8} \times \frac{1}{400}\right)} = 0.24 = 2.4\%$$

$$p = \frac{3}{8} = 37.5\%$$

The percentage of children who are under-weight can vary around 37.5% by 3 times the standard error.

Thus the limits, within which the percentage of under-weight children will probably lie are  $37.5 \pm 3 \text{ S.E.}$  or  $37.5 \pm 3 \times 2.4$ , i.e. 30.3 and 44.7 per cent. The probability of his statement, being correct is 99.73% since 99.73 area is covered between mean  $3\sigma$ . If we wanted the probability of our statement to be correct by 99%, the limits would have been  $36.5 \pm 2.5758 \times 2.4$   $37.5 \pm 19.6 \times 2.4$  would give us limits which are true in 95% cases.

### 13.3.1 Standard Error and Size of the Sample:

We know that

$$S.E. = \sqrt{\left(\frac{pq}{n}\right)}$$

The value of S.E. thus depends on  $p$  and  $n$  ( $q$  is automatically covered in  $p$  since  $q$  is always  $(1-p)$ ). The size of the universe is not important in the value of S.E. It is affected by the size of the sample. Given the value  $p$ , greater the value of  $n$ , smaller shall be the value of S.E. and vice versa. The value of S.E. varies inversely as the square root of  $n$ . If  $n$  increase 4 times the value S.E. is

reduced by  $\frac{1}{2}$  (i.e.  $\frac{1}{\sqrt{4}}$ ). If  $n$  remains  $\frac{1}{9}$ th of its former values, of S.E. shall go up, 3 times if  $p = \frac{1}{2}$

$q = \frac{1}{2}$ ,  $n = 100$ ,  $S.E. = \sqrt{\left(\frac{1}{2} \times \frac{1}{2} = \frac{1}{100}\right)} = .05$  or 5% if  $n$  becomes 400.

$$S.E. = \sqrt{\left(\frac{1}{2} \times \frac{1}{2} - \frac{1}{400}\right)} = .005 \text{ or } 2.5\%$$

### 13.3.2 Standard Error and Precision :

From what we have studied already it should be clear that greater the standard error, greater is the departure of observed value from expected values.

**In Example 10.** Suppose  $p$  remains  $\frac{3}{8}$   $q = \frac{5}{8}$  but  $n = 135$ . Then  $S.E. = \sqrt{\left(\frac{3}{8} \times \frac{5}{8} \times \frac{1}{135}\right)}$

4.16% app.

The limits of under-weight children will then be  $37.5 \pm 3 = 4.16$  i.e. 25.22 and 49.9%. Making the statement that limits of under-weight children are 25.02 and 49.98% is much less precise than that they are 30.3 and 44.7%. In this way standard error gives us an idea about the unreliability

of an estimate in a sample. The estimate and  $\frac{1}{S.E.}$  is sometimes called precision. This reciprocal of

standard error.  $\frac{1}{S.E.}$  is a measure of reliability measure Precision, however, is not very much used in sampling studies.

**Example 11.** A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad, Estimate the proportion of bad pineapples in the consignment, as well as the standard of the estimate. Deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5 (UPSC)

The proportion of bad pineapples in the sample  $\frac{65}{500} = 13\%$ .

In the absence of any other information, the proportion of bad pineapples in the sample can be taken as an estimate of bad pineapples in the consignment as a whole.

∴ the estimated percentage of bad pineapples in the consignments = 13%.

$$\begin{aligned} p \text{ the probability of bad pineapples} &= 0.13 \\ q &= 0.87 \\ n &= 500 \end{aligned}$$

Standard error of bad pineapples in sampling distribution.

$$\text{or S.E.} = \sqrt{\left(\frac{pq}{n}\right)} = \sqrt{\left(\frac{13 \times 87}{500}\right)} = \sqrt{\left(\frac{113}{500}\right)} = \sqrt{0.00226} = .015 \text{ or } 1.5\%.$$

Whenever the true proportion of bad pineapples in the consignment, 99.73% changes are (which means it is almost certain) that it lies within a limit of three times the standard error on either side of the estimate.

∴ The limits within which the proportion of bad pineapples almost certainly lie are.

Estimate  $\pm 3$ .S.E.

or  $13\% \pm 3 \times 1.5\%$

or  $13\% \dots 4.5\%$

or  $8.5\% \text{ and } 17.5\%$

### 13.4 Sampling Distribution

To understand properly the ideas and assumptions on which studies of this type are based, it is necessary to develop and understand some theoretical considerations.

Please understand that in sampling of variables there is no question of  $p$  and  $q$ , or success and failure. Our members, of samples can now take any value out of (theoretically at least) infinite values, and not one of the two attributes as in sampling of attributes.

### 13.4 Sampling Distribution.

Suppose we take a random sample of 121 Persons from adult male population of India and find the Arithmetic Mean of their heights. The A.M. could be some value say 65". We take suppose 200 such samples, each of 121 persons. We shall have 200 values of Arithmetic mean, like 67", 66", 64", 66", 68", 67", 65", 66", 67", and so on (Some values will be in fractions also). These values can be classified and grouped in a frequency distribution. This distribution will be called sampling Distribution of the Arithmetic Mean.....

We can calculate standard deviation of heights of 121 persons in each sample. We shall thus get 200 standard deviations. The distribution of these values will be called Sampling Distribution of standard deviation. Coefficient of correlation between heights and weights in each sample could be calculated. We would get 200 coefficients of correlation. Their distribution would be called Sampling Distribution of correlation coefficient. We can thus have .sampling Distribution of any statistical measure, e.g. sampling distributions of median, mode, quartile mean deviation etc.

There is one very important characteristic of all. sampling distributions and it is that they give a more or less, normal distribution. If the number of samples used in, the sampling distribution is large and the size of each sample is large the sampling distribution would be a normal distribution even though the parent distribution from which the samples have been drawn is not normal. This is an important and useful characteristic which forms the basis of sampling studies.



Since the sampling distribution resembles a normal distribution, it can be used to estimate the values of the population from the values of sample and we can then lay down the limits within which the observed or actual 'values will probably lie. You will remember that there is a mathematical relationship' between the area, covered within the mean ordinates and the ordinates at various distances from the mean ordinate, 99.73% area is covered within  $(\text{mean} \pm 3\sigma)$ , 99% area is enclosed within  $(\text{mean} \pm 2.5758\sigma)$ , 95% area is enclosed within  $(\text{mean} \pm 1.96\sigma)$ . The area outside these limits are only 27%, 1% and 5% respectively. In sampling distribution of means, 99.73% chances are that the actual mean lies within  $\text{mean} \pm 3\sigma$  (or 27% or 0.27 that the actual mean lies outside these limits); 99% chances are that the actual mean lies within  $\text{mean} \pm 2.5758\sigma$  (or 1% or .01 that it lies outside these limits)! or chances are 95% that the actual mean lies within  $\text{mean} \pm 1.96\sigma$  for 5% (or .05 that it lie outside these limits).

#### 12.4.2 Simple Sampling of variables :

As in sampling of attributes, our study here can be made, only on the assumption of simple sampling. The conditions of 'simple sampling' in the case of variables are same as in attributes. Only they, have to be worded differently in this context.

1. The drawing of each number of the sample is independent of draws of all other member, and each member of our sample is drawn 'from the same records.
2. The different samples are being drawn from the ' same record.

The standard deviation of a sampling distribution is called the Standard Error, here also.

In actual practice, 'sampling distribution are not available and we have to estimate parameters or statistical measures for the universe from the values of one sample only. In such cases we shall estimate the mean of standard deviation or other measures of the universe and shall then establish the limits with in which these (values of the universe) can be expected to vary with some specified probability.

#### 13.5.1. Standard Error of the Mean :

Standard error of the mean is the standard deviation of the sampling distribution of means. It is calculated by the formula

$$\text{Standard Error of the Mean} = \frac{\sigma(\text{population})}{\sqrt{n}}$$

$n$  here stands for the number of items in the sample.

If standard deviation  $\sigma$  of the population is not known, standard deviation of the sample can be substituted in its place, provided the Sample size is large say  $n \geq 30$ , Then

$$\text{S.E. of the mean} = \frac{\text{S.D. (Sample)}}{\sqrt{n}} = \frac{S}{\sqrt{n}}$$

Suppose in a sample of 121 persons, the average height is 68" and standard deviation is 5.5". Within what limits do we expect the average height of the, population to exist ?

$$\text{S.E. of the Mean} = \frac{s}{\sqrt{n}} = \frac{5.5''}{\sqrt{(121)}} = \frac{5.5}{11} = .5''$$

We can now say that the average height of the population from which the sample was taken at random is expected to lie within the range (Mean  $\pm$  3 S.E., *i.e.* 68"  $\pm$  3  $\times$  5 which is between 66.5 and 69.5". (Please note here we do not say that the height of the members of the population shall lie with, in 66.5" and 69.5". Only the average of their heights will be within this range. The heights of the individuals may well lie outside this range.

### 13.5.2 Sampling errors and levels of significance :

The term sampling error is used to indicate the error at a certain level of significance. Let us first, be clear what we mean by 'level of significance'.

When we take the limits (Mean  $\pm$  3 S.E. 99.7% of the cases are covered in these limits; and only .27% cases are left out. The probability of parameter or population value lying within (mean  $\pm$  3 S.E.) is 99.73% and the probability of its lying outside the limits is 27%. This .27% is the level of significance when the limits are (Mean  $\pm$  3. S.E.) And at 27% level of significance, 3 is the critical value. Similarly, the probability of parameter lying between (Mean  $\pm$  2.5758 S.E.) is 99%. The level of significance of these limits is 1%. The critical value at 1% level of significance is 2.5758. For .5% level of significance the limits would be given by (Mean  $\pm$  1.96 S.E.), 1.96 is the critical value at 5% level of significance.

You will note that the probability of our statement being correct is 95% at 5% level of significance, 99% at 1% level of significance and 99.73% at 27% level of significance. This may seem paradoxical and our commonsense may not easily accept it, but it is true that the level of significance is inversely related to the extent of precision. Further sometimes the term level of confidence is used in place of level of significance. When our level of confidence is 5% we shall be correct in 95% of cases, but when our level of confidence is less say 1% we shall be correct in 99% cases. The limits which are obtained at a certain level of significance or confidence are called the Confidence Intervals. In our illustration of heights in the previous section, the confidence intervals at 27% level of significance are 68"  $\pm$  (3  $\times$  5) *i.e.* 66.5" and 69.5". At say 5% level of Confidence the confidence interval shall be (68"  $\pm$  1.96  $\times$  5) *i.e.* 67.02" and 68.98".

With the explanation of such terms as level of significance or confidence, critical values and confidence intervals, sampling error becomes an easy thing to understand. Sampling error is simply equal to the critical value multiplied by the standard error. Critical value at 5% level of confidence is 1.96.

Therefore, sampling error at 5% level of confidence is (1.96  $\times$  S.E). Sampling error at 1% level of significance is (2.5751  $\times$  S.E.), and at 27% level of confidence" (3  $\times$  S.E.). Find the sampling errors at the various levels of Significance for the illustration given in the previous section. By the way confidence intervals would be given by Mean  $\pm$  Sampling error, which is the same thing as (Mean + Critical value = S.E.).

The most difficult part of the sampling of variables is over. Please go over these pages again in order to understand the theoretical functions of the practical problems that follow.

**Example 12.** An investigator wants to make a survey of the mean weekly wage of 10,000 workers of an industry. Since the study of all the workers is impossible, a representative sample of 400 workers is selected. The mean weekly wage of 400 workers is Rs. 30 and the standard deviation Rs. 2.50, If additional samples were taken by how much would the results differ from the above sample ?

**Solution :** The last sentence of the problem amounts to this : “Within what range is the mean weekly wage of the population expected to lie?”

Size of the universe. In the case 10,000 workers is irrelevant and shall not be used anywhere. Standard error of the mean weekly wages of the sample is S.E. of Mean

$= \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{400}} = \frac{2.5}{20} = \frac{1}{8} = 0.125$  Re 99.73% chances are that the mean weekly wages of the additional samples would lie between Rs.  $(30 \pm 3 \times .125)$  or Rs.  $(30 \pm .375)$  or Rs. 29.625 and Rs. 30.375.

The confidence interval will be different if we consider 5% or 1% level of significance instead of 0.27% as above.

**Example 13.** It has been determined that the average pulse rates of male in the 20-25 age group is 72 beats per minute and that the standard deviation is 8 beats per minute. If a group of 100 distance runners, all in the given age group of 20-25 were examined and found to have an average pulse rate of 68, should this be regarded as significant deviation from the general average ?

**Solution:** Standard deviation of the population, 8 is available should be used.

S.E. of the average pulse rate of 100 distance runners is

$$\text{S.E. of mean} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{100}} = \frac{8}{10} = .8.$$

The difference between the two average pulse rates in 4 beats per minute which is 5 times the standard error. Hence the deviation of average pulse rate of distance runners from the general average is significant.

**Example 14.** The following are the data of height measurements for a random sample or individual from a certain population.

$$\text{Mean} = 60'', \quad n = 81, \quad \sigma = 4.5''.$$

What would be the limits of the 1% confidence interval for the true mean ? (It is given that  $\text{Mean} \pm 2.57580\sigma$  covers 99% area of a normal curve).

**Solution:** Standard error of the mean height of 81 individuals is

$$\text{S.E. mean} = \frac{\sigma}{\sqrt{n}} = \frac{4.5}{\sqrt{81}} = \frac{4.5}{9} = .5$$

At 1% level of confidence, the limits of confidence intervals will be given by  $(\text{Mean} \pm 2.6778 + \text{S.E.})$  or  $66'' \pm 2.5758 \times .5$  or  $66'' \pm 1.2879$  or  $64.712''$  and  $67.2879''$ . They can be approximated to  $64.7''$  and  $67.3''$ .

### 13.5.3 Standard error of coefficient of correlation:

The standard error of the coefficient of correlation, or

$$\text{S.E. (r)} = \frac{1 - r^2}{\sqrt{n}}$$

**Example 15.** A sample of 400 fathers and sons gives a correlation coefficient between their heights as +.8. If additional samples were taken from the same universe, between what limits would the coefficients of correlation in the case of those samples vary ?

The standard error of the coefficient of correlation between the height of the father and the son is

$$\text{S.E. } (r) = \frac{1-r^2}{\sqrt{n}} = \frac{1-8^2}{\sqrt{400}} = \frac{1-64}{20} = \frac{.36}{.20} = 0.18$$

The correlation coefficient of other samples would lie between  $r \pm 3 \text{ S.E.}$  or  $8 \pm 3 \times .018$  i.e. between .746 and 84.

Standard errors of almost every statistical measures are given in your text books.

**Example 16.** If 60 new entrants in a given university are found to have a mean height of 68.60 inches, and 50 seniors a mean height of 69.51 inches, is the evidence conclusive that the mean height of the seniors is greater than that of the new entrants ? Assume the standard deviation of height to be 2.48 inches.

**Solution:** The observed difference between the mean heights of the two samples (new entrants and seniors) is

$$69.51 - 68.60 = .91''$$

The two independent samples come from the same universe. Standard error of the difference of the two mean heights or

$$\text{S.E. } (m_1 = m_2) = \left[ p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right]$$

$\sigma p$  is given as 2.48"

$$n_1 = 60, n_2 = 50$$

$$\text{S.E. } = (m_1 - m_2) = \sqrt{\left\{ (2.48)^2 \times \left( \frac{1}{60} + \frac{1}{50} \right) \right\}}$$

$$\sqrt{\left[ 6.1504 \left( \frac{5+6}{300} \right) \right]} = \sqrt{(.2255)} \\ = 47.$$

The observed difference .91 is less than two times this standard error and could, therefore have arisen due to fluctuations of simple sampling.

### 13.6. The Student 't' Test and its Properties

The probabilities of the 't' distribution have been tabulated by W.S. Gosset, who wrote under the pseudonym Student which gave, the name to the 't' W.S. Gosset found that the distribution of standard deviation of small samples departs systematically from a normal distribution. Therefore the technique, meant for large samples cannot be applied to small samples. It is well known that if the sample is sufficiently large ( $n > 30$ ) the estimates are adequate for the application of the Z

transformation  $Z = \frac{\hat{b}_1}{\text{standard error of } (\hat{b}_1)}$  when the sample is small  $n < 30$  and provided that the population of the parameter is normal, another test can be applied based on the student's Y distribution.

The general formula which transforms the value of any variable  $X$  into 't' units is similar to the  $Z$  transformation, but the 't' values depends in addition on the number of degrees of freedom and it includes the variance estimates  $S^2x$ . The transformation formula ( $t$  statistic) is

$$t = \frac{x_1 - \mu}{S_x} \text{ with } n - 1 \text{ degrees of freedom}$$

where  $\mu$ , = value of the population mean

$S^2x$  = Sample estimate of the. population variance<sup>3</sup>

$$S^2x = \Sigma(x_1 - \bar{x})^2 \frac{1}{n-1}.$$

$n$  = sample size.

The sampling distribution in this case, that is the distribution of the sample mean, is  $\bar{x} = N$

$(\mu, S^2 \bar{x})$  and the transformation statistic is  $(\bar{x} - \mu) / \sqrt{S^2 \frac{x^2}{N}}$ , and has at 't' distribution with  $(n - 1)$  degrees of freedom.

The 't' distribution is always symmetric, with  $m$  equal to zero and variance  $(n - 1)/n - 3$ , which approaches unity when  $n$  is large. Clearly as  $n$  increases, the 't' distribution approaches the standard Normal distribution  $Z \approx N(0, 1)$ .

The  $t$  distribution is independent of population parameters  $\mu$  and  $\sigma^2$ . The definition of  $t$  is independent of  $\sigma^2$  and therefore, it will not be necessary to know  $\sigma^2$  (before, using this distribution. This is the chief merit of the  $t$  distribution. Before its discovery it was usual to replace the unknown by  $S$  in.

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ to get } t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

and still to regard  $t$  as normally distributed. .

**13.6.1 Student 't' Distribution:** Let  $x_1, x_2, \dots, x_n$  be the members of a random sample drawn from a normal population with mean  $\mu$  and variance  $\sigma^2$ . Then we define the statistic  $t$  as :

$$t = \frac{\bar{X} - \mu \sqrt{n}}{S}$$

where 
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

i.e., 
$$(n-1) S^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = ns^2$$

Hence 
$$\frac{t^2}{v} = \frac{(\bar{x} - \mu)^2 + \frac{\sigma^2}{n}}{\frac{ns^2}{\sigma^2}} \text{ where } v = n - 2$$

Thus  $\frac{t^2}{v}$  is the ratio of two independent  $X^2$  variates, distributed with 1 and  $v$  degrees of freedom respectively, and therefore  $\frac{t^2}{v}$  is a  $\beta \left[ \frac{1}{2}, \frac{v}{2} \right]$  variate and its distribution is

$$\rho_p = \frac{\left( \frac{t^2}{v} \right) - \frac{1}{2} d \left( \frac{t^2}{v} \right)}{\beta \left( \frac{1}{2}, \frac{v}{2} \right) \left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}} (v+1)}, 0 \leq t^2 < \infty$$

or 
$$d_p = \frac{\partial t}{\sqrt{v} \beta \left( \frac{1}{2}, \frac{v}{2} \right) \left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}} (v+1)}, -\infty \leq t < \infty.$$

the factor 2 disappearing since the integral  $-\infty$  to  $+\infty$  must be equal to unity. The distribution is known as the  $t$  distribution with  $v$  degrees of freedom.

### 13.6.2. Chief Features of the $t$ Probability Curve:

The equation to the ' $t$ ' probability curve is

$$y = \frac{1}{\sqrt{v} \beta \left( \frac{1}{2}, \frac{v}{2} \right)} \frac{1}{\left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}} (v+1)}$$

The curve is symmetrical about the line  $t = 0$  since only even powers of  $t$  appear in the equation of the curve. Further since  $\frac{1}{1+t^2/v}$  degrees rapidly as  $(r)$  increases the curve will fall off to zero on each side of the origin.

The curve has a maximum ordinate at  $t = 0$

The probability that the value of  $t$  from a random sample will be between two fixed samples  $t_1$  and  $t_2$  is given by integrating  $y$  with respect to  $t$  from  $t_1$  to  $t_2$ . Similarly, if  $P$  be the probability that the value of  $|t|$  from a random sample will exceed ' $t$ ' then we have

$$P = 2 \int_{t_0}^{\infty} \frac{1}{\sqrt{v} \beta \left( \frac{1}{2}, \frac{v}{2} \right)} \frac{\partial t}{\left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}} (v+1)}$$

values of ' $t$ ' have been tabulated for various fixed values of  $P$  for values of  $v$  from 1 to 60. For values of  $v > 60$ , the fact  $t$  is a symbolically a standard, normal variate can be used.

### 13.6.3. Properties off H' Distribution

(1) The graph of the t distribution is lower, at the centre and high at tails as shown in fig. 12.1.

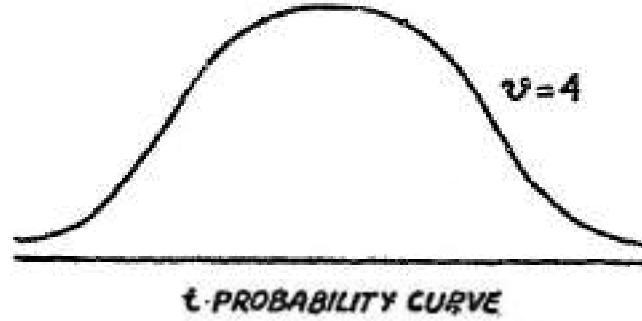


Fig 12.1

(2) The t statistic ranges from  $-\infty$  to  $+\infty$ .

(3) For values of  $v > 60$ , the fact, t is asymptotically a Standard normal variate can be used. As N approaches unity, the t distribution approaches, normal curve.

(4) The t distribution is similar to the normal curve since it is single peaked at and symmetrical about, a zero mean, for the case in which are under the distribution is unity.

(5) The table of 't' falls rapidly as the number of degrees freedom from 1 to 10 and vary, slightly as degrees of freedom increases from 10 to 40.

(6) All the moments of odd order about the origin vanish. The moment or order  $2r$  about the origin (which is also, the mean) is

$$\begin{aligned}\mu_{2r} &= 2 \int_0^{\infty} \frac{t^2}{\sqrt{v} \beta \left( \frac{1}{2}, \frac{v}{2} \right) \left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}(v+1)}} \frac{\partial t}{\partial t} \\ &= \int_0^{\infty} \frac{t^2 r \left( \frac{t^2}{v} \right)^{-1/2} \partial(t^2/v)}{\beta \left( \frac{1}{2}, \frac{v}{2} \right) \left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}(v+1)}} \\ \mu_{2r} &= \frac{v^r}{\beta \left( \frac{1}{2}, \frac{v}{2} \right)} \int_0^{\infty} \frac{\left( \frac{t^2}{v} \right)^{r+1/2-1} \partial}{\left( 1 + \frac{t^2}{v} \right)^{\frac{1}{2}(v+1)}} \left( \frac{t^2}{v} \right) \\ &= \frac{v^r}{\beta \left( \frac{1}{2}, \frac{v}{2} \right)} \beta \left( r + \frac{1}{2}, \frac{v}{2} - r \right), r < \frac{v}{2}\end{aligned}$$

$$\left[ \text{On letting } 1 + \frac{t^2}{v} = \frac{1}{Y} \text{ i.e. } \frac{t^2}{v} = \frac{1-Y}{Y} \right]$$

$$\text{Hence } \mu_{2r} = \frac{(2r-1)(2r-3)\dots\dots 1}{(v-2)(v-4)\dots\dots(v-2r)}$$

### 13.7. Snedecor's F Distribution

In the preceding section we discussed the methods of determining whether two samples have come from the same universe or from two universes which are significantly different from each other. One of the method by which this study is done is by the calculation of standard error of the difference of the means of two samples; another method is the  $x^2$  test' and in case of small samples the method that is generally followed is that of t test. Here we shall discuss Snedecor's 'F' distribution which shows that the distribution of the ratio of independent estimates of the population variance. We know that variance

$$V = \sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$

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

degrees of freedom in such cases are equal to  $n - 1$ . In other words, in small samples.

$$\sigma = \frac{\sqrt{\Sigma(x - \bar{x})^2}}{n-1} \quad \& \quad V = \frac{\Sigma(x - \bar{x})^2}{n-1}$$

There are two types of variations in the data. One between the various samples and the other within the various samples. Now if the variations within the samples and between the various samples are not significantly different from, each other then the samples belong to the same universe.

Suppose the values of the items in the four samples were as follows.

	Sample 1	Sample 2	Sample 3	Sample 4
	4	6	12	10
	6	8	16	10
	2	6	14	10
	6	10	8	6
	2	0	20	4
Total	20	30	70	40
Mean	4	6	14	8

Total number of items in the sample or  $N = 20$ .

$$T = 20 + 30 + 70 + 40 = 160$$

The grand mean of all the items of all the samples =  $160/20 = 8$ .



## Total Variations

Squares of the deviations of various items from the grand average of 8

	Sample 1	Sample 2	Sample 3	Sample 4
	16	4	16	4
	4	0	64	4
	36	4	36	4
	4	4	0	4
	36	64	144	16
Total	96	76	260	32

Grand total of squares =  $96 + 76 + 260 + 32 = 464$

Degrees of freedom =  $20 - 1 = 19$

### 13.7.1 Variance Between The Samples:

We shall calculate the square of the deviations of the means of the various samples from the grand average, if the value of each item in the first sample be taken as 4, for the second sample as 6, in the third samples as 4 and in the fourth as, 8 and the squares of the deviations of those values of the grand average are calculated, they would be below:

	Sample 1	Sample 2	Sample 3	Sample 4
	16	4	36	0
	16	4	36	0
	16	4	36	0
	16	4	36	0
	16	4	36	0
Total	80	20	180	0

Variance between the samples is  $\frac{280}{4-1} = \frac{280}{3} = 93.7$

### 13.7.2 Variance within Samples

Thus in sample 1, the deviation would be taken from 4, in sample 2 from 6, in sample 3 from 14 and in sample 4 from 8. These deviations would be squared & totaled.

	Sample 1	Sample 2	Sample 3	Sample 4
	0	0	4	4
	4	4	4	4
	4	0	0	4
	4	16	36	4
	4	36	36	16
Total 16	56	80	32	

$$\begin{aligned}\text{Grand total of the sum of squares} \\ = 16 + 56 + 80 + 32 = 184\end{aligned}$$

Variance with in the samples

$$= \frac{184}{20-4} = \frac{184}{16} = 11.5$$

All these results can be tabulated as follows:

Source of variation	Sum of squares	D.F.	Variance
Between Sample	280	3	$\frac{280}{3} = 93.3$
Within Sample	184	16	$\frac{184}{16} = 11.5$
Total	464	19	$F = \frac{93.3}{11.5} = 8.1$

If should be remembered that

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

It has been noted that variance between samples is generally greater than variance within samples. Now if we look at Snedecor's table for the value  $F$  for the given degrees of freedom at 50% level of significance. The calculated value of  $F$  is higher than this and as such the difference is significant.

- (1) The variance ratio of  $F$  has a very important, property that its value remains unchanged, if all the figures are either multiplied or divided by a common factor or if a common factor is added to or subtracted from each figure.
- (2) The numerator and denominator of the second member are independent  $X^2$  variates with  $\nu_1$  and  $\nu_2$  degrees freedom respectively.
- (3) The distribution of  $F$  is independent of the population variance  $\sigma^2$  and depends on  $\nu_1$  and  $\nu_2$  only. The  $F$  curve is J shaped if  $\nu_2 \leq 2$  and bell shaped for  $\nu_1 > 2$ . For  $\nu_1 > 4$  the shape of the  $F$  curve is as shown in figure below.
- (4) The distribution has highly positive skewness.

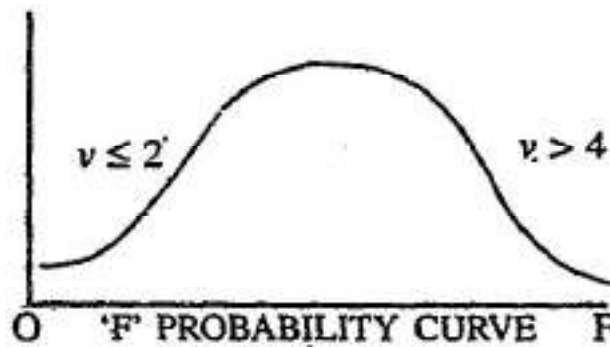


Fig. 12.2

- (5) The probability density of the distribution increases steadily at first reaching at the highest peak (corresponding to the model value) and then goes on decreasing slowly so as to become tangential at infinity. But mode exists if and only, if  $v_1 > 2$  and is equal to  $\frac{v_2}{v_2 + 2} \frac{v_1 - 2}{v_1}$ .

Hence mode of F distribution is always less than 1.

- (6) For  $F_{v_1, v_2}$  distribution mean  $d(\mu_1) = \frac{v_2}{v_2 - 2}$  and variance  $(\mu_2 \cong 2) \left( \frac{1}{v_1} + \frac{1}{v_2} \right)$ .

- (7) The distribution free from population parameters.

- (8) When  $v_1 = 1$ , the F distribution becomes equal to that of  $t^2$  with  $v_2$  degrees of freedom.

- (9) When  $v_2 = \infty$ . it means  $S_2^2 = \sigma^2$ . So that  $v_1, F = xv_1^2$ .

- (10) When  $v_1 = \infty$ . it means  $S_1^2 = \sigma^2$ . So that  $\frac{v_2}{F} \frac{v_2}{F} = xv_2^2$ .

- (11) For large  $v_1$  and  $v_2$   $F \approx N \left[ 1 \sqrt{\left\{ 2 \left( \frac{1}{y} + \frac{1}{v} \right) \right\}} \right]$

### 13.7.3 Some Properties of the F Distribution and F curve

Let  $x_i$ , ( $i = 1, 2, \dots, n_1$ ) and  $x_j$  ( $j = 1, 2, \dots, n_2$ ) be the values of two independent random samples drawn from the same normal population with variance  $\sigma^2$ . Let  $\bar{x}_1$  and  $\bar{x}_2$  be sample means and let.

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2$$

and 
$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^n (x_{2j} - \bar{x}_2)^2$$

Then we define the statistic F by the relation

$$F = \frac{S_1^2}{S_2^2}$$

Hence 
$$\frac{v_1 F}{v_2} = \frac{(n_1 - 1) \frac{S_1^2}{\sigma^2}}{(n_2 - 1) \frac{S_2^2}{\sigma^2}}$$

where  $v_1 = n_1 - 1, v_2 = n_2 - 1$ .

The numerator and denominator of the second member are independent  $X^2$  variates with  $v_1$  and  $v_2$  d.f. respectively. Hence  $v_1 F/v_2$  is a  $\beta_2 (v_1/2, v_2/2)$  variate so that the probability that a random value of F will fall in the interval  $\partial F$  is

$$\partial_p = \frac{\frac{v_1}{2} v_2 \frac{v_2}{2} F^{\frac{v_1-2}{2}} \partial F}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right) (v_1 F + v_2)^{\frac{1}{2}(v_1+v_2)}}$$

This distribution is called the distribution of the variance ratio F with  $v_1$  and  $v_2$   $\partial f$ .  
The  $r$  th moment about the origin is given by

$$\begin{aligned} \mu_r^1 = E(F) &= \frac{\left(\frac{v_1}{v_2}\right)^{\frac{v}{2}}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \\ &= \frac{\int_0^1 F + \frac{v}{2} - 1 \left(1 + \frac{v_1}{v_2} F\right)^{\frac{-v_1+v_2}{2}} \partial F}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \\ &= \frac{\left(\frac{v_1}{v_2}\right)^{-r}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \int_0^1 Y + \frac{v^2}{2} r - 1 (1-y)^{y+\frac{v_1}{2}-1} \partial y \end{aligned}$$

where  $1 + \frac{v_1}{v_2} F = \frac{1}{Y}$  and  $\partial F = \frac{v_2}{v_1}, \frac{\partial y}{Y^2}$

$$\mu_r^1 = \frac{\left(\frac{v_1}{v_2}\right)^{-r}}{\beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \frac{\beta\left(\frac{v_1}{2} + y, \frac{v_2}{2} - y\right)}{\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} + y}, y < \frac{v_2}{2}$$

or 
$$\mu_r^1 = \frac{\sqrt{\frac{v_1}{2} + r} \sqrt{\frac{v_2}{2} + r}}{\sqrt{\frac{v_1}{2}} \sqrt{\frac{v_2}{2}}} \left(\frac{v_2}{v_1}\right)^y, r < \frac{v_2}{2}$$

in particular

$$\mu_1^1 = \frac{\sqrt{\frac{v_1}{2} + 1} \sqrt{\frac{v_2}{2} - 1}}{\sqrt{\frac{v_1}{2}} \sqrt{\frac{v_2}{2}}} - \left(\frac{v_2}{v_1}\right) = \frac{v_2}{v_2 - 2}$$

which is independent of  $v_1$ , and is always greater than unity.

$$\mu_2^1 = \frac{\sqrt{\frac{v_1}{2} + 2} \sqrt{\frac{v_1}{2} - 2}}{\sqrt{\frac{v_1}{2}} \sqrt{\frac{v_2}{2}}} \left( \frac{v_2}{v_1} \right)^2 = \frac{(v_2 + 2)v_2}{(v_2 - 2)(v_2 - 4)v_1}$$

$$\mu_2 = \mu_2^1 - \mu_1 - 2 = \frac{2v_1^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \\ 2 \left( \frac{1}{v_1} + \frac{1}{v_2} \right)$$

### 13.8 Interrelationship Between $\frac{x}{\sigma}$ , $t$ , $x^2$ , and F

As has been noted earlier in 't' distribution properties that  $t$  distribution approaches the normal distribution as  $n$  approaches infinity. The normal distribution is therefore a special case of the  $t$  distribution.

For the same set of data normal distribution yield the same, probabilities as do  $x^2$  values when  $n = 1$  for of  $x^2$ . More specifically, comparing *Areas in Two Tails of the Normal Curve at Selected Values of  $\frac{x}{\sigma}$  or  $\frac{x}{\sigma}$*  from the Arithmetic mean and values of  $x^2$ , that for a given probability

$$\left( \frac{x}{\sigma} \right)^2 = x^2, \text{ when } n = 1 \text{ for } x^2.$$

For any given probability  $\frac{x^2}{n} = F$ . when  $n$  for  $X^2$  equals  $n_1$  for  $F$  and when  $n_2 = \infty$  for  $f$ . This can be seen by comparing values of  $X^2$  and values of  $F$  (from their respective tables).

It is also to be noted that for any given probability,  $t^2 = F$ . when  $n$  for  $t$  equals  $n_2$  for  $F$  and when  $n_2$  for is 1. This is apparent from an examination of values of  $t$  and values of  $F$  (from their respective tables).

$F$  distribution is an inclusive distribution in that the other three distributions are merely special cases of  $F$ .

### APPLICATIONS

#### 13.9. Special Tests

For large samples the sampling distributions of many statistics are normal distributions with mean  $\mu_s$ , and standard deviation  $\sigma_s$ . In each case the result hold for infinite populations or for sampling with replacement.

**(1) Means.** Here  $s = \bar{x}$ , the sample mean :

$\mu = 4 \bar{x} = \mu$  the population mean :

$\sigma_s = \sigma_{\bar{x}} = \sigma / \sqrt{N}$  where  $\sigma$  is the population standard deviation and  $N$  is the sample size;  
The  $Z$  score is given by

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

when necessary the sample deviations  $s$  or  $\hat{s}$  is used to estimate  $s$ .

**(2) Proportion:** Here  $S = P$ , the proportion of “successes” in a sample  $\mu_s = \mu_p = P$  where  $p$  is the population of proportion of success  $N$  = the sample size;

$$\sigma_s = \sigma_p = \sigma / \sqrt{PN / N}$$

where  $q = 1 - P$ . The  $Z$  score is given by

$$Z = \frac{P - p}{\sqrt{PN / N}}$$

In case  $P = \frac{x}{N}$  where  $x$  is the actual number of successes in a sample, the  $z$  score becomes

$$Z = \frac{X - NP}{\sqrt{NPq}}$$

i.e.,  $\mu_a = \mu = N p$   $\sigma_a = \sigma = \sqrt{NPq}$ , and  $S = x$ .

### I. Tests of Significance Involving sample Difference

Let  $\bar{x}_1$  and  $\bar{x}_2$  be the means obtained in large sample of size  $N_1$  and  $N_2$  drawn from respectively, populations having means  $\mu_1$  and  $\mu_2$  and standard deviation  $\sigma_1$  and  $\sigma_2$ .

Consider the null hypothesis that there is no difference between the populations means, i.e.

$$\mu_1 = \mu_2$$

$$\mu_{\bar{x}} - x_2 = \mu_{\bar{x}} - \mu_{\bar{x}} = \mu_1 = \mu_2 \text{ and}$$

$$\rho_{\bar{x}} - \bar{x}^2 = \sqrt{\rho_x^2 + \rho_{\bar{x}}^2} = \sqrt{\frac{\rho_1^2}{N_1} + \frac{\rho_2^2}{N_2}} \quad \dots 1$$

The above equation is when  $S_1$  and  $S_2$  are the Sample means from the two populations  $S$ , which we denote by  $\bar{x}_1$  and  $\bar{x}_2$ , then the sampling distribution of the differences of means is given for infinite population with, mean and standard deviation  $\mu_1$ ,  $\sigma_2$  and  $\mu_2$  respectively as in equation I.

$$\mu_{\bar{x}} - \bar{x}_2 = 0 \text{ and}$$

$$\sigma_{\bar{x}} - \bar{x}_2 = \sqrt{\left(\frac{\sigma_1^2}{N_1}\right) + \left(\frac{\sigma_2^2}{N_2}\right)}$$

Where we can, if necessary, use the sample standard deviations  $S_1$  and  $S_2$  (or  $\hat{s}_1$  and  $\hat{s}_2$ ) as estimates of  $\sigma_1$  and  $\sigma_2$ .

By using the standardized variable or Z score given by

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

$$= \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

We can test the null, hypothesis, against alternative hypothesis (or the significance of an observed difference at an appropriate level of significance).

## 2. Differences of Proportions

Corresponding results can be obtained for the sampling distributions of differences of proportions from two binomially distribute populations with parameters  $P_1, q_1$ , and  $P_2, q_2$  respectively. In this case  $S_1$  and  $S_2$  correspond to the proportion of successes,  $P_1$  and  $P_2$ , and equation  $\mu_s - S_2 = \mu_s - \mu_s$  and  $\sigma_s - S_2$ .

$\sqrt{\sigma_{S1}^2 + \sigma_{S2}^2}$  yield the results.

$$\mu_{p1-p2} = \mu_{p1-p2} = P_1 - P_2.$$

and  $\sigma_{p1-p2} = \sqrt{\sigma_{p1}^2 + \sigma_{p2}^2} = \sqrt{\frac{P_1 q_1}{N_1} + \frac{P_2 q_2}{N_2}}$

If  $N_1$  and  $N_2$  are large ( $N_1, N_2 \geq 30$ ) the distributions of differences of means or proportions are very closely normally distributed.

We noted that the sampling distribution of differences in proportions is approximately normally distributed with mean and standard deviation given by

$$\mu_{p1-p2} = 0 \text{ and } \sigma_{p1-p2} = \sqrt{pq \left( \frac{1}{N_1} + \frac{1}{N_2} \right)}$$

Where  $P = \frac{N_1 P_1 + N_2 P_2}{N_1 + N_2}$  is used an estimate of the population proportion and  $q = 1 - P$ .

By using the standardized variable

$$Z = \frac{P_1 - P_2 - 0}{\sigma_{P1-P2}} = \frac{P_1 - P_2}{\sigma_1 - \sigma_2}$$

And can test observed differences at an appropriate level of significance and thereby test the null hypothesis.

## Tests of Means and Proportions Using Normal Distributions

**Example 1.** Find the probability of getting 40 and 60 heads inclusive in 100 tosses of a fair coin.

**Solution :** (a) according to binomial distribution the required probability is

$$100C_{40} \left(\frac{1}{2}\right)^{40} \left(\frac{1}{2}\right)^{60} + 100C_{41} \left(\frac{1}{2}\right)^{41} \left(\frac{1}{2}\right)^{59} + \dots + 100C_{60} \left(\frac{1}{2}\right)^{60} \left(\frac{1}{2}\right)^{40}$$

Since  $N_p = 100 \left( \frac{1}{2} \right) = nq = 100 \left( \frac{1}{2} \right)$  are both greater than 5, the normal approximation to the binomial distribution can be used in evaluating this sum.

The mean and standard deviation of the number of heads in 100 tosses are given by

$$\mu = N_p = 100 (1/2) = 50$$

$$\sigma = \sqrt{N p q} = \sqrt{100 \times 1/2 \times 1/2} = 5$$

On a continuous scale, between 40 and 60 heads inclusive is the same as between 39.50 and 60.5 heads.

$$39.5 \text{ in a standard units} = \frac{(39.5 - 50)}{5} = -2.10$$

$$60.5 \text{ is standard units} = \frac{(60.5 - 50)}{5} = 2.10$$

Required probability = area under normal curve between  $Z = -2.10$  and  $Z = 2.10$ .

$$= 2 (\text{area between } Z=0 \text{ and } Z = 2.10)$$

$$= 2 (0.4821) = 0.9642.$$

(b) To test the hypothesis that a coin is fair, the following rule of decision is adopted:

(1) Accept the hypothesis if the number of heads in a single of 100 tosses is between 40 and 60 inclusive.

**Example 2. Reject the hypothesis otherwise**

(a) Find the probability of rejecting the hypothesis when it is actually correct.

(b) Interpret 'graphically the decision rule and the results of part (a).

(c) What conclusions would, you "draw if the samples of 100 tosses yielded 53 heads? 60 heads?

(d) Could you be wrong in your, conclusions to (c)? Explain.

**Solution:**

(a) According to above problem the probability of not getting between 40 and 60 heads inclusive if the coin is fair =  $1 - 0.9642 = 0.0358$ . Then the probability of rejecting the hypothesis when it is correct = 0.0358.

(b) If a single sample of 100 tosses yields a Z score between -2.10 and 2.10. We accept the hypothesis otherwise, we reject the hypothesis and decide that the coin is not fair. The error made in rejecting the hypothesis when it should be accepted is the Type I error of the decision rule.



Fig. 12.3



If a single sample of 100 tosses yields a number of heads whose Z score (or Z statistic) lies in the shaded region, which shows that the score differed significantly from what would be expected if the hypothesis were true. For this reason the total shaded area (i.e., Probability of a type I error) is called the level of significance of the decision rule and equals 0.0358 in this case. Thus we speak of rejecting the hypothesis at 0.0358 or 3.58% level of significance.

- (c) According to the decision rule, we would have to accept the hypothesis that the coin is fair in both cases.
- (d) Yes, we could accept the hypothesis when it actually should be rejected, as would be the case. For example, when the probability of the need is really 0.7 instead of 0.5.

\*\*\*\*\*

## LESSON—15

### TESTING HOMOGENEITY OF SEVERAL INDEPENDENT ESTIMATES OF POPULATION VARIANCE

Dear Student,

**Example 1:** A random sample of 9 steel beams has an average compressive strength of 55,815 pounds , per square inch and a standard deviation of 200 pounds per square inch. Test the hypothesis that the true average strength of the Steel beams from which this sample was taken is 56,000 pounds per square inch, using a two sided alternative and a level of significance 0.05.

**Solution:** We treat the hypothesis to be tested namely  $\mu = 56,000$  pounds per square inch (PSi) as the null hypothesis.

The alternative hypothesis, being two sided is  $\mu \neq 56,000$  pSi.

Since we are using a small sample consisting of 9 steel beams, the appropriate distribution is the t distribution having (9.1) or 8 degrees of freedom, shown in figure 13.1 below :-

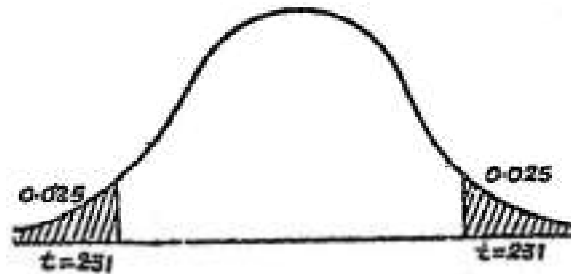


Fig. 13.1

The  $t$  value corresponding to a level of significance 0.05 ‘two tailed’ test’ is found to be 2.31, from the table in which ‘ $t$ ’ values are for different levels of significance and degrees of freedom are given.

The difference between the sample mean and null hypothesis mean is now expressed in term.

$$t = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n-1}}}$$
 where  $\bar{x}$  and  $\mu$  are the sample and hypothesis means respectively and  $\sigma$ , the sample standard deviation.

$$\begin{aligned}
 &= \frac{55815 - 56000}{\frac{200}{\sqrt{8}}} \\
 &= \frac{-185 \times \sqrt{8}}{200} = -2.62
 \end{aligned}$$

Since this ‘ $t$ ’ value exceeds that  $t_{0.05}$  value for 8 degrees of freedom we reject the null hypothesis and conclude that the true mean compressive strength of the lot of the steel beams from which the sample has been taken is not, 56,600 psi.

### Tests of Significance based on t distribution

We shall consider five tests of significance based on the t distribution.

**Example 2.** Show that 95% fiducially limits for the mean  $\mu$ , of the population are  $\bar{x} \pm St_{0.05}/\sqrt{n}$ . Deduce that for a random sample of 16 values with mean 41.5 inches and the sum of the squares of the deviations from the mean (135) (inches)<sup>2</sup> and drawn from a normal population 95% fiducially limits for the mean of population are 39.9 inches and 43.1 inches.

**Solution:**

$$\text{We have } p \left[ \left| \frac{\bar{X} - \mu/\sqrt{n}}{S} \right| \leq t_{0.05} \right] = .95$$

$$\text{or } p \left[ \bar{X} - St \frac{0.05}{\sqrt{n}} \leq \mu \leq \bar{X} \right] = .95$$

So that we can say with a confidence .95 that the confidence interval  $\bar{X} \pm \left| \frac{S}{\sqrt{n}} \hat{t}_{0.05} \right|$  contains the population mean  $\mu$ . The limits of this confidence or fiducially limits for  $\mu$ .

In the particular case

$$n = 16, v = n - 1 = 15.$$

$$S = \sqrt{\frac{1}{15} \times 135} = 3$$

Also from the table,  $t_{0.05} = 2.13$

$$\sqrt{\frac{n}{S}} t_{0.05} = \frac{3}{4} \times 2.13 = 1.6 \text{ Approx.}$$

Therefore the required fiducial limits are  $41.5 \pm 11.6$  i.e. 39.6 and 41.3 inches.

**Example 3.** Two yields of the types ‘Type A’ and ‘Type B’ of cereals in kgs per hectare in 6 replications are given below. What comments would you make on the difference in the mean yields? You may assume that if there be 5 degrees of freedom and  $P = 0.2$ ,  $t$  is approximately 1.476.

Replication	‘A’ Yield in kgs	‘B’ Yield in kgs
1	20.50	24.86
2	24.60	26.39
3	23.06	28.19
4	29.98	30.75
5	30.37	29.97
6	23.83	22.04

**Solution:—**

Replication	A (yield in kgs)	B (yield in kgs)	$\partial$	Deviation from mean ( $\partial - \bar{\partial}$ )	Sq. of deviations ( $\partial - \bar{\partial}$ ) <sup>2</sup>
1	20.50	24.86	4.36	+2.717	7.38
2	24.60	26.39	1.79	+0.147	0.02
3	23.06	28.19	5.13	+3.487	12.15
4	29.18	30.75	0.77	- 0.873	0.76
5	30.37	29.97	-0.40	- 2.043	4.17
6	23.83	22.04	-1.79	- 3.433	11.79
					<u>36.27</u>
			<u><math>\Sigma \partial = 9.86</math></u>		

$\bar{\partial}$  = mean of difference of the yields

$$= \frac{\Sigma \partial}{N} = \frac{9.86}{6} = 1.643 \text{ kgs.}$$

S = S.D. of difference of yields

$$= \sqrt{\frac{\Sigma(\partial - \bar{\partial})^2}{n - 1}}$$

$$= \sqrt{\frac{36.7}{5}} = 2.69 \text{ kgs.}$$

We set up the null hypothesis, that the difference in type has no effect on yield *i.e.*, the population mean of the difference is zero then

$$t = \frac{1.643 - 0}{2.69} \sqrt{\sigma} = 1.489$$

This value of  $t$  is less than  $t_{0.05}$  for 5 *d.f.* and therefore, the difference is not significant at 20% level.

Give two independent random samples from normal population with the same variance we have to test the hypothesis that the population means are  $\mu_1$  &  $\mu_2$  respectively.

For this case

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with  $(n_1 + n_2 - 2)$  *d.f.* and the test of significance is carried out as  $t = \frac{\bar{\partial} - 0}{\left(\frac{S}{\sqrt{n}}\right)}$ .

**Note about assumption for the test:**

For testing significance of a sample mean,  $t$  test has a greater range of validity and precision than the normal approximation tests of  $\left| \frac{x - \mu}{\sigma} \right| \sqrt{n} < 1.96$  but this is achieved by placing an important restriction viz., that the present population must be normal. If the parent population is not normal the distribution of the statistics  $\sqrt{n}(\bar{x} - \mu/S)$  may be quite different from the  $t$  distribution, since, in that case, the distribution of  $\bar{x}$  and  $S^2$  will not be independent and also the distribution will depend on parameters giving the departure of parent distribution from normality.

For testing the significance of the between two sample, means we make an additional assumption that the variances of the two populations are the same. Before applying the  $t$  test it may be desirable to test this assumption by applying the  $F$  test. If the two variance, are different  $t$  test is no longer valid and another test 'd' based on confidence intervals applies by Behren can be used

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2 + S_2^2}}$$

then we carry out the test of significance by using the tables of Sukhative and Fisher for various values of  $n_1, n_2$  and  $S_1^2/S_2^2$ , the values of  $d$  corresponding to various probability levels.

**Testing the significance of observed correlation coefficient**

Given a random sample  $(x_1, y_1), (x_1, y_2) \dots (x_n, y_n)$  from a bivariate normal population, here we have to test the hypothesis that the correlation coefficient of the population is zero *i.e.*, variables are independent.

If me hypothesis is true, it is foe case that in an uncorrelated population *i.e.* when  $P = 0$ ; the distribution of correlation-coefficient,  $r$ , can be obtained easily as follows:

Let the  $n$  values  $i$  ( $i = 1, 2, \dots, n$ ) be subjected to an orthogonal transformation yielding  $n$  independent varieties  $n_1, n_2, \dots, n_n$  so that

$$\sum_{i=1}^n n_i^2 = \sum_{i=1}^n y_i^2$$

choosing  $n_1 = \frac{1}{\sqrt{n}} \sum_i y_i = \sqrt{xy}$  (the sum of the squares of fee coefficient of  $y_1'$  (s is unity) we get

$$\begin{aligned} \sum_{i=1}^n n_i^2 &= \sum_{i=1}^n y_i^2 = \sum_{i=1}^n (y_i - \bar{y} + \bar{y})^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = n\bar{Y}^2 = nS_2^2 + n_1^2 \end{aligned}$$

or 
$$nS_2^2 = \sum_{i=1}^n n_i^2$$

Also  $\sqrt{n} rs_3 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{n} S^1} \sqrt{n} rs_2$

Hence  $\sum_{i=3}^n n_1^2 = n(1-r^2)S_2^2$

Since  $n_1/\sigma^2$  are independent standard normal variates we conclude that  $\frac{nr^2S_2^2}{\sigma_2^2}$  and

$\frac{n(1-r^2)S_2^2}{\sigma_2^2}$  are distributed independently like  $X^2$  with 1 and  $n-2$  degrees of freedom respectively.

now 
$$r^2 = \frac{r^2 S_2^2}{S_2^2} = \frac{nr^2 S_2^2 / \sigma_2^2}{nr^2 S_2^2 / \sigma_2^2 + n(1-r^2)S_2^2 / \sigma_2^2}$$

$$= \frac{\lambda_1^2}{X_1^2 + X_2^2}$$

where  $x_1^2$  and  $x_2^2$  are distributed like  $X^2$  with 1 and  $(n-2)$  d.f. respectively and therefore  $r^2$  is

$\beta \left[ \frac{1}{2}, \frac{n-2}{2} \right]$  variate.

**Hence the distribution of  $r^2$  is**

$$\partial \rho = \frac{\left(r^2\right)^{\frac{1}{2}} - 1(1-r^2)^{\frac{n-4}{2}}}{\beta \left[ \frac{1}{2}, \frac{n-2}{2} \right]} \partial(r^2) \quad 0 \leq r^2 \leq 1$$

Thus the distribution of  $r$  is

$$\partial \rho = \frac{1}{\beta \left[ \frac{1}{2}, \frac{n-2}{2} \right]} (1-r^2)^{\frac{n-4}{2}} \partial r \quad -1 \leq r \leq 1.$$

If the hypothesis is true the critical ratio 't' is defined by the expression.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

That the 't' is a variate with  $(n-2)$  d.f. If our calculated value of (t) exceeds, to .05 for  $(n-2)$  d.f. we say that the value of  $r$  is significant at 5% level of significance, if  $|t| \leq$  to .05 the sample is consistent with the hypothesis of an uncorrelated population.

**Example 4.** A random sample of 15 from a normal universe gives a correlation coefficient of + 0.60. Is of the existence of correlation in the population significant?

Here  $n = 15, r = - 0.60$

$$|t| = \frac{+0.60\sqrt{15-2}}{\sqrt{1-(-0.60)^2}} = +2.70$$

Also for 13 d.f.  $t_{0.05} = 2.16$

$\therefore$  Sample correlation coefficient is significant as the calculated value of t is more than the table value. Hence our hypothesis that sample has been taken from an uncorrelated bivariate normal population appears to be incorrect.

**Example 5.** A random sample of 18 pairs from a bivariate normal population showed a correlation coefficient of 0.3. Is this value significant of a correlation in the population ?

**Solution :** We set up the hypothesis that the variables were uncorrelated in the normal population.

$$\begin{aligned} t &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{3\sqrt{18-2}}{\sqrt{1-(.3)^2}} = \frac{1.2}{\sqrt{1-.09}} \\ &= \frac{1.2}{.91} = 1.26. \end{aligned}$$

The number of degrees of freedom

$$= n - 2 = 18 - 2 = 16$$

The value of t from the table for 15 d.f. at 5% level of significance is 2.12. Thus the calculated value of t is less than the table value.

Hence our hypothesis that sample has been taken from an uncorrelated bivariate normal population appears to be correct.

**Example 6.** It was found that the correlation coefficient between two variables, calculated, from a sample of size 25 was 0.40. Does this show evidence of having come from a population with zero correlation?

**Solution:** We set up the hypothesis that the sample has come from a normal population with zero correlation.

$$\begin{aligned} t &= \frac{r-0}{\sqrt{1-r^2}} \times \sqrt{n-2} \\ &= \frac{.40}{\sqrt{1-(.40)^2}} \times \sqrt{25-2} \\ &= \frac{.40}{\sqrt{1-.16}} \times 4.7958 \\ &= \frac{.40}{\sqrt{.84}} \times 4.7958 \\ &= \frac{1.9183}{\sqrt{.84}} \times 2.09 \text{ Approx.} \end{aligned}$$

The number of degrees of freedom = 25 - 2 = 23.

The table value of t for 23 d.f. at 5% level of significance is 2.07 so that the calculated value is higher than the table value.

Hence our hypothesis appears to be incorrect. We are likely to conclude that the sample has not come from a population with zero correlation.

**Example 7.** Two samples of sizes 23 and 28 give 'r' as 0.5 and 0.85 respectively. Is there any significant difference between the two correlation coefficients ?

$$\begin{aligned}\text{Solution : } Z_1 &= 1.1513 \log_{10} \left( \frac{1+r}{1-r} \right) \\ &= 1.1513 \log_{10} \left( \frac{1+1/2}{1-1/2} \right) = 0.55 \\ Z_2 &= 1.1513 \log_{10} \left( \frac{1+r}{1-r} \right) \\ &= 1.1513 \log_{10} \left( \frac{1+8}{1-8} \right) = 1.10 \\ t &= \frac{Z_1 - Z_2}{\sqrt{\frac{1}{r_1 - 3} + \frac{1}{r_2 - 3}}} \\ &= \frac{1.10 - .55}{\sqrt{\frac{1}{20} + \frac{1}{25}}} = \frac{.55}{0.3} \\ &= 1.83 < 6.96\end{aligned}$$

Hence the difference is not significant at 5% level.

**Example 8.** Find the least value of r in a sample of 18 pairs from a bivariate normal population significant at 2% level.

**Solution :** Substituting the value of 'n' in the formula

$$t = \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \times r$$

we get

$$t = \frac{4r}{\sqrt{1-r^2}}$$

Now for significance of 'r' at 2% level, t should be greater than the value of t from the table at 2% level for 16 degrees of freedom which is 2.58.

$$\frac{4r}{\sqrt{1-r^2}} > 2.58$$



$$r^2 > (.645)^2 (1-r^2)$$

$$r^2 > 0.416025 (1-r^2)$$

$$1.416025 r^2 > 0.416025$$

$$|r| > \sqrt{\frac{0.416025}{1.416025}}$$

$$> 0.54$$

Hence the required least value of r is 0.54 Ans.

**Example 9.** Twelve pictures submitted in a competition were remarked by shown in the table below.

Picture	A	B	C	D	E	F	G	H	I	J	K	L
Rank assigned by first judge	5	9	6	7	1	3	4	12	2	11	10	8
Rank assigned by second judge	5	8	9	11	3	1	2	10	4	12	7	6

Calculate  $\rho$ . Is there a lack of independence in these rankings?

(Assume that on the hypothesis of independence of two sets of  $n$  ranking

$$t = \rho \frac{\sqrt{n-2}}{\sqrt{1-\rho^2}} \text{ follows the t distribution with } (n-2) \text{ degrees of freedom}$$

Given that

Degrees of freedom	10	11	12
Value of t significant at 5% level of probability	2.23	2.20	2.18

**Solution :** Calculation of  $\rho$ , the rank correlation coefficient

Rank difference  $d$ :

$$0.1 -3, -4, -2, 2, 2, 2, -2, -1, 3, 2.$$

$$\therefore \sum d^2 = 0 + 1 + 9 + 16 + 4 + 4 + 4 + 4 + 4 + 1 + 9 + 4 = 60$$

$n$  - number of pictures ranked = 12

$$\therefore \rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{12(12^2-1)} = 1 - \frac{360}{143 \times 12} = .791$$

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} = .791 \sqrt{\frac{12-2}{1-(.791)^2}}$$

$$= .791 \times 5.17 = 4.089$$

number of degrees of freedom =  $n - 2 = 12 - 2 = 10$

The value of t for 10 degrees of freedom at 5% level of significance is 2.3.

$\therefore$  The calculated value of t is greater than the table value and hence deviation is significant.

Therefore, the ranking is significant.

### Testing the significant of an observed partial Correlation

Fisher has shown that the sampling distributor of a partial correlation coefficient of order K (*i.e.*, with K secondary scripts) is of the same form as that of correlation coefficient from a bivariate normal population with the sample size  $n$  reduced by K. In particular; if we are given a random sample from a multivariate normal distribution and we have to test the hypothesis that a particular partial correlation coefficient of order K in the population is zero, we use of the fact.

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-k-2}$$

is a  $t$  variate with  $(n - k - 2)$  d.f.

**Example 10.** Show that in a sample of size 20 from a normal population, a correlation coefficient  $r_{123} = 0.5$  is significant at 5% level.

Here  $r = 0.5$ ,  $n = 20$ ,  $k = 2$

$$t = \frac{0.5}{\sqrt{1-(.5)^2}} \sqrt{16} = \frac{2}{\sqrt{.75}} = \frac{4\sqrt{3}}{3} = \frac{6.928}{3}$$

$= 2.31$  Approx.

As for 16 d.f. to .05 = 2.12.

Also  $|t| > t_{0.05}$  the value of  $r_{123}$  is significant at 5% level.

### Testing the Significance of an observed regression coefficient:

Suppose we are given a random, sample  $x_1 y_1, x_2 y_2, \dots, (x_n y_n)$  from a bivariate normal population. The regression equation of  $y$  on  $x$  is obtained from the sample be

$$y - \bar{y} = b(x - \bar{x})$$

So that the estimated value  $y$  corresponding to given  $x_1$  is

$$y - \bar{y} = b(x_1 - \bar{x})$$

We have to test the hypothesis that the regression coefficient of  $y$  on  $x$  or in the population is  $\beta$ .

If the hypothesis is correct the  $t$  statistic is

$$t = \frac{(b - B) \left[ (n-2) \left( \sum_j x_1 - \bar{x} \right)^2 \right]}{\sum_j (y_1 - \bar{y})^2}$$

Conforms to  $t$  distribution with  $(n - 2)$  degrees of freedom.

**Example 11.** For a sample of size 30 where  $x$  takes the values 1,2, 3.... 30 it is found that.

$$\Sigma(x_1 - \bar{x})(y_1 - \bar{y}) = 599.62 \text{ and } \Sigma(y_1 - \bar{y})^2 = 10.206.$$

Test the significance of the regression coefficient of  $y$  and  $x$ .

$$\text{Solution : Here } \Sigma(x_2 - \bar{x})^2 = \frac{n(n^2 - 1)}{12} = \frac{30 \times 899}{12} = 22475$$

$$b = \frac{\Sigma(x_1 - \bar{x})(y_1 - \bar{y})}{\Sigma(x_1 - \bar{x})^2} = \frac{599.62}{2247.5} = .2668$$

Also  $\Sigma(y - y_1) = \Sigma(y_1 - \bar{y})^2 - b^2 \Sigma(x_1 - \bar{x})^2$   
 $= 1020.6 - 159.97 = 860.59$

Also the hypothesis gives  $B = 0$ .

Substituting in

$$t = \frac{(b - B) \left[ (n - 2) \Sigma(x_1 - \bar{x})^2 \right]}{\Sigma[(y_1 - \bar{y})^2]^{\frac{1}{2}}}$$

$t = 2.28$  with  $\nu = 28$

The value of  $t$  is significant at 5% level but not at 1% level.

**Example 12.** The data showing the aptitude test scores of a random sample of salesman and their first year sales in rupees is given in Table 4 below.

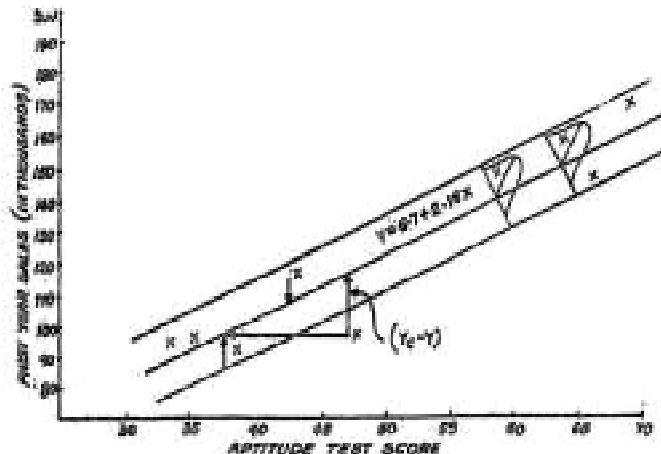
**Table 4 - Aptitude Test Scores and first year sales of Salesman**

Salesman	Aptitude test score	First year sales (in 000's)
A	52	108
B	49	92
C	66	136
D	44	109
E	71	107
F	59	149
G	53	84
H	74	190
I	38	82
J	66	157
K	76	154
L	32	84

We can plot these values on a graph showing the test scores along the X-axis and the corresponding sales on the Y-axis, this diagram is called, “the scatter diagram.”

**Worksheet for Use estimation of constants a and b.**

$n$	Aptitude test score $X_i$	First year Sale in (000 Rs) $y_i$	$x_i$ (original value-55)	$x_i^2$	$x_i y_i$	$y_i^2$
1	52	108	-3	9	-324	11,604
2	49	92	-6	36	-552	8464
3	66	136	11	121	1496	18,496
4	44	109	-11	121	-1199	11,881
5	71	167	16	256	2672	27,889
6	59	149	4	14	596	22,201
7	33	84	22	484	-1848	-7056
8	74	190	19	361	3610	36,100
9	38	82	-17	289	-1394	6,724
10	66	157	11	121	1661	22,801
11	76	154	21	441	3234	23,716
12	32	84	-23	529	-1932	7056
<hr/>						
	$\Sigma y_i = 1506$	$\Sigma x_i = 0$	$\Sigma x = 2784$	$\Sigma x_i y_i = 6020$	$\Sigma y_i^2 = 204,048$	



A straight line can be represented in the general form  $y = a + bx$  where  $a$  is the  $y$  intercept corresponding to  $x = 0$  and ' $b$ ' is the slope of the line, rate of change of  $y$  for unit change in ' $x$ '. if we can obtain the values of the constants ' $a$ ' and ' $b$ ' in this equation, the relationship is determined

- (i)  $\Sigma y_i = n a + b \Sigma x_i$
- (ii)  $\Sigma x_i y_i = a \Sigma x_i + b \Sigma x_i^2$

An intuitively understandable explanation of these equation is as under:

$$\begin{aligned}
 y_1 &= a + bx_1 \\
 \text{i.e. } y_1 &= a + bx_1 \\
 y_2 &= a + bx_2 \\
 &\vdots \\
 &\vdots \\
 y_n &= a + bx_n
 \end{aligned}$$

Summing up both sides

$$Y_1 + Y_2 + \dots + Y_n = na + b(X_1 + X_2 + X_3 + \dots + X_n)$$

$$\Sigma Y_i + na + b \Sigma x_i \text{ the first equation}$$

Similarly, if we multiply each equation by  $X$ , we get,

$$X_1 Y_1 = ax_1 + bx_1^2$$

$$X_2 Y_2 = ax_2 + bx_2^2$$

$$X_n Y_n = ax_n + bx_n^2$$

Summing up both sides

$$\Sigma x_i y_i = a \Sigma x_i + b \Sigma x_i^2, \text{ the second equation.}$$

Solving these two equations we can obtain

$$a = \frac{\Sigma x_i^2 \Sigma Y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$a = \frac{n \Sigma x_i y_i - \Sigma x_i y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$a = \frac{\Sigma x_i^2 \Sigma Y_i - 0 \Sigma x_i y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\Sigma x_i^2 \Sigma Y_i}{n \Sigma x_i^2} = \frac{\Sigma Y_i}{n}$$

$$b = \frac{n \Sigma x_i y_i - 0 \Sigma y_i}{n \Sigma x_i^2 - 0} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

Substituting these values

$$a = \frac{\Sigma Y_i}{n} = \frac{1506}{12} = 125.5$$

$$b = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{6020}{2784} = 2.16$$

The regression equation can be written as

$Y = 125.5 + 2.16 X$  where 125.5 is the value of the intercept on  $Y$  axis when  $x^1 = 0$ , i.e., corresponding to the value 55 in the original data.

The standard deviation (denoted by symbol  $S_{y.x}$ ) of the  $Y$  variable is given by the expression.

$$S_{y.x} = \sqrt{\frac{\Sigma (y - y_0)^2}{n - 2}}$$

A more convenient form of the same relationship for case of calculation is given by

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b \Sigma xy}{n - 2}}$$

The denominator ( $n - 2$ ) shows the number of degrees of freedom, since two degrees of freedom are lost because two quantities  $a$  and  $b$  have been estimated. The suffix  $S_y$  means that we are referring to the variability of the value of  $Y$  corresponding to given values of  $x$ .  $S_{ys}$  for data is worked out as under

$$\begin{aligned}
 S_{y:x} &= \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}} \\
 &= \sqrt{\frac{2.04 \times 0.48 - 125.5 \times 1506 - 2.16 \times 6020}{10}} \\
 &= \sqrt{\frac{2.042}{10}} \\
 &= \sqrt{204.2} \\
 &= 14.3
 \end{aligned}$$

We can estimate his first year sales by substituting  $X = 60$  in the regression equation

$$\begin{aligned}
 Y &= 6.7 + 2.16 \times .60 \\
 &= 136.3 \text{ (in thousand rupees).}
 \end{aligned}$$

For example we can compute the 95.5 percent confidence for the first year sales of salesman whose aptitude test score was 60. The 95.5 percent confidence interval will be

$$136.2 \pm 2 \times 14.3 \text{ (in thousands of rupees)}$$

Or  $136.3 \pm 28.6$  (-do-)

*i.e.* between 107, 700 and 164,900.

\*\*\*\*\*

## LESSON-16

### ANALYSIS OF VARIANCE

**Dear Student,**

The analysis of variance (ANOVA) is a statistical method developed by R.A. Fisher for the analysis of experimental data. Initially, the main application (ANOVA) was confined to the analysis of agricultural experiments but later on, the use of this technique expanded to many other fields of scientific research.

The total variance of a variable can be decomposed into different additive components with the help of analysis of variance which may be attributed to various separate factors. Let us explain with the help of an example that there are twenty plots of land on which rice is cultivated. It is proposed to study the yield per unit of land. Let us further suppose that different seeds, different fertilizers and different means of irrigation are used. Thus the variation in yields may logically be attributed to the three factors.

$x_1$  = type of seed

$x_2$  = type of fertilizer

$x_3$  = type of irrigation.

The total variation in yield can be broken down into three separate components : a component due to  $X_1$ , another due to  $X_2$  and third due to  $X_3$  with the method of analysis of variance: Analysis of variance is conceptually the same as regression analysis, if we go by the above definition of (ANOVA). In regression analysis total variation in the explained variable is split into two components : the variation explained by the regression line, and the unexplained variations shown by the scatter of point around the regression line.

However, it must be noted that there are significant differences between regression analysis and analysis of variance. The main difference is that regression analysis provides numerical values for the influence of the various explanatory factors on the dependent variable, in addition to the information concerning the breaking down of the total variance of  $Y$  into additive components, while the analysis of variance provides only, the latter type of information.

The main objective of regression analysis and ANOVA is the determination of the various factors causing variations of the dependent variable. The method of ANOVA is used in regression analysis for conducting various tests of significance, the most important being :

- (1) The test of the overall significance of the regression.
- (2) The test of the significance, of the improvement in fit obtained by the introduction of additional explanatory variables in the function.
- (3) The test of equality of coefficients obtained from different samples.
- (4) The test of the extra-sample performance of a regression, or test of the stability of the regression coefficients.
- (5) The test of restrictions imposed on coefficient of a function.

An experiment with four samples : During Cooking doughnuts absorb fat in various amounts. Mrs. X wished to learn if the amount absorbed depends on the type of fat use. For each of four fats, six of batches of doughnuts were prepared batch consisting of 24 doughnuts. Data of this kind are called a single or one way classification each fat representing one class.

Before we start, it is to be noted that the totals for the four fats differ quite a lot; from 372 for fat 4 to 510 for fat 2.

**Grams of Fat Absorbed Per Batch**

Fat	1	2	3	4	Total
	64	78	75	55	
	72	91	93	66	
	68	97	78	49	
	77	82	71	64	
	56	85	63	70	
	95	77	76	68	
$\Sigma X$	432	510	456	372	1770 = G
$\bar{X}$	72	85	76	62	295
$\Sigma X^2$	31,994	43652	35,144	23,402	134,192
$\frac{(\Sigma X)^2}{N}$	31,104	43,350	34,656	23,064	132,174
$S_1^2 = \left( \Sigma x^2 - \frac{(\Sigma X)^2}{N} \right)$	890	302	488	338	2018 Pooled $S_1^2$
d.f.	5	5	5	5	20
Pooled $S^2 = 2,018/20 = 100.9$					
$S_p = \sqrt{252/n} = \sqrt{2(100 - 9)/6} = 5.80$					

The analysis of variance is of great utility and flexibility and was developed by Fisher in the 1920's. The analysis of variance performs two functions :

- (1) It is an elegant and slightly quicker way of computing the pooled  $S^2$ . In a single classification on this advantage in speed is minor, but in the more complex classifications, the analysis of variance, is the only simple and reliable method of determining the appropriate pooled error variances  $S^2$ .
- (2) It provides a new test, the F-test. This is a single test of null hypothesis that the population means  $\mu_1, \mu_2, \mu_3, \mu_4$ , for the four fats are identical. This test is often useful, in a preliminary inspection of the results and has many subsequent applications. We can compute the total sum of squares of deviation for the 24 observations as

$$64^2 + 72^2 + 68^2 + \dots \dots \dots 70^2 + 68^2 - \left( \frac{1770}{24} \right)^2$$

$$= 134,192 - 130,538 = 3654. \quad \dots \dots \dots 1.1$$



This sum of squares has 23 degrees of freedom. The mean square,  $3654/23 = 158.9$ , is the first estimate of  $\sigma^2$ .

The second estimate is pooled  $S^2$  already obtained. Within each fat, we computed the sum of squares between batches (890, 302 etc.), each with .5 d.f. These sum of squares were added to give

$$890 + 302 + 488 + \dots = 2018 \text{ 1.2}$$

Source of variation	Degrees of Freedom	Sum of Squares
Between fats	3	1,636
Between batches within fats	20	2,018
Total	23	3,654

The degrees of freedom and the sum of squares for the two components (between fats and within fats) add to the corresponding total figures. These results hold in any single, classification, the result for the difference is not hard to verify with classes and  $n$  observations per class, the  $d.f.$  are  $(a-1)$  for between fats and  $a(n-1)$  for with in fats, and  $(an-1)$  for the total. But

$$(a-1) + a(n-1) = a-1 + an-a-1 = an-1$$

The standard practices in the analysis of variance is to compute only the total sum of squares and the sum of squares between fats. The sum of squares within fats, leading to the pooled  $s^2$  is obtained by subtraction.

The symbol  $T$  denotes a typical class' total, while  $G = \sum T = \sum \sum X$  (summed over both rows and columns) is the grand total. The first step is to calculate the correction for the mean,

$$C = \frac{G^2}{an} = \frac{(1770)^2}{24} = 130.538.$$

This quantity is called the sum of squares between batches with in fats, more concisely the sum of squares within fats. The sum of squares is divided by its degrees of freedom, 20 to give me second estimate  $s^2 = 2,018/20 = 100.9$ .

For the third estimate, consider the mean for the four fats, 72, 85, 76 and 62. These are also estimates of  $\mu$  but have variances  $\frac{\sigma^2}{6}$  since they are means of samples of 6, their sum of squares of deviations is

$$76^2 + 85^2 + 76^2 + 62^2 - \frac{(295)^2}{4} = 272.75$$

with 3 degrees of freedom. The mean square,  $\frac{272.75}{3}$  is an estimate of  $\mu$  consequently if we multiply by 6, we have the third estimate of  $\mu$  we shall accomplish this by multiplying the sum of squares by 6, giving.

$$6 \{72^2 + 85^2 + 76^2 + 62^2 - \frac{(295)^2}{4}\} = 272.75\} = 1636 \quad \dots\dots 1.3$$

the mean square being  $\frac{1636}{3} = 545.3$ .

Since the total for any fat is six times the fat means, this sum of squares can be computed from the fat totals as

$$\frac{432^2 + 510^2 + 456^2 + 372^2}{6} - \frac{(1770)^2}{24}$$

$$= 132, 174 - 130, 538 = 1656 \quad \dots\dots 1.4$$

This sum of square is called the sum of squares between fats.

This is done because c occurs both in formula 1.1 for the total sum of squares and in formula for the sum of squares between fats. The remaining steps should be clear from Table 3.

**Formula for calculate the Analysis of variance Table**

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square
‘Between classes (fats)	a - 1 = 3	$\left(\frac{\Sigma T^2}{n}\right) - C = 1.636$	543.3
Within classes (fats)	a (n-1) = 20	Subtract 2,018	100.9
Total	an - 1 = 23	$\Sigma \Sigma X^2 - C = 3,654$	

#### The variances ratio F

$$F = \frac{\text{Mean square between classes}}{\text{Mean squares within classes}}$$

should be good criterion for testing the null hypothesis that the population means are the same in all classes. The value of F should be around 1 when the null hypothesis holds, and should become large when the  $r_i$  differ substantially. The distribution was first tabulated by fisher in form  $2 = \log e \sqrt{F}$ . In honour of Fisher, the criterion was named F by Snedecor. Fisher and Yates designate F as the variance ratio, It should be noted that when there are only two classes, the F test is equivalent to the t test, which is used to compare the two means. With two classes, the relation  $F = t^2$  holds.

**An experiment comparing two groups of equal size.** Mr. X compared the 15 day mean comb weights of two lots of male chicks, one receiving hormone A, the other C. Day old chicks, 11 in number, were assigned at random to each of the treatments) To distinguish between the two lots, which were caged together, the heads of chicks were stained green and black respectively. The individual comb weights are given in Table. 1 Prove that Hormone A gives higher average comb weight than hormone C.

**Table 1**  
**Testing the Differences Between the Means of Two Independent Samples**  
**Weights of Comb (rags)**

	<b>Harmone</b>	<b>Harmone</b>
	<b>A</b>	<b>C</b>
	57	89
	120	30
	101	82
	137	50
	119	39
	117	22
	104	57
	73	32
	53	96
	68	31
	118	88
Total	1,067	616
N	11	11
Mean	97	56
$\Sigma X^2$	111,971	42,244
$\frac{(\Sigma X)^2}{n}$	103,499	34,496
$S^2 = \left  \Sigma x^2 - \frac{(\Sigma X)^2}{N} \right $	8,472	7,748
d.f.	10	10

$$\text{pooled } s^2 = \frac{8472 + 7748}{10 + 10} = 811 \text{ d.f} = 20$$

$$S_{x_1-x_2} = \sqrt{\frac{252}{n}} = \sqrt{\frac{2 \times 811E}{11}} = 12.14\text{mg.}$$

$$t = \frac{(X_1 - X_2)}{S_{X_1} - x_2}$$

$$= \frac{41}{12.14} = 3.38$$

Analysis of Variance with only two classes.

The period  $S^2 = \frac{16220}{2} = 811$ , has already been computed. To complete the analysis of variance, compute the between samples sum of squares. Since the sample totals were 1067 and 616, with  $n = 1$ , the Sum of squares is

$$\frac{(1067)^2 + (616)^2}{11} - \frac{(1683)^2}{22} = 9245.5$$

With only two samples, this sum of squares is obtained more quickly as

$$\frac{(\Sigma x_1 - \Sigma x_2)^2}{2n} = \frac{(1067 - 616)^2}{2 \times 11} = 9245.5$$

#### Analysis of Variance

Source of variation	Degrees of freedom	Sam of square	Mean square
Between samples	1	9,245.5	9,245.5
within samples	20	16.2200	811.0

$$F = \frac{9245.5}{811.0} = 11.40 \quad \sqrt{F} = 3.48$$

The value of F is significant at 1% level showing that Harmonic A gives higher average comb weights than harmonic C. The two sum of squares of deviations 8,472 and 7,748, make the assumption of equal  $\sigma^2$  appear reasonable.

The 95% confidence limit for  $(\mu_1 - \mu_2)$  are

$$x_1 - x_2 \pm t_{0.005} S_{x1-x2}$$

or, in this example

$$41 - (2.086)(12.14) = 16\text{mg.}$$

and  $41 + (2.086)(12.14) = 66\text{mg.}$

**Example :** Below are given the yield of three strains of rice planted in five randomized blocks. Prepare the table of analysis of variance.

Blocks					
Strains	I	II	III	IV	V
A	20	21	23	16	20
B	18	20	17	15	25
C	25	28	22	28	32

We set up the hypothesis that there is no difference between the strains. Taking 20 as the origin, the given table becomes.

Blocks						
Strains	I	II	III	IV	V	Total
A	0	1	3	-4	0	0
B	-2	0	-3	-5	5	-5
C	5	8	2	8	12	35
Total	3	9	2	-1	17	30

Here  $T = 30$

$$\sum \sum ij^2 = 4 + 25 + 1 + 64 + 9 + 9 + 4 + 16 + 25 + 64 + 144 = 390$$

$$N = 15$$

Sum of squares between strains

$$= \frac{\sum T_i^2}{n_j} = \frac{T^2}{n}$$

$$= \frac{0 + 25 + (122)^2}{5} - \frac{(30)^2}{15}$$

$$= 250 - 60 = 190$$

Total sum of squares

$$= \sum \sum u_{ij}^2 - \frac{T^2}{N} = 390 - \frac{(30)^2}{15} = 330$$

Sum of squares within strains

$$= \text{Total S.S} - \text{S.S between strains}$$

$$= 330 - 190 = 140$$

**Analysis of Variance Table**

Source of variation	D.F.	S.S.	M.S.	F	F at Level	
					1 %	5%
Between strains	2	190	95	8.14	693	389
Within strains	12	140	11.67			
Total	14	-	-	-	-	-

$$M.S. = \frac{S.S}{D.F}$$

The observed F being greater than the value of F at, 1% for 2, 12, d.f. we reject our hypothesis  
The difference between the strains is significant.

## Two criteria of classification

We classify blocks not only, according to type of soil but also according to the type of seed used on them. The question then arises whether the crop yield varies with the type of block or with the type of seed. Here the variations in the yield of crop may be attributed to

- (i) Variation in crop yield due to type of soil.
- (ii) Variation in crop yield due to type of seed.
- (iii) Variation in crop yield due to error term.

For carrying out the analysis of variance, we could compute total variations and variation between columns (type of soil). There is no variation with the columns (type of soil). But there is a variation between rows (variation due to type of seed) and residual variation :

1. Total variation (Total sum of square's)

$$\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{N}$$

2. Variation between columns (S.S between, columns)

$$\sum_1^c [N_c(\bar{X}_c - \bar{X})^2] = \sum_1^c \left[ \frac{\left( \sum_i^n x \right)^2}{N_c} - \frac{(\sum X)^2}{N} \right]$$

Where  $\sum_1^c$  stands for summation of 'C' columns

$\sum_1^m$  stands for the summation of N items in a column

$N_c$  stands for the no of items in a column

$\bar{X}_c$  stands for the column mean

$\bar{X}$  stands for the grand mean

3. Variation between rows (sum of squares between rows)

$$\sum_1^R N_r(\bar{X}_r - \bar{X})^2 = \sum_1^R \left[ \frac{\left( \sum_1^n x \right)^2}{N_r} - \frac{(\sum x)^2}{N} \right]$$

where

$\sum_1^R$  stands for the summations of 'R' rows.

$\sum_1^N$  stands for the summation of 'N' item's in a row.

Nr stands for the no of items in a row.

$\bar{X}_r$  stands for the mean of row

4. Residual variation = Total variation - Variation between columns Variation within rows.

**Table of Analysis of Variance - Two criteria of classification**

Sources of Variations (1)	Sum of Squares (2)	Degrees of Freedom (3)	Estimates of Variance (4)	Variance Ratio-I (5)
Bet columns	S.S.C.	C-1	$\frac{SSC}{C-1} = m_1$	$F_1 = \frac{m_1}{m_3}$
Bet Rows	S.S.R.	R-1	$\frac{SSC}{C-1} = m_2$	$F_1 = \frac{m_2}{m_3 m}$
Residual	S.S.Re.	(C-1) (R-1)	$\frac{SSRe}{(C-1)(R-1)} = m_3$	
Total	T.S.S.	N-1	-	-

**Example :** Three varieties A, B, C, of a crop are tested in a randomized block design with four replications, the lay out being given in diagram appended. The plot yield in kgs. are also indicated therein. Analyze the experimental yield and state your conclusions.

A	6	C	5	A	8	B	9
C	8	A	4	B	6	C	9
B	7	B	6	C	10	A	6

**Solution**

Blocks					
Varieties	I	II	III	IV	Total
A	6	4	8	6	24
B	7	6	6	9	28
C	8	5	10	9	32
Total	221	15	24	24	84

$$T = 84$$

$$\Sigma \Sigma x_{ij}^2 = 36 + 49 + 64 + 16 + 36 + 25 + 64 + 36 + 100 + 36 + 81 + 81 = 624$$

$$\text{Correction Factor} = \frac{T^2}{n} = \frac{(84)^2}{12} = 588$$

$$\text{Total sum of squares} = 624 - 588 = 36$$

Sum of squares between varieties

$$= \frac{\sum V_i^2}{4} - \frac{T^2}{N}$$

$$= \frac{24^2 + 28^2 + 32^2}{4} - 588$$

$$= 596 - 588 = 8$$

$$\text{Sum of squares between blocks} = \frac{\sum \beta_i^2}{3} - \frac{T^2}{n}$$

$$= \frac{21^2 + 15^2 + 24^2 + 24^2}{3} - 588$$

$$= 606 - 588 = 18$$

$$\therefore \text{Error sum of squares} = 36 - (18 + 8) = 10$$

Analysis of Variance table					
Sources of Variation	D.F.	S.S.	M.S.	F	F at 5% level
				5.143	4.757
Between varieties	2	8	4	2.4	
Between Blocks	3	18	6	3.6	
Error	6	10	1.667		
Total	11				

Since the calculated value of F is less than the table value in both cases, we conclude that the variation between the varieties and between blocks are significantly not different from the variance due to random errors.

### Comparison of Regression Analysis and Analysis of Variance

As has been explained in the beginning that the basic difference between regression analysis and ANOVA is that while the former provides numerical values for the influence of the various explanatory factors on the dependent variable, in addition to the information concerning the breaking down of the total variance of y into additive components, while the latter provides only the breaking down of the total variance into additive components.

*Firstly* in both methods the total variation in y is split into two additive components:

(a) *Regression analysis*

$$\sum Y^2 = \sum \hat{Y}^2 + \sum e^2$$

Total variation = (Explained by regressors) + (unexplained or residual)



(b) Analysis of variance

$$\sum_j^k \sum_i^{n_j} (Y_{ji} - \bar{Y})^2 = \sum_j^k n_j (\bar{Y}_j - \bar{Y})^2 + \sum_j^k \sum_i^{n_j} (Y_{ji} - \bar{Y}_j)^2$$

Total variation = Between + Within

*Secondly.* The test performed in the method of analysis of variance concerns the equality between means of sub-groups of sub-samples of an enlarged population. That is, the null hypothesis being tested is

$$H_0 = \mu_1 = \mu_2 = \dots \mu_n$$

and alternative hypothesis is

$$H_1 : \mu_1 \text{ not all equal.}$$

The  $F^*$  ratio is a test of significance of  $R^2$

$$F^* = \frac{\Sigma \hat{Y}^2 / K - 1}{\Sigma e^2 / N - K} = \frac{R^2 YX_1 / K - 1}{(1 - R^2 YX_1 / N - K)}$$

If  $R^2$  is not statistically significant it means that there is no linear relationship between Y and X.

*Thirdly.* In both methods we obtain an analysis of variance table, from which F ratios, can be computed and used for testing 'hypothesis related to the aim of study.

*Fourthly :* As has been proved earlier that the individual regression coefficients that t and F tests are formally equivalent, the relationship between them being.

*Lastly.* Regression analysis is more powerful method than the ANOVA method when studying economic relationship from market data which are not experimental but stochastic: It is generally believed that ANOVA method is more appropriate for the study of the influence of qualitative factors on a certain Variable, because qualitative variables do not possess numerical values, and hence their influence, cannot be assessed by regression analysis, while the ANOVA technique does not require information of the values of X's but it is based solely on the values of Y. This argument loses its power due to the increasing use of dummy variables in regression analysis. However, the ANOVA technique may be incorporated into regression analysis, for carrying out tests of various hypothesis.

**Merits**

- (1) The calculations are comparatively simple in ANOVA than in regression analysis.
- (2) The technique of analysis of variance makes use of all data.
- (3) It does not involve a high degree of abstraction.
- (4) It may be applied to all important planning phase of the enquiry as well as to the interpretive phase of enquiry.

## SUGGESTED READINGS

1. Murray it Spiegel : *Theory and Problem of Statistics.*
2. P.S. Grewal : *Method of Statistical Analysis* (Starting Publishers, Pvt, Ltd, New Delhi).
3. Olive Jean Dunn & Virginia A Clak : *Applied Statistics Analysis of Variance and Regression* (John Wiley & Sons, New York)
4. Harold Cramer : *Mathematical Methods of Statistics* (Asia Publishing House. New Delhi)
5. George W. Snedecor and William G. Cochran. : *Statistical Methods* (Oxford & IBH Publishing Co. New Delhi)
6. Alpha C. Chiang : *Fundamental Methods of Mathematical Economics* (McGraw-hill).

\*\*\*\*\*

## **LESSON-17**

### **INDEX NUMBERS**

Dear Student,

Index numbers are the indicators which reflect changes over a specified period of time in (i) prices of different commodities, (ii) industrial production, (iii) sales, (iv) imports and exports, (v) cost of living etc. These indicators are of paramount importance to the management personnel of any government organization or industrial concern, for the purpose of reviewing position and planning action if necessary and in the formulation of executive decisions. They reflect the pulse of an economy and serve as indicators of inflationary or deflationary tendencies. Economic index numbers measure the pressure of economic behaviour and are rightly termed as economic barometers or ‘barometers of economic activity’, since a look at some of the important indices like index numbers of whole sale prices, industrial production, agricultural production etc; gives a fairly good idea as to what is happening to the economy of a country.

“Index numbers as statistical devices designed to measure the relative change, in the level of a phenomenon (variables or a group of variables) with respect to time, geographical location or other characteristics such as income profession etc.” In other words, these are the numbers which express the value of a variable at any given date called the given period as a percentage of the value of that variable at same standard date called the base period. The variable may be:

- (i) The price of a particular commodity e.g. silver, iron etc. or a group of commodities like consumer goods, foodstuffs, etc.
- (ii) The volume of trade, exports and imports, agricultural or industrial production, sales in a departmental store etc.
- (iii) The national income of a country or cost of living of persons belonging to particular income group/profession etc.

For example, suppose we want to measure the general changes in the price level of consumer goods. Obviously, these changes are not directly measurable as the price quotations of various commodities are available in different units, e.g. wheat and sugar in Rs. per quintal, petrol and kerosene oil in Rs. per liter, cloth in Rs. per metre etc. An average price of all these items expressed in different units is obtained by using the technique of index numbers.

#### **Problems Involved in the Construction of Index**

The methods of Construction of index numbers warrant a careful study of the following problems:

##### **1. The Purpose of Index Number:**

An index number which is properly designed for a purpose can be most useful and powerful tool, otherwise it can be equally misleading and dangerous. Thus the first and foremost problem is to determine the purpose of index number without which it is not possible to follow the steps in its construction.

##### **2. Selection of Commodities :**

Having defined the purpose of index numbers, select only those commodities which are relevant to the index. For example, if the purpose of an index is to measure the cost of living of low

income group (poor families) we should select only those commodities or items which are consumed/ utilized by persons belonging to this group and due care should be taken not to include the goods/ services which are, ordinarily consumed by middle-income or high income group. For such an index, selection of commodities like cosmetics and other luxury goods like scooters cars; refrigerators, television sets etc, will be absolutely useless.

### **3. Data for Index Numbers:**

The data, usually the set of prices and of quantities consumed of the selected commodities for different periods, places, etc., constitute the raw material for the construction of index numbers. The data should be collected from reliable sources such as standard trade journals, official publications periodical special reports from the producers, exporters, etc; or through field agency. The principles of data collection, viz., accuracy, comparability, sample representativeness and adequacy should be borne in mind. In any case the data should strictly pertain to what is being measured.

### **4. Selection of Base Period :**

The period, with which the comparisons of relative changes in the level of a phenomenon are made is termed as base period and the index for this period is always taken as 100. The following are basic criteria for the choice of the base period.

- (i) The 'base period' must be a 'normal period', i.e. a period free from all sorts of abnormalities or chance fluctuations such as economic boom or depression, labour, strikes, wars, floods, earthquakes etc. If the base period be taken as a period of economic instability or depression in which the prices of various commodities and goods, due to their scarcity have been abnormally high then the comparison of price relatives in any given year will not be of much practical utility.
- (ii) The base period should not be too distant from the given period. Since index numbers are Essential tools in business planning and in formulation of executive decisions, the base period should not be too far back in the past relative to the given period because due to dynamic pace of events these days, distant base period is likely to be entirely different from the given period. Moreover, if the base year is shifted far away from the given period, it is possible that the pattern of consumption of commodities may change appreciably.

### **5. Type of Average to be used :**

Since index numbers are specialized averages a judicious choice of average to be used in their construction is of great importance. Usually the following averages are used :

- (i) Arithmetic Mean (A.M.) : Simple or weighted,
- (ii) Geometric Mean (G.M): Simple or weighted,
- (iii) Median.

Since in the construction of index numbers we deal with ratios or relative changes and since geometric mean gives equal weights to equal ratios of change, does not give undue weightage to extreme observation, Geometric mean is preferred over all other averages.

### **6. Selection of Appropriate weights:**

Generally, various items/commodities say wheat, rice, kerosene, clothing, etc, included in the index are not of equal importance and proper weights should be attached to them to take into account their relative importance. There are two types of indices :

- (i) 'Unweighted Indices', in which no specific, weights are attached to various commodities, and
- (ii) 'Weighted indices', in which appropriate weights are assigned to various items.

## **7. Choice of Formulae :**

A large number of formulae have been devised for constructing the index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index number but also on the data available. Fisher has suggested that an appropriate index is that which satisfies time reversal test and factor reversal test. Theoretically Fisher's method is considered as "Ideal" for constructing index numbers.

## **CLASSIFICATION OF INDEX NUMBERS**

Index numbers may be classified into three categories depending on the nature of phenomena under study.

### **1. Price Index Numbers :**

Price index numbers show the changes in the prices of commodities produced or consumed in a given period with reference to base period. These indices permit comparison of the prices, of commodities between regions, between cities of the same region and between two time periods. These are of two types :

- (a) Whole sale price index numbers.
- (b) Retail or Consumer price index numbers.

Quantity index numbers show the changes in the quantity of goods produced or consumed or purchased in a given period with reference to base period. These indices permit comparison of the quantity produced or purchased of different commodities.

### **3. Value Index Numbers**

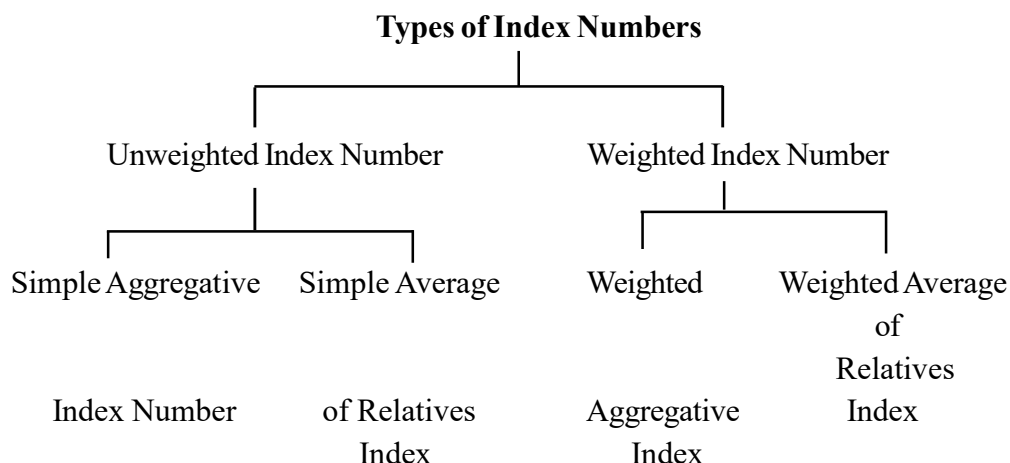
Value index numbers show, the changes in the value of commodities in a given period with reference to base period.

A large number of formulae have been devised for constructing index numbers. Broadly speaking they can be grouped under two heads :

- (a) Unweighted Indices: and
- (b) Weighted indices

In the unweighted indices weights are not expressly assigned whereas in the weighted indices weights are assigned to the various items. Each of these types may be further divided under two heads :

- (i) Simple Aggregative and
- (ii) Simple Average of Relatives.



### Notations and Terminology:

The notations and terminology used in constructing index numbers are given below:

Base year—The year selected for comparison. It is denoted by the suffix zero, ‘0’.

Current year—The year for which comparison is sought or required. It is denoted by the suffix ‘1’.

$P_0$ -Price of a commodity in the base year.

$P_1$ -Price of a commodity in the current year.

$Q_0$ -Quantity of a commodity in the base year.

$Q_1$ -Quantity of a commodity in the current year.

W-Weight assigned to a commodity according to its relative importance in the group.

$P_{01}$  - Price index number for the current year.

$P_{10}$ -Price Index number for the base year.

$Q_{01}$ -Quantity index number for fee current year.

$Q_{01}$ -Quantity index number for the base year.

The various methods of constructing index numbers are discussed below.

#### 1. Simple Aggregative Index:

Under this method, the total of current year price for the various commodities or items is divided by the total of base year prices and the quotient is multiplied by 100.

Symbolically

$$P_{10} = \frac{\sum p_1}{\sum p_0} \times 100$$

#### Example.

With the data given below construct price index for 1988 taking 1986 as base.

Commodities	P rice	
	1986	1988
A	50	70
B	40	65
C	80	95
D	110	130
E	20	18
F	15	14
G	10	12

We have,  $\Sigma p_0 = 325$  and  $\Sigma p_1 = 404$

Price index for 1988.

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{404}{325} \times 100$$

$$= 124.31$$

Hence, there has been 24.31 percent increase in prices of commodities in 1988 as Compared to 1986 prices.

### Simple Average of Relatives Index :

The average of relative price index is the ratio expressed in percentages of the price of a commodity in given period called current period to its price in another period called the base period. If  $P_0$  and  $P_1$  represent the price of commodity during the base period and the given period (current period) respectively then by definition:

$$\text{Price Relatives} = \frac{P_1}{P_0}$$

The relatives get by their average either by the arithmetic mean method or geometric mean method By the arithmetic, mean method, the average of price relatives index is:

$$P_{01} = \frac{\Sigma \left( \frac{P_1}{P_0} \times 100 \right)}{N}$$

Where N refers to the number of items (commodities) whose price relatives are thus averaged. When geometric mean, is used for averaging the price relatives, the formula for obtaining the Index become

$$\log P_{01} = \frac{\Sigma \log \left[ \frac{P_1}{P_0} \times 100 \right]}{N} \text{ or } \frac{\Sigma \log P}{N} \text{ where } P = \frac{P_1}{P_0} \times 100$$

$$\text{or } P_{01} = \text{antilog} \left[ \frac{\left( \frac{\sum \log \frac{P_1}{P_0} \times 100}{N} \right)}{N} \right] = \text{antilog} \frac{\sum \log P}{N}$$

**Example :**

From the data given below, construct index for 1988 taking 1987 as base year, by the average or relatives method using (a) arithmetic, mean and (b) geometric mean methods.

Commodities	Price 1987	Price 1988
A	70	90
B	60	95
C	123	135
D	150	180
E	20	20
F	15	16

**Solutions.**

(a) Construction of index number by arithmetic mean method.

Commodities	Price 1987 ( $P_0$ )	Price 1988 ( $P_1$ )	Price relatives ( $P_1/P_0 \times 100$ )
A	70	90	128.57
B	60	95	158.33
C	120	135	112.50
D	150	180	120.00
E	20	20	100.00
F	15	16	106.67

$$\sum P_1/P_0 \times 100 = 726.07$$

$$\begin{aligned} \text{Price Index for 1988 } P_{01} &= \frac{\sum p_1 / p_0 \times 100}{N} \\ &= \frac{726.07}{6} = 121.01 \end{aligned}$$



(b) Construction of index number by geometric mean method.

Commodities	Base year Price ( $P_0$ )	Current year Price ( $P_1$ )	Price relative ( $p_1/p_0 \times 100$ )	Log P
A	70	90	128.57	2.1091
B	60	95	158.33	2.1996
C	420	135	112.50	2.0511
D	150	180	120.00	2.0792
E	20	20	100.00	2.0000
F	15	16	106.67	2.0280

$$\Sigma \log p = 12.4670$$

$$P_{01} = \text{Antilog } \frac{\Sigma p_1 / p_0}{N} \times 100$$

$$\begin{aligned} \text{or } P_{01} &= \text{Antilog } \Sigma \log p / N \\ &= 12.4670 / 6 = \text{antilog } 2.0778 \\ &= 119.62 \end{aligned}$$

### Weighted Index Numbers

The index numbers constructed by the methods of simple aggregative and simple average of relatives are unweighted in the true sense of term. An equal importance assigned to all items is included in the index computed by the above two methods. In other words, implicit weight is given, in the sense each price is assumed to be of equal importance. Implicit weighting is far from realistic in many cases; Construction of indices which are said to be useful, requires a conscious and obvious effort to assign to each commodity or item a weight in accordance with the importance in the total phenomena so as to describe the activity. Index numbers with weight are of two types, namely weighted aggregative index and weighted average of relative index.

The index numbers calculated by the weighted aggregative method are of the simple aggregative type with the fundamental difference that weights are assigned to the items included in the index. In other words, the base period quantities or current period quantities of commodities involved in the computation are taken as weights, there are different methods of assigning weights; and as such a large number of formulae for constructing index numbers have been devised, of which some important or popular ones are :

1. Laspeyres method
2. Paasche method
3. Darbish and Bowley's method
4. Fisher's Ideal method
5. Marshal Edgeworth method
6. Kelly's method.

### 1. Laspsype's Method :

In this method weights are determined by quantities in the base period. This method is widely used because it keeps the quantities consumed in the base period as constant in the current period also, and finds the change in aggregate value. The main limitation of this method is that, it works under assumption that there is no change in quantities of the current year. The formula for constructing the index is

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Where  $P_{01}$  = price index,  $p_0$  = price in the base year,  $p_1$  = prices in the current year,  $q_0$  = quantity in the base year and  $q_1$  quantity in the current year.

### 2. Paasche Method :

In this method the current year quantities are taken as weigths. The formula for constructing the index is :

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

*i.e.* multiply current year prices of various commodities. with current year weights and obtain  $\sum p_1 q_1$  also multiply the base year prices of various commodities with current year weights and obtain  $\sum p_0 q_1$ . Divide  $\sum p_1 q_1$  by  $\sum p_0 q_1$  and multiply the quotient by 100.

### 3. Darbish and Bowley's Method :

Darbish and Bowley have suggested a practical method of constructing index using weighted aggregates by combining the Lspeyre's and Paasche's methods. In other words, under this method, the index is the simple arithmetic mean of the Laspeyre's and Paasche method. This method takes into account the influence of both the periods i.e., current as well as base period. The formula for constructing this index is :

$$P_{01} = \frac{L + P}{2}$$

where L = Laspeyris Index and P = Paasche Index

or

$$P_{01} = \frac{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}}{2} \times 100$$

### 4. Fisher's ideal method:

Living Fisher has successfully combined the Laspeyre's and Paasche's methods to develop ideal index. According to him, an ideal i.e. true or unbiased index number must satisfy two basic tests, namely time reversal and factor reversal tests. The geometric mean of the Laspeyre's and Paasche methods satisfies these tests. Hence Fisher consider his method of constructing index number as an ideal one.

The Fisher's Ideal Index is given by the formula:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P}$$

This method is assumed to be free from bias and takes into account both the constant and fluctuating weights.

### 5. Marshall-Edgeworth Method ?

In this method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is :

$$P_{01} = \frac{\sum (p_0 + q_1) P_1}{\sum (p_0 + q_1) P_0} \times 100$$

or

$$P_{01} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

### 6. Kelly's Method:

T.L. Kelly has suggested the following formula for constructing index number

$$P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

Here weights are the quantities which may refer to some period, not necessarily the base year or current year. Thus, the average quantity of two or more years may be used as weights. If in the Kelly's formula, the average of the quantities of two years is used as weights, the formula becomes.

$$P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100 \quad \text{where} \quad q = \frac{q_0 + q_1}{2}$$

Similarly the average of the quantities of three or more years can be used as weights. This method is known as fixed weight aggregative index. An important advantage of this formula is that like Laspeyres or Paasche's index, it does not demand yearly change necessitating corresponding change in the weights.

### Weighted Average of Relatives:

In the weighted aggregative methods discussed above price relatives were not computed. However, like unweighted, relative method it is possible to compute weighted average of relatives. For purposes of averaging we may use either the arithmetic mean or the geometric mean. In order to compute the weighted arithmetic mean of the relatives. Express each *item of the period for* which the index number is being calculated as a percentage of the same item in the base period. Then multiply the percentage as obtained for each item by the weight which has been assigned to that item. Symbolically the index number is

$$P_{01} = \frac{\sum PV}{\sum V}$$

where P = Price relative and

V = a Value weights i.e.  $p_0 q_0$

Instead of using arithmetic mean, the geometric mean may be used for averaging relatives. The weighted geometric mean of relatives is computed in the same manner as the unweighted, geometric mean of relative index number except that weights are introduced by applying them to the logarithms of the relatives. When this method is used the formula for computing the index is:

$$P_{01} = \text{Antilog} \left[ \frac{\sum V \log P}{\sum V} \right]$$

where  $P = \frac{P_1}{P_0} \times 100$

and  $V = \text{Value weight i.e. } P_0 q_0 \text{ for each item.}$

**Example:**

From the following data compute price index by applying weighted average of price relatives method using:

(a) arithmetic mean, and

(b) geometric mean.

Commodities	$P_0$ (Rs)	$q_0$ (kg.)	$P_1$ (Rs.)
Sugar	3.0	20	4.0
Flour	1.5	40	1.6
Milk	1.0		101.5

**Solution:**

(a) Index Number Using Weighted Arithmetic Mean of price Relatives

Commodities	$P_0$	$q_0$	$P_1$	$T = P_0 q_0$	$P = \times 100$	PV
Sugar	Rs.3.0	20 kg	Rs. 4.0	60	$4/3 \times 100$	8,000
Flour	Rs.1.5	50 kg	Rs. 1.6	60	$1.6/1.5 \times 100$	6,400
Milk	Rs. 1.0	10kg	Rs.1.5	10	$1.5/1.0 \times 100$	1,500
				$\sum V = 130$	$\sum PV = 15,900$	

$$P_{01} = \frac{\sum PV}{\sum V} = \frac{15,900}{130} = 122.31$$

This means that there has been a 22.3 per cent increase in prices over the base level.

(b) Index Number Using Geometric Mean of Price Relatives

Commodities	$P_0$	$q_0$	P	V	$P_1$	Log P	V Log P
Sugar	Rs. 3.0	20kg	Rs.4.0	60	133.3	2.1249	127.494
Flour	Rs/ 1.5	40kg	Rs. 1.6	60	106.7	2.0282	121.692
Milk	Rs. 10	10kg	Rs. 1.5	10	150.0	2.1761	21.761
				$\sum V = 130$	$\sum V \log P = 270.947$		

$$P_{01} = \left[ \frac{\Sigma V \cdot \log P}{\Sigma V} = \text{Antilog} \left[ \frac{270.947}{130} \right] \right]$$

$$= \text{Antilog } 2.084 = 121.3$$

### **Suggested Readings**

1. Nagar, A.L & Das R.K. : Basic Statistics, Chapters 10 to 10.6.
2. Gupta, S.P. : Statistical Methods, Chapter 13.
3. Grewal, P.S. : Methods of Statistical Analysis, Chapter 16.
4. Gupta S.C. & V.K. Kapoor : Fundamentals of Applied Statistics, Chapter 3.

\*\*\*\*\*

## LESSON-18

### TESTS FOR CONSISTENCY OF INDEX NUMBER

A large number of formulas have been devised by different statisticians for constructing index numbers and the problem is that of selecting the most appropriate in a given situation. Prof. Irving Fisher, a famous statistician, has laid down two tests for a good index number. The tests are:

#### 1. Time Reversal Test:—

According to this test, an index number should show the same relative movement from one period to another whichever may be taken as base. In other words, an index number should work both ways as well as backward. An index number for current year on the basis of base year should be reciprocal of the index number for base year on the basis of current year.

In the words of Fisher, “The test is that the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base.” When the time periods are reversed, one index number is the reciprocal of the other index and their product is always equal to one. Let  $P_{01}$  be an index number for the current year (1) based on base year (0) and  $P_{10}$  be an index number for base year (0) based on current year (1) then :

$$P_{01} \times P_{10} = 1$$

If the product of two indices is not equal to unity then there is a bias in the formula being used. Below we discuss which formula satisfies time reversal test:

#### (a) Laspeyres Formula:

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

Changing 1 to 0 and 0 to 1, we get

$$P_{10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}$$

$$P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1.$$

The product of two indices is not equal to unity, therefore, this formula does not satisfy time reversal test:

#### (b) Paasche Formula:

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

Changing 1 to 0 and 0 to 1, we get :

$$P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

$$P_{01} \times P_{10} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0} \neq 1.$$

Like the Laspeyres formula, this formula also does not satisfy time reversal test

**(c) Fisher's Formula :**

$$P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

Changing 0 to 1 and 1 to 0, we get:

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_0}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = 1.$$

Thus this formula satisfies the time reversal test and is called the ideal formula. There are five methods which do satisfy the test :

- (1) The Fisher's Meal Formula
- (2) Simple Geometric Mean of price relatives
- (3) Aggregates with fixed weights.
- (4) The weighted geometric mean of price relatives if we used fixed weights.
- (5) Marshall-Edgeworth method.

**2. Factor Reversal Test :**

According to this test, the product of the price index and quantity index should be equal to the corresponding value index.

In symbols:

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

In the words of Fisher. "Just as our formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results *i.e.* the two results 'multiplied together should give the true value ratio.'" Put it in other words, the change in price multiplied by the change in quantity should be equal to the total change in value. The total value of a given commodity in a given year is the product of the quantity and the price per unit (value =  $p \times q$ ). If  $p_1$  and  $p_0$  represent prices and  $q_1$  and  $q_0$  the quantities in the current year and the base year respectively; and if  $P_{01}$  represents the change in price and  $Q_{01}$ , represents the change in the current year then

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

In simple, words the tests is satisfied if the product of the price index and the quantity index computed from same data is equal to the ratio of the aggregate value in the current year to aggregate value in the base year. In case the product is not equal to the value index, there is a bias in the formula being used. Below we examine which formula satisfies this test :

**(a) Laspayres Formula :**

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

and  $Q_{01} = \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \neq \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Thus this formula does not satisfy the factor reversal test :

**(b) Paasche Formula :**

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$$

and  $Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}$

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0} \neq \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Thus this formula also does not satisfy factor reversal test.

**(c) Fisher's Formula:**

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

and  $Q_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$

(by changing p to q and q top)

$$\begin{aligned} \text{No'v } P_{01} \times Q_{01} &= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}} \\ &= \sqrt{\left( \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \right)^2} \\ &= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \end{aligned}$$

Thus Fisher's formula satisfies factor reversal test.



**Example:**

The following figures relate to the prices and quantities of certain commodities. Construct an appropriate index, number and show if it satisfies the time reversal test.

	1973		1974	
Commodities	Price	Quantities	Price	Quantities
A	30	50	32	50
B	25	40	30	35
C	18	50	16	55

Index Number by Fisher's Ideal Method

	1973		1974					
Commodities	$p_0$	$q_0$	$p_1$	$q_1$	$p_1q_0$	$p_0q_0$	$p_1q_1$	$p_0q_1$
A	30	50	32	50	1600	1500	1600	1500
B	25	40	30	35	1200	1000	1050	875
C	18	50	16	55	800	900	880	990
					$\Sigma p_1q_0$ =3600	$\Sigma p_0q_0$ =3400	$\Sigma p_1q_1$ =3530	$\Sigma p_0q_1$ =3365

$$\begin{aligned}
 P_{01} &= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 \\
 &= \sqrt{\frac{3600}{3400} \times \frac{3530}{3365}} \times 100 \\
 &= \sqrt{1.111} \times 100 = 1.054 \times 100 = 105.4
 \end{aligned}$$

Time reversal test is satisfied when  $P_{01} \times P_{10} = 1$

Substituting the values of  $\Sigma p_1q_0$ ,  $\Sigma p_0q_0$  etc;

$$P_{01} = \sqrt{\frac{3600}{3400} \times \frac{3530}{3365}}$$

$$P_{10} =$$

Hence time reversal test is satisfice the proof of the formula.

Test the adequacy of index by the reversal factor and factor reversal for the following data:

Items	1983		1988	
	Price (Rs.)	Quantity (kg.)	Price (Rs.)	Quantity (kg.)
Rice	3	5	4	6
Oil	24	1	19	1
Tea	6	1	7	1
Washing Powder	4		5	1
Sugar	3	4	5	5
Milk	2	2	3	3

**Solution:**

First, compute the required values.

Items	1983		1988					
	$p_0$	$q_0$	$p_1$	$q_1$	$p_1q_0$	$p_1q_1$	$p_0q_0$	$p_0q_1$
Rice	3	5	4	6	20.0	24.0	15.0	18.0
Oil	24	1	19	1	19.0	19.0	24.0	24.0
Tea	6	1	7	1	7.0	7.0	6.0	6.0
Washing Powder	4		5	1	2.5	5.0	2.0	4.0
Sugar	3	4	5	5	20.0	25.0	12.0	15.0
Milk	2	2	3	3	6.0	9.0	4.0	6.0

Time reversal test is  $P_{01} \times P_{10} = 1$ .

According to Fisher formula,

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

$$P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

**Laspeyres Formula,**

$$P_{01} \times P_{10} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

$$= \frac{\sum p_1 q_1 \times \sum p_0 q_0}{\sum p_0 q_1 \times \sum p_1 q_0}$$

**Paasche Formula,**

$$P_{01} \times P_{10} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

$$= \frac{\sum p_1 q_1}{\sum p_1 q_0} \times \frac{\sum p_0 q_0}{\sum p_0 q_1}$$

**Marshall Edgeworth,**

$$P_{01} \times P_{10} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

$$= \frac{\sum p_1 q_1}{\sum p_1 q_0} \times \frac{\sum p_0 q_0}{\sum p_0 q_1}$$

**Simple Aggregative Method,**

$$P_{01} \times P_{10} = \frac{\sum p_1}{\sum p_0} \times \frac{\sum p_0}{\sum p_1}$$

$$= \frac{\sum p_1}{\sum p_1} \times \frac{\sum p_0}{\sum p_0}$$

**Factor Reversal Test:**

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

$$= \frac{\sum p_1 q_1}{\sum p_1 q_0} \times \frac{\sum p_0 q_0}{\sum p_0 q_1}$$

$$= \frac{\sum p_1 q_1}{\sum p_1 q_0} \times \frac{\sum p_0 q_0}{\sum p_0 q_1}$$

**Fixed Base and Chain Base Index Numbers**

Index numbers may be constructed by

- (1) Fixed Base or
- (2) Chain Base

**1. Fixed Base:**

In fixed baas method a definition or a parted of years is taken as base and prices of subsequent years are compared directly or independently with the price in base year. The same base year should be a normal year free from any abnormalities. It should also be not far in the past. The following formula used for computing fixed base index number.

**Index Number for a particular year**

$$= \frac{\text{Price of the Current}}{\text{Price of the Previous year}} \times 100$$








**Example :**

Construct index numbers for 8 years taking 1961 as base from the following data.

Year	Price (Rs.)	Year	Price (Rs.)
1961	65	1965	86
1962	70	1966	90
1963	74	1967	95
1964	80	1968	98

**Solution:**

Construction of Index Number Taking 1961 as base.

Year	Price (Rs.)	Index Number 1961= 100
1961	65	100
1962	70	 $\times 100 = 107.69$
1963	74	 $\times 100 = 113.84$
1964	80	 $\times 100 = 123.07$
1965	86	 $\times 100 = 132.30$
1966	90	 $\times 100 = 138.46$
1967	95	 $\times 100 = 138.46$
1968	98	 $\times 100 = 150.76$

This method, though convenient, has certain limitation. As time elapse conditions were once important become less significant and it becomes more difficult to compare accurate present conditions with those of a remote part. New items may have to be included and may have to be deleted in order to make the index more representative. In significant may be desirable to use the chain base number when this method is used to compare is made with s fixed base; rather than comparisons from year to year.

### Step in constructing a Chain Base index

In chain base method, there is no fixed base to compare the price of subsequent year, but price of each is compared with the price of the preceding year. In this way, we compute price relatives, there price relatives are called link relatives. These link relatives are chained together to get the Chain index. The following formulae may be used for computing link relatives and chain index:

$$\text{Link Relative} = \frac{\text{Price of the Current year}}{\text{Price of the Previous year}} \times 100$$

To get the chain indices, these link relatives are to be chained by the following formula:

$$\text{Chain Index} = \frac{\text{Link relative for Current Year} \times \text{Chain Indices of previous year}}{100}$$

- (1) Find the link relative for each year by employing the above formula.
- (2) Chain the link relatives to get the Chain Indices by the above formula.

### Example:

Construct Chain base index numbers for the following data:

Year	Price (Rs.)	Year	Price (Rs.)
1961	65	1965	86
1962	70	1966	90
1963	74	1967	95
1964	80	1968	98

### Solution:

#### Construction of Chain Index Numbers

Year	Price	Link Relatives	Chain Index
1961	65	100	100
1962	70	$\frac{70}{65} \times 100 = 107.7$	$100 \times \frac{107.7}{100} = 107.7$
1963	74	$\frac{74}{70} \times 100 = 105.7$	$107.7 \times \frac{105.7}{100} = 113.8$
1964	80	$\frac{80}{74} \times 100 = 108.1$	$113.8 \times \frac{108.1}{100} = 123.0$
1965	86	$\frac{86}{80} \times 100 = 107.5$	$123.0 \times \frac{107.5}{100} = 132.2$
1966	90	$\frac{90}{86} \times 100 = 104.6$	$132.2 \times \frac{104.6}{100} = 138.3$
1967	95	$\frac{95}{90} \times 100 = 105.5$	$138.3 \times \frac{105.5}{100} = 145.9$
1968	98	$\frac{98}{95} \times 100 = 103.2$	$145.9 \times \frac{103.2}{100} = 150.6$

### Conversion for Fixed Base and Chain Base Index Numbers

Index numbers computed with fixed base method can be converted into chain base index numbers and index numbers computed by chain base method can be changed into fixed base index numbers.

In fixed base method year to year comparison is not possible because index numbers are tied to a distant past. In many situations, year to year comparison is most essential to study the behavior of the variables. To accomplish this purpose fixed base index numbers are converted into chain base numbers by the following rule

$$\text{Chain index for 1936} = \frac{\text{Fixed base index for 1936}}{\text{Fixed base index for 1935}} \times 100$$

and

$$\text{Chain index for 1937} = \frac{\text{Fixed base index for 1937}}{\text{Fixed base index for 1936}} \times 100$$

and so on.

In chain base index numbers year to year comparison is possible but in many situations comparison with remote past becomes, necessary study in the growth pattern of the data. To achieve and objectives chain base index numbers are converted into fixed base index numbers by the following rule.

$$\text{Fixed Base Index for 1951} = \frac{\text{Fixed index for 1950} \times \text{Chain index for 1951}}{100}$$

and

$$\text{Fixed Base index for 1952} = \frac{\text{Fixed Base Index For 1951} \times \text{Chain Base Index for 1952}}{100}$$

and so on.

#### Example:

From the following fixed base index number to find chain base index numbers.

Year :	1965	1966	1967	1968	1969
Fixed Base Index :	425	446	457	480	496

#### Solution:

Conversion of Fixed Base Index Nos. to Chain Base Index Numbers.

Year	Fixed base Index Nos.	Fixed base index Nos. converted to Chain indices	Chain base index numbers
1965	425	-	100.00
1966	446	<div> </div> × 100	107.94
1967	457	<div> </div> × 100	102.46
1968	480	<div> </div> × 100	105.03
1969	496	<div> </div> × 100	103.33

**Example:**

Calculate fixed base index numbers of the following series of chain base index:

Year	:	1950	1951	1952	1953	1954	1955	1956
Chain Indices	:	100	115	120	93	102	156	83

**Solution:**

Conversion of Chain Base Index Numbers to Fixed Base Index Numbers.

Year	Chain Base Index Nos.	Chains Indices converted to Fixed base = 1950	Fixed Base Index Nos.
1950	100	-	100.00
1951	115		115.00
1952	120		138.00
1953	93		128.34
1954	102		130.90
1955	156		204.2
1956	83		169.48

**Merits of the Chain Base Method:**

1. The chain base method has a great significance in practice because in economic and business data we are more often concerned with making comparisons with the previous period, and not with any distant past. The link relatives obtained by chain base method serve this purpose.
2. Chain base method permits the introduction of new commodities and the deletion of old ones without necessitating either the recalculation of entire series or other drastic changes. Because of this flexibility, Chain index is used in many types of indices such as the consumer price index and the wholesale price index.
3. Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.
4. Index numbers calculated by the chain base method are free to a greater extent from seasonal variations than those obtained by the other method.

**Limitations of the Chain Index:**

The limitation of the chain index is that the whole percentages of previous year figures give accurate comparisons of year to year changes, the long range comparisons of chained percentages are not strictly valid. However, when the index number user wishes to make year to year comparisons, as is so often done, by the businessman, the percentages of the preceding year provide a flexible and useful tool.

\*\*\*\*\*



## LESSON-19

### ANALYSIS OF TIME SERIES

In business and economics, it is not only sufficient to study the past pattern of changes but also essential to measure, analyse and understand the forces which are operating in a firm or in an industry or the economic system as a whole. There are many approaches for studying as well as forecasting the operating forces but one of these approaches is known as 'Analysis of time series.'

A time series is a set of observations taken at specified times, usually at equal intervals. For example, when quantitative data regarding agricultural production, national income, birth rate are arranged in order of their occurrence, the resulting statistical series is called time series.

Mathematically, a time series is defined by the values  $X_1, X_2, X_3, \dots, X_n$  of the variable  $X$  at times  $t_1, t_2, t_3, \dots, t_n$ . Here the variable  $X$  is a function of time. ( $t$ )

$$i.e. \quad X = f(t)$$

It means that the variable  $X$  depends upon time. The problem of time series analysis can best be appreciated with the help of the following example.

Year	Sales of Firm A (thousand units)	Year	Sales of Firm a (thousand units)
1970	40	1974	73
1974	42	1975	48
1972	47	1976	45
1973	41	1977	44

If we observe the above series we find that generally the sales have increased but for two years a decline is also noticed. There may be several causes responsible for increase or decrease from one period to another such as changes in the tastes and habits of people, of population availability of alternate products etc. It may be very difficult to study, the effect of various factors that have led either to an increase or decrease in the sales. The statistician, therefore, tries to analyse the effect of the various forces under four broad heads :

- (1) Changes that have occurred as a result of general tendency of the data to increase or decrease, known as 'secular movements'.
- (2) Changes that have taken place during a period of 12 months as a result of change in climate, weather conditions, festivals etc. Such changes are called 'seasonal variations'.
- (3) Changes that have taken place as a result of booms and depressions. Such changes are classified under the head 'cyclical variations.'
- (4) Changes that have taken place as a result of such forces that could not be projected like floods, earthquakes famines etc. Such changes are classified under the head irregular or erratic variations'.

These are called components of time series and shall be discussed in detail.

#### Uses of Time Series Analysis :

The time series analysis is of great significance to the economist and businessman and researcher etc. for the reasons given below:

- (i) One can understand by observing data over a period of time the changes that have taken place in the past. Such analysis will be extremely useful in predicting the future behaviour.
- (ii) Time series, analysis helps us in forecasting events, with which we can plan for the future. In other words, it helps in planning future operation.
- (iii) Time series analysis facilitates comparison of different, points of time or different units performance and thereby drawing important conclusions.

### **Components of Time Series :**

The combined force that caused fluctuations in the value of a phenomenon in time series may be broadly classified into four categories, commonly known as the components of time series. All or some of which are present in a given time series at varying degree, the components of time series, which often called elements of series are:

- (a) Secular Trend
- (b) Seasonal Variations
- (c) Cyclical Variations
- (d) Irregular Variations'.

#### **(a) Secular Trend:**

The tendency of time series data to increase or decrease or stagnate during a long period. It is called the secular trend or simple trend. Thus, the trend is secular or long term due to the basic tendency to grow or decline over a period of time. Either increase or decrease would not be in the same direction throughout the given period, but different tendencies of increase, decrease or stability would be taken in different sections of time. Such tendencies are the result of the forces in an evolutionary manner and do not reflect sudden change. For example the growth or decline in economic time series is the interaction of forces like advances in production technology, large scale production, improved marketing management and , business organization—all of which are continuous but gradual process. The secular trend is broadly divided into two heads, namely, linear or straight line trend and non-linear trend. If the time series values plotted on graph appear more or less straight, then it is termed as linear trend, otherwise non-linear.

#### **(b) Seasonal Variations :**

A number of forces which repeat periodically over a 12 month period give rise to seasonal variations. In other words, a variation which is of periodic in nature and whose repeating cycle is of relatively short duration, say week, month or quarter. The amplitude of seasonal variations may vary, but their period is fixed being one year. The factor that cause seasonal variations are :

- (a) natural forces ; and
- (b) man-made conventions.

The climatic change, plays an important role in seasonal movements. Changes in natural forces like weather, climate, rainfall, humidity, heat etc; act on different, products and industries differently. While the nature is primarily response for seasonal variations in time series, the people's habits, fashions, customs, conventions also have their impact on seasonal variations.

The study of seasonal variations is very useful to decision making in the sense, planning future operations regarding purchase, production, inventory control, personnel needs, selling and

advertising programmes. In the absence of knowledge of seasonal variations, a seasonal upswing or seasonal slump may be mistaken as indicator of better business or deteriorating business conditions respectively. Thus, to understand the behaviour of the phenomenon in a time series properly, the time series data must be adjusted for seasonal variations.

**(c) Cyclical Variations:**

In most of the economic and business time series data, there is periodic up and down movement in the sense that recurrent variation usually last longer than a year and these are more or less regular. These variations are known as cyclical variations. In other words the cyclical variations are long term variation that recur in, rises and declines, in activities and may or may not follow exactly since patterns after equal intervals of time. On the complete period which normally lasts from a years is termed as a 'cycle'. These oscillation in any business activity are the outcome of the so called 'Business Cycles' which are four phase cycles consisting prosperity, decline, depression and improvement.

**(d) Irregular Variations :**

Irregular variations also called erratic accidental or random refer to variations, which do not recur in a definite pattern. In other words, these fluctuations are purely random and are the result of such unforeseen as well as unpredictable forces which operate in erratic and irregular manner. The non-recurring factors like floods, famines, droughts, wars, earthquakes etc, cause the fluctuations in a most powerful manner.

**Measurement of Trends:**

Trend component in time series can be studied and measured with the help of the following four methods, namely.

- (i) Freehand or Graphic Method.
- (ii) Semi-average Method,
- (iii) Moving-average Method,
- (iv) Least Square Method.

This method is simple and flexible in studying trend. The procedure to obtain straight line trend is to plot the given time series on graph and draw a straight line, carefully on the plotted dots which will best fit to the data. The line drawn should be smooth.

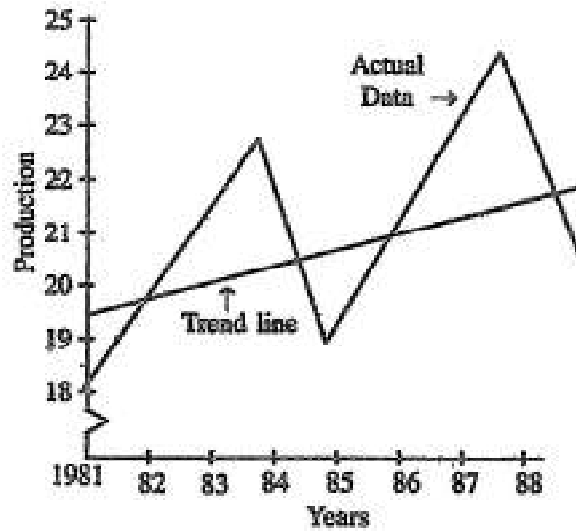
**Example:**

Fit a trend line 'to the following data by the graphic method:

Year	Production (in '000 tones')
1981	18
1982	20
1983	22
1984	19
1985	21
1986	25
1987	23
1988	21

**Solution :**

Fitting a trend line by the graphic method

**Semi-Average Method:**

This method has more objective method as compared with graphic method. In this method the whole time series is divided into several equals with reference to time. If we 'are given data from 1966' to 1985 i.e. 20 years, the two equal parts will be the first ten years i.e. from 1966 to 1975 and the second part from 1976 to 1985. In case of odd number of years, say 9, 11, 13, 15 etc., two equal parts will be made by ignoring the middle year. For example, if data are given for 11 years. fbr 1976 to 1986, the two equal parts would be from 1976 to 1980 and from 1982 to 1986, in this the middle year 1981 will be omitted.

After dividing the given series into two parts, next calculate the arithmetic mean of time series for each part. These arithmetic means are called 'Semi-average'. These semi-averages are plotted as points against the middle points of the respective time period covered by each part. The line joining points gives the straight line trend fitting the given data :

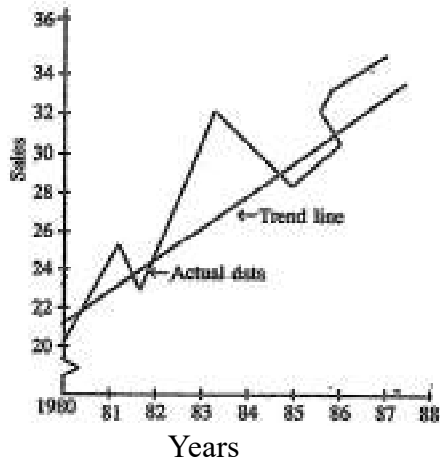
Determine trend of the following data by the method of semi-averages.

Year	Seles (000 Units)
1980	20
1981	24
1982	22
1983	30
1984	28
1985	30
1986	34
1987	36

**Solution:**

Year	Sales ('000 Units)	Semi-Average
1980	20	24
1981	24	
1982	22	
1983	30	
1984	28	32
1985	30	
1986	34	
1987	36	

The first part semi-average 24 is to be plotted against the mid-years or the first part *i.e.* 1981 and 1982, and the second semi-average 32 is to be identified against the mid-years of the second part *i.e.* 1985 and 1986. The trend line is shown in diagram given below :

**Moving Average Method:**

The trend is computed by smoothening fluctuations of the time series data by means of a moving average. The moving average is a series of successive, averages secured from a series of values by averaging 'groups' of 'n' successive values of the series. It is necessary to select a period for moving average such as 3 yearly moving average 5 yearly moving average, 8 yearly moving average, etc.

**Example:**

Calculate trend values taking a 3 yearly and 5 yearly period of moving average from the following data:

Year	:	1970	1971	1972	1973	1974	1975	1976	
Production									
'000 units	:	5	7		9	10	11	8	
Year	:	1977	1978	1979	1980	1981	1982	1983	1984
Production									
'000 units	:	12	13	17	20	10	15	12	14

**Solution:**

Computation of trend values by 3 yearly and 5 yearly moving average method.

Year	Production (‘000 units)	3 Yearly Total	Moving Average	5 Yearly Total	Moving Average
1970	5	-	-	-	-
1971	7	21	7.00	-	-
1972	9	28	9.33	43	8.60
1973	12	31	10.33	49	9.80
1974	10	33	11.00	50	10.00
1975	11	29	9.67	53	10.00
1976	8	31	10.33	54	10.80
1977	12	33	11.00	61	12.20
1978	13	42	14.00	70	14.00
1979	17	20	16.67	80	16.00
1980	20	55	18.33	83	16.60
1981	18	53	17.67	82	16.40
1982	12	41	13.67	-	15.80
1984	14	-	-	-	-

**Merits:**

- (i) The moving average method is simple as compared with least squares method.
- (ii) The moving average if happens to coincide with the cyclical movements, such variation are automatically removed.

**Limitation:**

- (i) In moving average method, the trend values cannot be computed for all the years.
- (ii) No predetermined or guiding principles are available to select the period of moving average. One has to use his own judgement. Therefore, great care has to be exercised in selecting the period of moving average.

**Method of Least Squares:**

The method of least squares is the most widely used method of fitting a line of the best fit to a series of data and is the most popular method of calculating trends in time series. It is a mathematical device employed for measuring the trend line which represents the movements of data most satisfactorily.

The method of least squares is given this name because its method of calculating gives its certain important mathematical properties which are not shared by other methods. These properties are:

1. The sum of the vertical deviations of the actual values of (y) from the fitted straight line when deviations above the trend line are given positive signs and deviations below the trend, line are given negative signs, is equal to zero,

In symbols :



2. The sum of the squared, deviations of actual values ( $y$ ) from trend values ( $y_c$ ), is less than the sum of squared deviations from any other straight line. In symbols:



= minimum.

This property is not shared by other straight lines. When a line is fitted to meet the first property is automatically met. It is because of second property that the name “Least Squares” is derived.

The major weakness of this method is that it does not indicate what type of trend should be fitted to the data. Generally this is to be decided by the investigator but once the decision has been made, the method of least squares supplies formula which may be used for estimating the time of the best fit. It is mathematical in character, therefore, it requires the solving of a set of simultaneous equations, called normal equations, to determine the values of constants involved in the equations.

#### **Fitting of Straight Line by the Method of Least Squares**

The simplest method of computing the trend values by the method of least squares is the straight line. The straight line series as the exact representation of the trend, in case the ‘series is increasing or decreasing by constant amounts. The equation of straight line may be written as

$$y_c = a + bx$$

Where,

$y_c$  = Computed value of the trend,

$x$  = Independent variable which represents time,

$a$  = Value of trend when  $x$  is zero ( $y$  intercept),

$b$  = Slope of the line

Here ‘ $a$ ’ and ‘ $b$ ’ are constants, once their values, are determined they do not change. The value of ‘ $b$ ’ represents the amount by which the trend increases or decreases for each unit of time. When ‘ $b$ ’ is positive, the trend increases by constant amounts and when ‘ $b$ ’ is negative, the trend decreases by constant amounts.

The values of ‘ $a$ ’ and ‘ $b$ ’ are computed by solving a set of simultaneous equations commonly called normal equations. The number of normal equations depends upon the number of constants in the equation.

In a straight line, there are two constants so we need two normal equations for computing the values of ‘ $a$ ’ and ‘ $b$ ’. Normal equations are :

$$\Sigma y = Na + b \Sigma x \quad \dots\dots (1)$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 \quad \dots\dots (2)$$

In above equations  $y$  represents original values in a time series,  $N$  stands for the number of years of  $x$  represents time.

There are two methods of fitting straight line trend:

### 1. Direct Method:

When this method is adopted for computing the values of constants 'a' and 'b', the first year is always taken as origin and its value is taken equal to zero, the next year is given value 1 and third, year 2 and so on. After computing the desired values, normal equations are solved simultaneously to get the values of 'a' and 'b'.

#### Example:

Fit a straight line trend to the following data:

Year	:	1947	1948	1949	1950	1951	1952	1953	1954	1955
Production	:	11	13	15	14	15	16	16	17	18
('000 tons)										

#### Solution:

Calculation of straight line trend by the method of least squares.

Year	Production	Origin, 1947		
	Y	X	XY	X <sup>2</sup>
1947	11	0	0	0
1948	13	1	13	1
1949	15	2	30	4
1950	14	3	42	9
1951	15	4	60	16
1952	16	5	80	25
1953	16	6	96	36
1954	17	7	119	49
1955	19	9	144	64
N = 9	ΣY = 315	ΣX = 36	ΣXY = 584	ΣX <sup>2</sup> = 204

The equation of straight line is

$$Y = a + bX$$

Normal Equations are

$$\Sigma Y = Na + b\Sigma x$$

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

Substituting the above computed values, we get

$$135 = 9a + 36b \quad \dots (1)$$

$$584 = 36a + 204b \quad \dots (2)$$

Multiplying equation (1) by 4 and subtracting it from equation (2), we get:

$$584 = 36a + 204b$$

$$540 = 36a + 144b$$

$$44 = 60b$$

or  $b = 0.73$



Substituting the value of (b) in equation (1)

$$135 = 9a + 36 \times 0.73$$

or  $9a = 108.72$

$$a = 12.8$$

The required equation is

$$Y = 12.8 + 0.73 X$$

## 2. The Short-cut Method

In this method, origin is always taken in the middle of the time series in such a way that  $\Sigma x$  becomes zero. The small letter 'x', representing time, has been substituted to distinguish it from capital letter 'X' which is employed in the direct method.

Normal equations are :

$$\Sigma y = Na + b \Sigma x$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2$$

By taking origin in the middle of time period covered,  $\Sigma x$  becomes zero. Therefore, the above equations are reduced to the form :

$$\Sigma y = Na$$

or  $a = \frac{\Sigma y}{N}$

the second equation becomes.

$$\Sigma xy = b \Sigma x^2$$

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

Hence, we can compute the values of 'a' and 'b' directly. If the series consists, of an odd number of years, the origin would be the middle of the series. On the contrary if the series consists of an even number of years, the origin falls in the centre of two middle years.

### (a) Odd Number of Years:

When we are supplied with the data for an odd number of years, the origin is always taken in middle of time period so that  $\Sigma x$  becomes zero.

The values of Parameters 'a' and 'b' calculated with the aid of formula :

$$a = \frac{\Sigma y}{N} \text{ and } b = \frac{\Sigma xy}{\Sigma x^2}$$

### (b) Even Number of Years:

If the time series consists of an even number of years the middle of the series falls between two years. This middle would necessitate the use of decimals to measure the distance of any years from the origin.

In case of even years also  $\Sigma x$  will be zero if the origin is placed, mid way between the two middle years. For example, if the years are 1973, 1974, 1975, 1976, 1977 and 1978 we can take deviations from the middle year 1975.5. If the deviations would be -2.5, -1.5, -0.5, +0.5, +1.5, -2.5 for the various years and the total  $\Sigma x$  would be zero. Hence both in odd as well as in even number of years we, can use the simple procedure of determining the values of the constant  $a$  and  $b$ .

**Example:**

Fit a straight, line trend by the method of least squares to the following data. Assuming that the same rate of change, continues, what would be the predicted earnings for the year 1972.

Year	:	1963	1964	1965	1966	1967	1968	1969	1970
Earnings									
(Rs. In lakhs)	:	38	40	65	72	69	60	87	95

**Solution:**

Fitting of straight line trend by the method of Least Squares.

Year (Rs. In lakhs) By 2	Earnings from 1966 X	Deviations multiplied XY	Deviations X <sup>2</sup>		
1963	38	-3.5	-7	-266	49
1964	40	-2.5	-5	-200	25
1965	65	-1.5	-3	-1.95	9
1966	72	-0.5	-1	-72	1
1967	69	+0.5	+1	+69	1
1968	60	+1.5	+3	+180	9
1969	87	+2.5	+5	+435	25
1970	95	+3.5	+7	+665	49
N = 8	ΣY = 526		ΣX = 0	ΣXY = 616	ΣX <sup>2</sup> = 168

$$Y_0 = a + bX$$

$$a = \frac{\Sigma Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{616}{168} = 3.667$$

$$Y = 65.75 + 3.667X$$

For 1972, X will be + 11

When X is + 11, Y will be

$$\begin{aligned}
 Y &= 65.75 + 3.667(11) \\
 &= 65.75 + 40.337 \\
 &= 106.087
 \end{aligned}$$

Thus the estimated earnings for the year 1972 are 106.087 lakhs.

**Merits and Limitation of the Method of Least Squares****Merits:**

1. This is a mathematical method of measuring trend and as such there is no possibilities of subjectiveness.

2. The line obtained by this method is called the line of best fit because it is the line from which the sum of the positive and negative deviations is zero and the sum of the squares of the deviations is least, *i.e.*  $\Sigma(y-y_c) = 0$  and  $\Sigma(y-y_c)^2$  is least.

**Limitations:**

Mathematical curves are useful to describe the general movement of a time series but it is doubtful whether any analytical significance should be attached to them, examine in special cases, it is seldom possible to justify on theoretical grounds any real dependence of a variable on the passage of time. Variables do change in a more or less systematic manner over time, but this can usually be attributed to the operation of other explanatory variables. Thus many economic time series show persistent upward trends over time due to a growth of population or to a general rise in prices, *i.e.*, national income and the trends element can to a considerable extent be eliminated by expressing these series; per capita or in terms of constant purchasing power. For these reasons mathematical trends are generally best regarded as tools for describing movements in time series rather than as theories of the causes of such movements.

**SUGGESTED READINGS**

1. Gupta, S.P. : Statistical Methods, Chapter 14.
2. Grewal, R.S. : Methods of Statistical Analysis Chapter 14.
3. Gupta, S.C. & Kapoor, V.K. : Fundamentals of Applied Statistics Chapter 2.
4. Nagar, A.L. and Das, R.K. : Basic Statistics, Chapters 11 to 11.7
5. Croxton and Cowden : Applied General Statistics, Chapters 15 and 16.

\*\*\*\*\*

## LESSON-20

### MEASUREMENTS OF NON-LINEAR TRENDS

The second degree equations alternatively called second degree parabola, is one of the most useful and simplest methods of fitting non-linear trends. The method of fitting second degree equations is only a little more, complicated than a straight line since it involves the addition of one more constant 'C'. The equation of second degree parabola may be written as:

$$y_c = a + bx + cx^2$$

Where,

- $y_c$  stands for trend values
- $a$  stands for the y-intercept
- $b$  stands for the slope at the origin
- $c$  stands whether the curve is concave-upwards or downwards
- $x$  stands for time

It is named second degree equation because of the fact that two is the highest power to which 'x' appears in the equation. It is the square value 'x' which gives curvature to the trend line whether the curve is concave upwards or downwards depends upon the value of 'c'.

- (i) If the value of 'c' is negative, the curve will have a downward bulge.
- (ii) If the value of 'c' is positive, the curve will have an upward bulge.

#### Fitting of Second Degree Equation

The second degree equation like the straight line may be fitted by two methods namely:

##### (a) Direct Method

In second degree parabola there are three constants, so we need three normal equations for computing the values of constants.

Normal Equations are :

$$\Sigma y = Na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

To find the values of constants a, b and c, the above equations may be solved simultaneously.

##### (b) Short-cut

By taking origin in the middle of the series "x becomes zero and the above equations are reduced to the form :

$$\Sigma y = Na + c\Sigma x^2 \quad \text{..... (1)}$$

$$\Sigma xy = a\Sigma x^2 \quad \text{..... (2)}$$

$$\Sigma x^2y = a\Sigma x^2 + c\Sigma x^4 \quad \text{..... (3)}$$

The value of 'b' is given by the equation (2)



The values of 'a' and 'b' may be determined by the simultaneous solution of equations (1) and (3). These values may also be computed directly :



**Example:**

Fit a second degree parabola to the following data and estimate to value for 1986.

Year	:	1979	1980	1981	1982	1983
Sale in ('000 Rs.)	:	10	12	13	10	8

**Solution:**

Computation of second degree parabolic trend equation.

Years	Sales	$x$	$xy$	$x^2$	$x^2y$	$x^2$	$x^2$	Trend value
	( '000 Rs.)	X - 1981						
	Y							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1979	10	-2	-20	-4	40	-8	16	10.086
1980	12	-1	-12	1	12	-1	1	12.057
1981	13	0	0	0	0	0	0	12.314
1982	10	+1	+10	1	10	+1	1	10.857
1983	8	+2	+16	4	32	+8	16	7.686
N= 5	y=53	x = 0	xy = - 6	$\Sigma x=10$	$\Sigma x^2y=94$	$\Sigma x^2 = 0$	$\Sigma x^4 =34$	



= 12.314

The second degree parabolic trend elation is :

$$y = 12.314 - 0.6x - 0.857x^2$$

Trend values shown in column (9) of the table are transits above equation in the following manner:

$$\begin{aligned}\text{For } 1979 : x = -2, y_c &= 12.314 - 0.6(-2) - 0.857(-2)^2 \\ &= 10.086\end{aligned}$$

$$\begin{aligned}\text{For } 1980 : x = -1, y_c &= 12.314 - 0.6(-1) - 0.857(-1)^2 \\ &= 12.057\end{aligned}$$

and soon

Trend value for 1986,

$$\begin{aligned}x = 5, y_c &= 12.31 - 0.6(5) - 0.857(5)^2 \\ &= 12.314 - 3.0 - 21.425 \\ &= -12.111.\end{aligned}$$

It is of the form :

$$y = a + bx + cx^2 + dx^3$$

For computing the values of  $a, b, c$ , and  $d$ , we need four normal equations as shown below :

$$\Sigma y = Na + b\Sigma x + c\Sigma x^2 + d\Sigma x^3 \quad \dots\dots (1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 + d\Sigma x^4 \quad \dots\dots (2)$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 + d\Sigma x^5 \quad \dots\dots (3)$$

$$\Sigma x^3y = a\Sigma x^3 + b\Sigma x^4 + c\Sigma x^5 + d\Sigma x^6 \quad \dots\dots (4)$$

The value of  $a, b, c$  and  $d$ , may be determined by solving a set of four equations simultaneous.

By making  $\Sigma x = 0$ , the above equations may be reduced to the form:

$$\Sigma y = Na + c\Sigma x^2 \quad \dots\dots (5)$$

$$\Sigma xy = b\Sigma x^2 + d\Sigma x^4 \quad \dots\dots (6)$$

$$\Sigma x^2y = a\Sigma x^3 + c\Sigma x^4 \quad \dots\dots (7)$$

$$\Sigma x^3y = d\Sigma x^4 + d\Sigma x^4 \quad \dots\dots (8)$$

The values of 'a' and 'c' may be computed by solving equations (5) and (7) and the values of 'b' and 'd' may be determined by solving equations (6) and (8).

### Logarithmic Straight Line Trend or Exponential Curve

The straight line trend is fitted when the data is 'increasing or decreasing by constant amounts from one period to another. Such a series show linear trend when plotted on a graph paper. If a series does not show a linear trend when plotted on an arithmetic graph paper but depicts a linear trend on a semi log paper then log arithmetic straight line trend is fitted.

The equation of log straight line is of the form :

$$y_c = ab^x$$

where :

- (a) If the value of 'b' is a positive number greater than one, then the trend of the series is upward and the amount of change is undergoing 'a constant percentage of increase.
- (b) If the value of 'b' is a positive number smaller than one then trend is downward and the amount of change shows a constant percentage of decrease.

These trends are also called exponential trends because  $x$  appears as an exponent in the equation. These trends are also named semi-logarithmic linear trends because we plot the values of  $y$  against absolute value of  $x$ .

### Method of Fitting the Exponential Curve:

The exponential curve is of the form :

$$y = ab^x$$

In log form, the equation may be written as :

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

Normal Equations are

$$\sum \log y = N \log a + \log b \sum x \quad \dots\dots (1)$$

$$\sum x \log y = \log a \sum x + \log b \sum x^2 \quad \dots\dots (2)$$

By taking origin in the centre so that  $\sum x$  becomes zero, the above equation may be reduced to the form:

$$\sum \log y = N \log a$$

Or  $\log \boxed{\phantom{000000}}$

and  $\sum x \log y = \log b \sum x^2$

or  $\log \boxed{\phantom{000000}}$

By taking antilog of  $\log a$  and  $\log b$ , we can find the values of  $a$  and  $b$ .

### Steps:

The following steps are involved in fitting an equation of the type  $y = ab^x$

- (i) Write the log values of  $y$ , the dependent variable ( $\log y$ ).
- (ii) Take origin in the middle of time period to make  $\sum x = 0$  and write down the deviations ( $x$ ).
- (iii) Multiply  $\log y$  values with deviations obtained in step 2 to get  $x \log y$ .
- (iv) Apply the rule of addition to the products obtained in step (iii) to get  $\sum x \log y$ .
- (v) Square the deviations to get  $\sum x^2$ .

### Example:

Fit a logarithmic straight line trend to the following data showing the production ('000 tons) of a sugar factory during 1979-85

Years	:	1979	1980	1981	1982	1983	1984	1985
Production								
('000 tons)	:	12	10	14	18	20	24	30

**Solution:**

Fitting of logarithmic straight line trend by the method of Least Squares.

Year	Production		Origin, 1979			
	(‘000 tons)				Trend Values	
	y	log y	x	x <sup>2</sup>	x log y	y <sub>c</sub>
1979	12	1.08	-3	9	-3.24	10.00
1980	10	1.00	-2	4	-2.00	12.02
1981	14	1.15	-1	1	-1.15	14.45
1982	18	1.26	0	0	0	17.38
1983	20	1.30	1	1	+1.30	20.89
1984	24	1.38	2	4	+2.76	25.12
1985	30	1.48	3	9	+4.44	30.20
N= 7		Σlog y = 8.65	Σx <sup>2</sup> = 28	Σ(x log y) = 2.11		

The logarithmic straight line is of the form

$$y = ab^x$$

or  $\log y = \log a + x \log b \dots\dots (i)$ 

Normal equations are

$$\Sigma \log y = N \log a + \log b (\Sigma x)$$

$$\Sigma (x \log y) = \log a \Sigma x + \log b (\Sigma x^2)$$

Since  $\Sigma x = 0$ , the above equations are reduced to the form

$$\log a = \frac{\Sigma \log y}{N} = \frac{8.65}{7}$$

$$\text{and } \log b = \frac{\Sigma (x \log y)}{\Sigma x^2} = \frac{2.11}{28}$$

Thus the required equation is

$$\log y = 1.24 + 0.08 x$$

**Computation of Trend Values**

$$\text{Year 1979, } \log y = 1.24 + 0.08 (-3) = 1.0$$

$$y = \text{Antilog } (1.0) = 10.00$$

$$\text{Year 1980, } \log y = 1.24 + 0.08 (-2) = 1.08$$

$$y = \text{Antilog } (1.08) = 12.02$$

$$\text{Year 1981, } \log y = 1.24 + 0.08 (-1) = 1.16$$

$$y = \text{Antilog } (1.16) = 14.45$$

$$\text{Year 1982, } \log y = 1.24 + 0.08 (0) = 1.24$$

$$y = \text{Antilog } (1.24) = 17.38$$



$$\text{Year 1983, } \log y = 1.24 + 0.08 \times 1 = 1.32$$

$$y = \text{Antilog}(1.32) = 20.89$$

$$\text{Year 1984, } \log y = 1.24 + 0.08 \times 2 = 1.40$$

$$y = \text{Antilog}(1.48) = 30.20$$

**Example:** Fit the curve  $y = ae^{bx}$  to the following data 'e' being 2.7183.

$x$	:	0	2	4
$y$	:	5.012	10	31.620

**Solution:** Fitting of the curve :

The form of the equation is

$$y = ae^{bx}$$

In logarithmic form this equation can be written as

$$\log y = \log a + (b \log e)x.$$

Putting  $y = \log 10 y$  and  $B = b \log 10e$

The above equation is reduced to the form

$$y = A + Bx$$

Normal equations are

$$\Sigma y = NA + B\Sigma x \quad \dots\dots (1)$$

$$\Sigma xy = A\Sigma x + B\Sigma x^2 \quad \dots\dots (2)$$

The required calculation are shown in the following table

$x$	$y$	$\log y = y$	$xy$	$x^2$
0	5.012	0.7	0.0	0
2	10.000	1.0	2.0	4
4	31.620	1.5	6.0	16
$\Sigma x = 6$	-	$\Sigma y = 3.2$	$\Sigma xy = 8.0$	$\Sigma x^2 = 20$

Substituting these values in equation (1) and (2)

$$3.2 = 3A + 6B \quad \dots\dots (3)$$

$$8.0 = 6A + 20B \quad \dots\dots (4)$$

Multiplying equation (3) by 2 and subtracting from equation (4), we get

$$8.0 = 6A + 20B$$

$$-6.4 = 6A + 12B$$

$$1.6 = 8B$$

$$B = 1.6/8 = 0.2$$

Substituting the value of B in equation (3), we get

$$3.2 = 3A + 6 \times 0.2$$

$$\text{or } 3A = 3.2 - 1.2$$

$$3A = 2.0$$

$$A = 0.67$$

Now  $A = \log 10^a$

$$0.67 = \log 10^a$$

Taking antilog, we get

$$a = 4.677$$

and  $B = 6 \log 10^e$

or  $0.2 = 0.4343 b$

or  $b = 0.46$  ( $\log 10e = -0.4343$ )

Hence the required equation is

$$0.46x$$

$$y = 4.677e$$

### The Gompertz Curve:

The Gompertz curve may be used to describe an increasing series which is increasing by a decreasing percentage of growth or a decreasing series which is decreasing by a decreasing percentage of growth. It is mainly employed in the study of economic and social trends because it portrays a process of cumulative expansion to a maximum value.

It is defined by the equation

$$y = ka b^x$$

After taking log of the equation it takes the shape of modified exponential form as shown below

$$\log y = \log k + b^x \log a$$

The shape of the Gompertz curve depends upon the values of  $k$ ,  $\log a$  and  $b$ . Actually its shape depends upon whether

1.  $\log a$  is positive or negative
2. The value of ' $b$ ' is greater or lesser than one.

The Gompertz curve like the modified exponential curve also takes four shapes depending upon the values of  $k$ ,  $\log a$  and  $b$ .

### The Method of Fitting the Gompertz Curve

Two methods may be employed for fitting this curve

1. The Method of partial total
2. The method of selected points.

The procedure of fitting the Gompertz Curve by these methods is the same as for fitting the modified exponential curve, except for a minor difference that we use log values of ' $y$ ' in place of given values of ' $y$ ' for computing the three constants, i.e.  $k$ ,  $a$  and  $b$ .

**Example :** Fit a Gompertz curve to the following

Year (‘000 Rs.)	Savings
1968	5.00
1968	7.07
1970	8.41
1971	9.17
1972	9.58
1973	0.79

**Solution:** Fitting of Gompertz Curve

The Gompertz Curve is stated in the form

$$y = ka b^x$$

in log form, it may be expressed as

$$\log y = \log k + b^x \log a$$

Year	$x$	Savings	$\log y$	$b^x$	$\log a(b)$	$\log y = \log k$ $+ \log a (b^x)$
1968	0	5.00	0.6990	1.00	-0.03	0.6992
1969	1	7.07	0.8494	0.50	-0.84	0.8492
$\Sigma_1 \log y$		1.5484				1.5484
1970	2	8.41	0.9248	0.25	-0.075	0.9242
1971	3	9.17	0.9624	0.125	-0.0375	0.9617
$\Sigma_2 \log y$		1.8872				1.8859
1972	4	9.28	0.9814	0.0625	-0.01875	0.9804
1973	5	9.79	0.9908	0.0312	-0.00937	0.9898
$\Sigma_3 \log y$		1.9722				1.9702

**Calculation of Constants**



where,  $\Sigma_3 \log y = 1.9722$ ,  $\Sigma_2 \log y = 1.8872$   
 $\Sigma_1 \log y = 1.5484$ ,  $n = 2$

**Substituting these values**



or

$$b = 0.25$$

$$\log a = (\Sigma_2, \log y - \Sigma_1, \log y)$$

$$(1.8872 \times 1.5484)$$

=

and

$$(1.5484 - 0.45)$$

$$= 0.9992.$$

Thus the required equation is

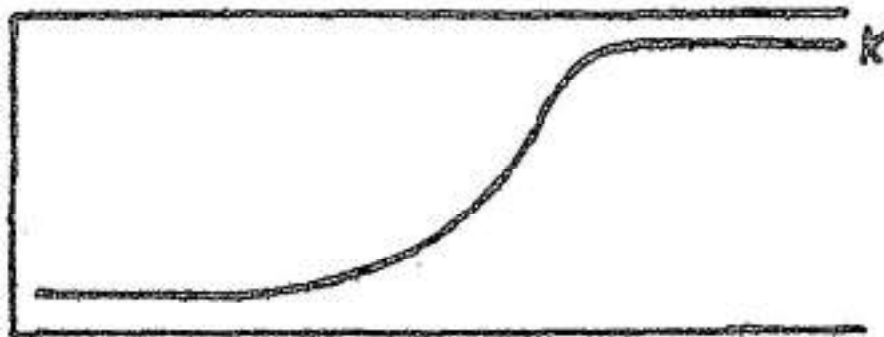
$$\log y = 0.9992 + (0.5)^x (.03)$$

Trend values computed with the aid of the above equation are shown in the last column of the above table.

### The Logistic Curve :

The logistic curve was introduced by Raymond Pearl and L.G. Reed, therefore, it is also called Pearl-Reed growth curve. It was employed by them for predicting the growth of population in the United States.

### The Logistic Curve



The logistic curve is stated in the form

$$1/y = k + ab^x$$

The logistic curve is identical with the modified exponential except that 'y' in that equation is changed in to 1/y in the case of logistic curve.

The shape of the logistic curve resembles an elongated 'S' rising from a lower asymptote of zero to an upper asymptote indicated by  $k$  as has been depicted.

### **The Method of Fitting the Logistic Curve**

The method of fitting the logistic curve is the same which has been employed for fitting the modified exponential curve with a minor modifications. Here we employ reciprocals of y rather than original values of y. First, we find the reciprocals of original values and then the series is divided into three equal parts. The observations falling in each group are totaled.

$S_1$  stands for the total of observations in the first part.

$S_2$  stands for the total of observations in the second part.

$S_3$  stands for the total for observations in the thud part.

The values of three constants  $a$ ,  $b$  and  $k$  may be combated by the following expression:

$$\frac{S_1}{S_2} = \frac{k + ab^x}{k + ab^{x+1}}$$

$$\frac{S_2}{S_3} = \frac{k + ab^{x+1}}{k + ab^{x+2}}$$

and

$$\frac{S_1}{S_3} = \frac{k + ab^x}{k + ab^{x+2}}$$

The logistic curve gives a fairly good representation of the stages of slow initial growth acceleration and retardation in the life history of an industry. This curve may be used for finding the tread of economic series that decreases at a constant rate at later stages.

The phenomenon for which this growth curve has been used is population growth, or the number of cells in an organism or the number of individuals in the region.

Both the Gompertz curve and the logistic curve are identical because both the curves can be employed to describe increasing series which are increasing by a decreasing percentage of growth or decreasing series which are decreasing by decreasing percentage of decline. But both the curves differ in one respect, the Gompertz curve involves, a constant ratio of successive first differences of log y values whereas for the logistic curve it is the constant ratio of successive first differences of 1/y values.

\*\*\*\*\*

# ASSIGNMENTS

## UNIT I

*Attempt any two questions.*

1. Discuss the usefulness of the measures of variability in economic analysis.
2. Show that the coefficient of determination is equal to the product of two regression-coefficients in a bivariate regression model.
3. Explain the difference between a parameter and a statistic.
4. State the assumptions of linear regression model and discuss their rationale.

## UNIT-II

*Attempt any two Questions*

1. Calculate the correlation coefficient between the following series :										
	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Road accidents	155	150	180	135	156	168	178	160	132	145
Consumption of	70	63	72	60	66	70	74	65	62	67
Liquor (in tons)										

Interpret your results.

2. The following table includes the ranking of preference of two housewives for ten different brands of lipsticks.

Brands of Lipstics	1	2	3	4	5	6	7	8	9	10
Ranking by Mrs. X	X	3	5	8	10	7	9	1	4	6
Ranking by Mrs. Y	Y	5	6	4	9	8	3	1	2	10

3. A random, sample of ten families had the following income the food expenditure (in Rs. per week)

Families	A	B	C	D	E	F	G	H	I	J
Families Income	20	30	33	40	15	13	26	38	35	43
Family Expenditure	7	9	8	11	5	4	8	10	9	10

Estimate the regression line of food expenditure on income and interpret your results.

5. The following results have been obtained from a sample of 11 observations on the values of sales (Y) of a firm and the corresponding prices (x)

$$x = 519.18 \quad y = 217.82$$

$$\Sigma X_1^2 = 3,134,543 \quad \Sigma X_1 Y_1 = 1,296,836$$

$$\Sigma Y = 539,512$$

- (i) Estimate the regression line of sales on price and interpret the results.
- (ii) What is the part of variation in sales which is not explained by the regression line ?

### UNIT-III

#### *Attempt any two Questions*

1. The following Table shows the values of expenditure on clothing (Y), total expenditure ( $X_1$ ) and the price of clothing ( $X_2$ ).

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
$X_2$	16	13	10	7	7	5	4	3	35	2
$X_1$	15	-20	30	42	50	54	65	72	85	9
Y	3.5	4.3	5	6	7	9	8	10	12	14

- (a) Find the least squares -regression equations of Y on  $X_1$  and  $X_2$ .
- (b) Find the explain and unexplained variation in Y.
2. The following results were obtained from a sample of 12 firms on their output (Y), labour input ( $X_1$ ) and capital input ( $X_2$ ), measured in arbitrary units.

$$\Sigma Y = 753 \quad \Sigma Y^2 = 48,139 \quad \Sigma Y X_1 = 40,830$$

$$\Sigma Y_1 = 643 \quad \Sigma X_1^2 = 34,843 \quad \Sigma Y X_2 = 6,796$$

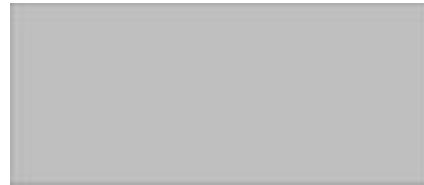
$$\Sigma X_2 = 106 \quad \Sigma X_2^2 = 976 \quad \Sigma Y_1 X_2 = 5,779$$

- (a) Find the least squares equation of Y on  $X_1$  and  $X_2$ . What is the economic meaning of your coefficients?
3. (a) Explain mathematical expectation.
- (b) A random sample of 700 units from large consignment should that 20 were damaged. Find.
  - (a) 95%
  - (b) 99% confidence limits for the proportion damaged units in consignment.
4. Differentiate, between parametric and non-parametric test. What are the advantages and limitations of parametric test.

## UNIT-IV

### *Attempt any two Questions:*

- (1) Define the following types of events occurring in the theory of probability and state (without proof) the theorems concerning them :
  - (i) Mutually exclusive                      (ii) Dependent
  - (iii) Independent                              (iv) Exhaustive
- (2) Write the properties of normal and  $\chi^2$  distribution, t and F distributions.
- (3) Write notes on
  - (i) Type I and Type II Errors                      (ii) Mathematical Expectation
- (4) The lengths of six randomly chosen sailors are in inches 63, 65, 68, 69, 71 and 72. Those of the randomly chosen soldiers are 61, 62, 65, 66, 69, 70, 71, 72, 73.
  - (a) Discuss in the light that the data throw that the soldiers are, on the average, taller than the sailors.
  - (b) Test whether correlation is significant  
if  $r = 6$  and  $n = 38$ .
- (5) What is the probability that a correlation coefficient of 0.75 or less can arise in a sample of 30 from a normal population in which the true correlation is + 0.9 ?



- (6) Four breeds of cattle  $B_1, B_2, B_3, B_4$ , were fed on three different rations  $R_1, R_2, R_3$ . Gain in weight in pounds over a period were recorded

	$B_1$	$B_2$	$B_3$	$B_4$
$R_1$	46.5	62	41	45
$R_2$	47.5	41.5	22	31.5
$R_3$	50	40	26.5	28.5

Is there a significant difference:

- (a) between breeds, (b) between rations ?

[Ans. no significant difference between breeds or between rations.]



- (7) (a) Find out the approximate probability that a correlation coefficient of 0.75 or less can arise in a sample of 18 from a normal population in which the true coefficient is + 0.9.
- (b) Test on fee 5% level the significance of the following:
- (1) A partial correlation -0.45 between  $x_1$ , and  $x_2$  (eliminate effect of  $x_3$  and  $x_4$ )
- (c) The correlation coefficient between mathematical aptitude linguistic aptitude for a group of 42 boys is 0.72 and for a group of 225 girls is 0.75. Is the difference significant?
- (8) The manufacturer of a patent medicine claimed that it was 90% effective in relieving an allergy for a period of 8 hours in a sample of 200 people who had the allergy, the medicine provided relief for 160 people. Determine whether the manufacture's claim is legitimate.

(9) Fit a straight line  $Y = \square + \square x$  to the following data:

$x$	1	2	3	4	5	6	7	8
$y$	14	27	40	55	68	70	76	80

Make use of method of least squares; test  $H_0 = \beta = 0$ .

[Hint: Calculation of Standard Error of Regression.



- (10) The intelligent quotient of 20 students from one college showed mean of 107 with a standard of 11 while the 9Q of 16 students from another college showed a mean of 112 with a standard deviation of 9. Is there a significant different between 9 Q.S. of the two groups at
- (i) 0.01 and (ii) 0.05 level of significance

## UNIT-V

**Attempt any two Questions:**

1. The number of units of product exported during 1960-67 is given below. Fit a straight line trend to the data. Find estimated for the year 1995.

Year	1987	1988	1989	1990	1991	1992	1993	1994
No. of Units	12	13	13	16	19	20	21	23
('000)								

2. Fit the second degree parabola to the following:

x	:	1	2	3	4	5	6	7	8	9
y	:	2	6	7	8	10	11	11	10	9

3. Below are given the figures of production (Lakh tonnes) of a sugar factory:-

Year	1980	1981	1982	1983	1984	1985
Production:	77	88	95	114	119	127

Fit a trend  $y = a^{bx}$  to this data and tabulate the trend values.

4. What are Index numbers ? How are they constructed ? Discuss the limitation of index numbers.
5. Discuss the problems encountered in the construction of Index Numbers. Bring put the relative merits and demerits chain-based and fixed based methods of constructing index numbers.
6. Construct Fisher's Ideal numbers show how it satisfies the factor reversal test with the help of the following data:

Commodity	$P_0$	$q_0$	$P_1$	$q_1$
x	56	70	50	26
y	32	107	35	85
z	41	62	30	50

7. Distinguish between 'Moving Average' and 'Least Squares' as methods of measuring trend in given time series. Which method is better and why ?
8. The following figures relate to the prices and quantities of certain commodity in base year and current year respectively. Construct appropriate index using these data:

Commodity	Base Year		Current Year	
	Quantity	Price	Quantity	Price
A	50	32	50	30
B	35	30	40	25
C	55	16	50	18

Check whether the Index-Number satisfies .the Time Reversal Test.

9. The price quotations for different items are given below Calculate index number by using weighted average of Relatives method (Base—1980)

Item	Price (1980)	Quantity (1980)	Price (1985)
A	6	50	9
B	8	40	12
C	10	30	12
D	2	100	2
E	4	60	6

10. (a) Explain the importance of different components of a time series.

(b) Compute the 4 years moving average trend of the following time series:

Year	:	1981	1982	1983	1984	1983	1986	1987	1988
Price	:	3	6	9	8	7	10	9	7

\*\*\*\*\*

*M.A. IInd Semester  
Economics*

*Course Code: DSC ECON-122*

# ***Basic Statistics***

*Lessons 1-20*

*By: Dr. K. Kaushik*

***International Centre for Distance Education & Open Learning  
Himachal Pradesh University, Gyan Path  
Summer Hill, Shimla - 171005***

# CONTENTS

SR. NO.	TOPIC	PAGE NO.
LESSON-1	MEASURES OF CENTRAL TENDENCY	3
LESSON-2	MEASURES OF DISPERSION AND SKEWNESS	28
LESSON-3	CORRELATION AND REGRESSION	53
LESSON-4	FITTING OF REGRESSION EQUATION AND STANDARD ERROR OF ESTIMATE	77
LESSON-5 & 6	THE GENERAL LINEAR REGRESSION MODEL : MATRIX FORMULATION & SOLUTION	87
LESSON 7 & 8	MULTIPLE AND PARTIAL CORRELATION	103
LESSON-9	A PROBABILITY THEORY & CONCEPT OF PROBABILITY DISTRIBUTION & A DENSITY FUNCTION	120
LESSON-10	MATHEMATICAL EXPECTATION	137
LESSON- 11	STATISTICAL HYPOTHESIS	144
LESSON-12	NON-PARAMETRIC TEST : THE SIGN TEST, RANK SUM TEST, THE MANN-WHITNEY U TEST, ADVANTAGES AND LIMITATIONS	152
LESSON- 13&14	STANDARD ERROR OF MEAN STUDENT'S 'T' DISTRIBUTION, CHI-SQUARE TEST	160
LESSON-15	TESTING HOMOGENEITY OF SEVERAL INDEPENDENT ESTIMATES OF POPULATION VARIANCE	192
LESSON-16	ANALYSIS OF VARIANCE	205
LESSON-17	INDEX NUMBERS	217
LESSON-18	TESTS FOR CONSISTENCY OF INDEX NUMBER	228
LESSON-19	ANALYSIS OF TIME SERIES	239
LESSON-20	MEASUREMENTS OF NON-LINEAR TRENDS	250
	ASSIGNMENTS	260