BUSINESS TOOLS FOR DECISION MAKING

B.Com.,

IV- SEMESTER

Dr. P.RAMAR

BHARATHIDASAN UNIVERSITY
CONSTITUENT COLLEGE,
NAGAPATTINAM

Module 1

INTRODUCTION

The term statistics seems to have been derived from the Latin word *'status'* or Italian word *'statista'* or the German word *'statistic*, each of which means *political state.*

The word 'Statistics' is usually interpreted in two ways. The first sense in which the word is used is a plural noun just refer to a collection of numerical facts. The second is as a singular noun to denote the methods generally adopted in the collection and analysis of numerical facts. In the singular sense the term 'Statistics' is better described as statistical methods.

Different authors have defined statistics in different ways. According to Croxton and Cowden statistics may be defined as *"collection, organisation presentation, analysis and interpretation of numerical data"*

Population and sample

Population

An aggregate of individual items relating to a phenomenon under investigation is technically termed as 'population'. In other words a collection of objects pertaining to a phenomenon of statistical enquiry is referred to as population or universe. Suppose we want to collect data regarding the income of college teachers under University of Calicut,, then, the totality of these teachers is our population.

In a given population, the individual items are referred to as elementary units, elements or members of the population. The population has the statistical characteristic of being finite or infinite. When the number of units under investigation are determinable, it is called finite population. For example, the number of college teachers under Calicut University is a finite population. When the number of units in a phenomenon is indeterminable, eg, the number of stars in the sky, it is called an infinite population.

Sample

When few items are selected for statistical enquiry, from a given population it is called a 'sample'. A sample is the small part or subset of the population. Say, for instance, there may be 3000 workers in a factory. One wants to study their consumption pattern. By selecting only 300 workers from the group of 3000, sample for the study has been taken. This sample is not studied just for its own sake. The motive is to know the true state of the population. From the sample study statistical inference about the population can be done.

Census and sample Method

In any statistical investigation, one is interested in studying the population characteristics. This can be done either by studying the entire items in the population or on a part drawn from it. If we are studying each and every element of the population, the process is called *census method* and if we are studying only a sample, the process is called sample survey, *sample method* or *sampling*. For example, the Indian population census or a socio economic survey of a whole village by a college planning forum are examples of census studies. The national sample survey enquiries are examples of sample studies.

Advantages of Sampling

1. The sample method is comparatively more economical.
2. The sample method ensures completeness and a high degree of accuracy due to the small area of operation
3. It is possible to obtain more detailed information, in a sample survey than complete enumeration.
4. Sampling is also advocated where census is neither necessary nor desirable.
5. In some cases sampling is the only feasible method. For example, we have to test the sharpness of blades-if we test each blade, perhaps the whole of the product will be wasted; in such circumstances the census method will not be suitable. Under these circumstances sampling techniques will be more useful.
6. A sample survey is much more scientific than census because in it the extent of the reliability of the results can be known where as this is not always possible in census.

Variables and Attributes

A quantity which varies from one person to another or one time to another or one place to another is called a variable. It is actually a numerical value possessed by an item. For example, price of a given commodity, wages of workers, production and weights of students etc.

Attribute means a qualitative characteristic possessed by each individual in a group. It can't assume numerical values. For example, sex, honesty, colour etc.

This means that a variable will always be a quantitative characteristic. Data concerned with a quantitative variable is called *quantitative data* and the data corresponding to a qualitative variable is called *qualitative data*.

We can divide quantitative variables into two (i) discrete (ii) continuous. Those variables which can assume only distinct or particular values are called *discrete* or *discontinuous* variables. For example, the number of children per family, number rooms in a house etc. Those variables which can take any numerical value with in a certain range are known as *continuous* variables. Height of a boy is a continuous variable, for it changes continuously in a given range of heights of the boys. Similar is the case of weight,: production, price, demand, income, marks etc.

Types of Frequency Distribution

Erricker states "frequency distribution is a classification according to the number possessing the same values of the variables''. It is simply a table in which data are grouped into classes and the number of cases which fall in each class is recorded. Here the numbers are usually termed as 'frequencies'. There are discrete frequency distributions and continuous frequency distributions.

1. Discrete Frequency Distribution

If we have a large number of items in the data it is better to prepare a frequency array and condense the data further. Frequency array is prepared by listing once and consecutively all the values occurring in the series and noting the number of times each such value occurs. This is called discrete frequency distribution or ungrouped frequency distribution.

Illustration: The following data give the number of children per family in each of 25 families 1, 4, 3, 2, 1, 2, 0, 2, 1, 2, 3, 2, 1, 0, 2, 3, 0, 3, 2, 1, 2, 2, 1, 4, 2. Construct a frequency distribution.

No of children	Tally marks	No of families
0	$\vert\vert\vert$	3
1	$\cancel{\vert\vert\vert\vert}\ \vert$	6
2	$\cancel{\vert\vert\vert\vert}\ \cancel{\vert\vert\vert\vert}$	10
3	$\vert\vert\vert\vert$	4
4	$\vert\vert$	2
Total		25

2. Continuous Frequency Distribution

An important method of condensing and presenting data is that of the construction of a continuous frequency distribution or grouped frequency distribution. Here the data are classified according to class intervals.

The following are the rules generally adopted in forming a frequency table for a set of observations.

1. Note the difference between the largest and smallest value in the given set of observations
2. Determine the number classes into which the difference can be divided.
3. The classes should be mutually exclusive. That means they do not overlap.
4. Arrange a paper with 3 columns, classes, tally marks and frequency.
5. Write down the classes in the first column.
6. Go though the observations and put tally marks in the respective classes.
7. Write the sum of the tally marks of each class in the frequency column.
8. Note that the sum of the frequencies of all classes should be equal to the total number of observations.

Concepts of a Frequency Table

i. Class limits: The observations which constitute a class are called class limits. The left hand side observations are called lower limits and the right hand side observations are called upper limits.

ii. Working classes: The classes of the form 0-9, 10-19, 20-29,... are called working classes or nominal classes. They are obtained by the inclusive method of classification where both the limits of a class are included in the same class itself.

iii. Actual classes: If we are leaving either the upper limit or the lower limit from each class, it is called exclusive method of classification. The classes so obtained are called ‘actual classes’ or ‘true classes’. The classes – 0.5 - 9.5, 9.5 - 19.5, 19.5 - 29.5,... are the actual classes of the above working classes. The classes of the type 0-10, 10 - 20, 20 - 30,... are also treated as actual classes. There will be no break in the actual classes. We can convert working classes to the corresponding actual classes using the following steps.

1. Note the difference between one upper limit and the next lower limit.
2. Divide the difference by 2.
3. Subtract that value from the lower limits and add the same to the upper limits.

For example

Working Classes	Frequency	Actual Classes
1-2.9	2	0.95-2.95
3-4.9	8	2.95-4.95
5-6.9	10	4.95-6.95
7-8.9	5	6.95-8.95

iv. Class boundaries: The class limits of the actual classes are called actual class limits or class boundaries.

v. Class mark: The class marks or mid value of classes is the average of the upper limit and lower limit of that class. The mid value of working classes and the corresponding actual classes are the same. For example, the class mark of the classes 0 - 9, 10 - 19, 20 - 29,... are respectively 4.5, 14.5, 24.5,...

vi. Class interval: The class interval or width of a class is the difference between upper limit and lower limit of an actual class. It is better to note that the difference between the class limits of a working class is not the class interval. The class interval is usually denoted by ‘c’ or i or ‘h’.

Example

Construct a frequency distribution for the following data

70	45	33	64	50	25	65	74	30	20
55	60	65	58	52	36	45	42	35	40
51	47	39	61	53	59	49	41	20	55
46	48	52	64	48	45	65	78	53	42

Solution

Classes	Tally marks	Frequency
20-29	III	3
30-39	~~IIII~~	5
40-49	~~IIII~~ ~~IIII~~ II	12
50-59	~~IIII~~ ~~IIII~~	10
60-69	~~IIII~~ II	7
70-79	III	3
Total		40

Cumulative Frequency Distribution

An ordinary frequency distribution show the number of observations falling in each class. But there are instances where we want to know how many observations are lying below or above a particular value or in between two specified values. Such type of information is found in cumulative frequency distributions.

Cumulative frequencies are determined on either a less than basis or more than basis. Thus we get less than cumulative frequencies (<CF) and greater than or more than cumulative frequencies (>CF). Less than CF give the number of observations falling below the upper limit of a class and greater than CF give the number of observations lying above the lower limit of the class. Less than CF are obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulation is started from the lowest size of the class to the highest size, (usually from top to bottom). They are based on the upper limit of actual classes.

More than CF distribution is obtained by finding the cumulation or total of frequencies starting from the highest size of the class to the lowest class, (ie., from bottom to top) More than CF are based on the lower limit of the actual classes.

Classes	f	UL	<CF		LL	>CF	
0-10	2	10	2	2	0	3+7+10+8+5+2	35
10-20	5	20	2+5	7	10	3+7+10+8+5	33
20-30	8	30	2+5+8	15	20	3+7+10+8	28
30-40	10	40	2+5+8+10	25	30	3+7+10	20
40-50	7	50	2+5+8+10+7	32	40	3+10	13
50-60	3	60	2+5+8+10+7+3	35	50	3	3

Module 2

MEASURES OF CENTRAL TENDENCY

A measure of central tendency helps to get a single representative value for a set of usually unequal values. This single value is the point of location around which the individual values of the set cluster. Hence the averages are known also as *measures of location*.

The important measures of central tendencies or statistical averages are the following.

1. Arithmetic Mean
2. Geometric Mean
3. Harmonic Mean
4. Median
5. Mode

Weighted averages, positional values, viz., quartiles, deciles and percentiles, also are considered in this chapter.

Criteria or Desirable Properties of an Average

1. *It should be rigidly defined:* That is, it should have a formula and procedure such that different persons who calculate it for a set of values get the same answer.

2. *It should have sampling stability:* A number of samples can be drawn from a population. The average of one sample is likely to be different from that of another. It is desired that the average of any sample is not much different from that of any other.

1. Arithmetic Mean

The arithmetic mean (AM) or simply mean is the most popular and widely used average. It is the value obtained by dividing sum of all given observations by the number of observations. AM is denoted by $\bar{x}$ (x bar).

Definition for a raw data

For a raw data or ungrouped data if $x_1, x_2, x_3, \ldots, x_n$ are n observations,

then $\bar{x} = \dfrac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$

ie., $\bar{x} = \dfrac{\sum x}{n}$ where the symbol Σ (sigma) denotes summation.

Example 1

Calculate the AM of 12, 18, 14, 15, 16

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{12+18+14+15+16}{5} = \frac{75}{5} = \mathbf{15}$$

Definition for a frequency data

For a frequency data if $x_1, x_2, x_3, \ldots, x_n$ are 'n' observations or middle values of 'n' classes with the corresponding frequencies $f_1, f_2, \ldots, f_n$ then AM is given by

$$\bar{x} = \frac{f_1 \times x_1 + f_2 \times x_2 + \ldots + f_n \times x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum fx}{\sum f}$$

ie., $\bar{x} = \dfrac{\sum fx}{N}$ where N = $\sum f$ = Total frequency

Example 2

The following data indicate daily earnings (in rupees) of 40 workers in a factory.

Daily earnings in ₹	:	5	6	7	8	9
No of workers	:	3	8	12	10	7

Calculate the average income per worker.

X d = x − 320

Solution

305 −15
(320) 0
332 12
350 30

Daily Earnings in ₹	No. of workers (f)	fx
5	3	15
6	8	48
7	12	84
8	10	80
9	7	63
Total	40	290

$$\bar{x} = \frac{\sum fx}{N} = \frac{290}{40} = \mathbf{7.25}$$

Average income per worker is ₹ **7.25**

Example 3

Calculate the AM of the following data

Class	:	04	4-8	8-12	12-16
Frequency	:	1	4	3	2

Solution

Class	f	Mid values (x)	fx
0-4	1	2	2
4-8	4	6	24
8-12	3	10	30
12-16	2	14	28
Total	10		84

$$\bar{x} = \frac{\sum fx}{N} = \frac{84}{10} = \mathbf{8.4}$$

Shortcut Method: Raw data

Suppose the values of a variable under study are large, choose any value in between them. Preferably a value that lies more or less in the middle, called arbitrary origin or assumed mean, denoted by A. Take deviations of every value from the assumed mean A.

Let d = x – A, Taking summation of both sides and dividing by n, we get

$$\bar{x} = A + \frac{\sum d}{n}$$

Example 4

Calculate the AM of 305, 320, 332, 350

Solution

X	d = x – 320
305	– 15
320	0
332	12
350	30
	27

$$\bar{x} = A + \frac{\sum d}{n}$$

$$= 320 + \frac{27}{4}$$

= 320+6.75

= **326.75**

Shortcut Method: Frequency Data

When the frequencies and the values of the variable x are large the calculation of AM is tedious. So a simpler method is adopted. The deviations of the mid values of the classes are taken from a convenient origin. Usually the mid value of the class with the maximum frequency is chosen as the arbitrary origin or assumed mean. Thus change x values to ‘d’ values by the rule,

$$d = \frac{x - A}{c}$$

where A-assumed mean, c-class interval, x-mid values. Then the formula for calculating AM is given by

$$\bar{x} = \mathbf{A} + \frac{\sum \mathbf{fd}}{\mathbf{N}} \times \mathbf{c}$$

Example 5

Calculate AM from the following data

Weekly wages	:	0-10	10-20	20-30	30-40	40-50
Frequency	:	3	12	20	10	5

Solution

Weekly wages	f	Mid value x	$d = \frac{x-25}{10}$	fd	
0-10	3	5	– 2	– 6	– 18
10-20	12	15	– 1	– 12	
20-30	20	25	0	0	
30-40	10	35	1	10	20
40-50	5	45	2	10	
Total	50			2	

$$\bar{x} = A + \frac{\sum fd}{N} \times c = 25 + \frac{2}{50} \times 10 = 25 + 0.4 = \mathbf{25.4}$$

Properties

1. *The AM is preserved under a linear transformation of scale.*

   That is, if x_i is changed to y_i by the rule

   $y_i = a + b\,x_i$, then $\bar{y} = a + b\,\bar{x}$, which is also linear.
2. *The mean of a sum of variables is equal to the sum of the means of the variables.*
3. *Algebraic sum of the deviations of every observation from the A.M. zero.*
4. *If n_1 observations have an A.M $\bar{x}_1$ and n_2 observations have an AM $\bar{x}_2$ then the AM of the combined group of $n_1 + n_2$ observations is given by* $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

Example 6

Let the average mark of 40 students of class A be 38; the average mark of 60 students of another class B is 42. What is the average mark of the combined group of 100 students?

Here $n_1 = 40$, $\bar{x}_1 = 38$, $n_2 = 60$, $\bar{x}_2 = 42$

Here $\bar{x} = \dfrac{n_1.\bar{x}_1 + n_2.\bar{x}_2}{n_1 + n_2} = \dfrac{(40\times 38)+(60\times 42)}{40+60}$

$= \dfrac{1520 + 2520}{100} = \dfrac{4040}{100} = \mathbf{40.4}$

Note

The above property can be extended as follows. When there are three groups, the combined mean is given by

5. *The algebraic sum of the squares of the observations from AM is always minimum. ie.,* is always minimum.

Merits and Demerits

Merits

The most widely used arithmetic mean has the following merits.

1. It is rigidly defined. Clear cut mathematical formulae are available.
2. It is based on all the items. The magnitudes of all the items are considered for its computation.
3. It lends itself for algebraic manipulations. Total of a set, Combined Mean etc., could be calculated.
4. It is simple to understand and is not difficult to calculate. Because of its practical use, provisions are made in calculators to find it.
5. It has sampling stability. It does not vary very much when samples are repeatedly taken from one and the same population.
6. It is very much useful in day-to-day activities, later chapters in Statistics and many disciplines of knowledge.
7. Many forms of the formula are available. The form appropriate and easy for the data on hand can be used.

Demerits

1. It is unduly affected by extreme items. One greatest item may pull up the mean of the set to such an extent that its representative character is questioned. For example, the mean mark is 35 for the 3 students whose individual marks are 0, 5 and 100.
2. Theoretically, it cannot be calculated for open-end data.
3. It cannot be found graphically.
4. It is not defined to deal with qualities.

Weighted Arithmetic Mean

In calculating simple arithmetic mean it was assumed that all items are of equal importance. This may not be true always. When items vary in importance they must be assigned weights in proportion to their relative importance. Thus, a weighted mean is the mean of weighted items. The weighted arithmetic mean is sum of the product of the values and their respective weights divided by the sum of the weights.

Symbolically, if $x_1, x_2, x_3, \ldots x_n$ are the values of items and $w_1, w_2, \ldots w_n$ are their respective weights, then

$$\text{WAM} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n} = \frac{\sum wx}{\sum w}$$

Weighted AM is preferred in computing the average of percentages, ratios or rates relating to different classes of a group of observations. Also WAM is invariably applied in the computation of birth and death rates and index numbers.

Example 7

A student obtains 60 marks in Statistics, 48 marks in Economics, 55 marks in law, 72 marks in Commerce and 45 marks in taxation in an examination. The weights of marks respectively are 2, 1, 3, 4, 2. Calculate the simple AM and weighted AM of the marks.

Solution

$$\text{Simple AM} = \frac{\sum x}{n} = \frac{60+48+55+72+45}{5} = \frac{280}{5} = \mathbf{56}$$

Marks (x)	Weights (w)	wx
60	2	120
48	1	48
55	3	165
72	4	288
45	2	90
	12	711

$$\text{WAM} = \frac{\sum wx}{\sum w} = \frac{711}{12} = \mathbf{59.25}$$

Geometric Mean

Geometric mean (GM) is the appropriate root (corresponding to the number of observations) of the product of observations. If there are n observations GM is the n-th root of the product of n observations.

Definition for a raw data

If $x_1, x_2, x_3, \ldots, x_n$ are n observations;

$$\text{GM} = \sqrt[n]{x_1, x_2, \ldots\ldots x_n}$$

Using logarithms, we can calculate GM using the formula,

$$\text{GM} = Anti\log\left(\frac{\sum \log x}{n}\right)$$

Definition for a frequency distribution

For a frequency distribution if $x_1, x_2, x_3, \ldots, x_n$ are n observations with the corresponding frequencies $f_1, f_2, \ldots, f_n$

$$\text{GM} = \sqrt[N]{x_1^{f_1}, x_2^{f_2}, \ldots\ldots x_n^{f_n}}$$

using logarithm,

$$\textbf{GM} = \textbf{Antilog}\left(\frac{\sum \mathbf{f \log x}}{\mathbf{N}}\right) \text{ where } N = \sum f .$$

Note

1. GM is the appropriate average for calculating index number and average rates of change.
2. GM can be calculated only for non zero and non negative values.
3. Weighted GM = $Anti\log\left(\frac{\sum w \log x}{\sum w}\right)$

where w's are the weights assigned.

Example 8

Calculate GM of 2, 4, 8

Solution

$$GM = \sqrt[n]{x_1, x_2, \ldots\ldots x_n} = \sqrt[3]{2 \times 4 \times 8} = \sqrt[3]{64} = \mathbf{4}$$

Example 9

Calculate GM of 4, 6, 9, 1 1 and 15

Solution

x	logx
4	0.6021
6	0.7782
9	0.9542
11	1.0414
15	1.1761
	4.5520

$$GM = Anti\log\left(\frac{\sum \log x}{n}\right)$$

$$= Anti\log\left(\frac{4.5520}{5}\right)$$

= Antilog0.9104

= **8.136**

Example 10

Calculate GM of the following data

Classes	:	1-3	4-6	7-9	10-12
Frequency	:	8	16	15	3

Solution

Classes	f	X	logx	f.logx
1-3	8	2	0.3010	2.4080
4-6	16	5	0.6990	11.1840
7-9	15	8	0.9031	13.5465
10-12	3	11	1.0414	3.1242
Total	42			30.2627

$$GM = \mathbf{Antilog}\left(\frac{\acute{g}\,\mathbf{f}\ \mathbf{log}\ \mathbf{x}}{\mathbf{N}}\right)$$

= Antilog(30.2627/42)

= Antilog 0.7205 = 5.254

Merits and Demerits

Merits

1. It is rigidly defined. It has clear cut mathematical formula.
2. It is based on all the items. The magnitude of every item is considered for its computation.
3. It is not as unduly affected by extreme items as A.M. because it gives less weight to large items and more weight to small items.
4. It can be algebraically manipulated. The G.M. of the combined set can be calculated from the GMs and sizes of the sets.
5. It is useful in averaging ratios and percentages. It is suitable to find the average rate (not amount) of increase or decrease and to compute index numbers.

Demerits

1. It is neither simple to understand nor easy to calculate. Usage of logarithm makes the computation easy.
2. It has less sampling stability than the A.M.
3. It cannot be calculated for open-end data.
4. It cannot be found graphically.
5. It is not defined for qualities. Further, when one item is zero, it is zero and thereby loses its representative character. It cannot be calculated even if one value or one mid value is negative.

Harmonic Mean

The harmonic mean (HM) of a set of observations is defined as the reciprocal of the arithmetic mean of the reciprocals of the observations.

Definition for a raw data

If $x_1, x_2, x_3, \ldots, x_n$ are 'n' observations

$$\text{HM} = \frac{1}{\dfrac{\left(\dfrac{1}{x_1}+\dfrac{1}{x_2}+\ldots+\dfrac{1}{x_n}\right)}{n}} = \frac{n}{\dfrac{1}{x_1}+\dfrac{1}{x_2}+\ldots+\dfrac{1}{x_n}} = \frac{\mathbf{n}}{\sum\left(\dfrac{\mathbf{1}}{\mathbf{x}}\right)}$$

Definition for a frequency data

If $x_1, x_2, x_3, \ldots, x_n$ are 'n' observations with the corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$

$$\text{then HM} = \frac{N}{f_1 \times \dfrac{1}{x_1} + f_2 \times \dfrac{1}{x_2} + \ldots + f_n \times \dfrac{1}{x_n}} = \frac{\mathbf{N}}{\sum\left(\dfrac{\mathbf{f}}{\mathbf{x}}\right)}$$

where $N = \sum f$

Note 1 HM can be calculated only for non zero and non negative values.

Note 2 HM is appropriate for finding average speed when distance travelled at different speeds are equal. Weighted HM is appropriate when the distances are unequal. HM is suitable to study rates also.

Note 3 Weighted HM = $\dfrac{N}{\sum\left(\dfrac{w}{x}\right)}$ where w's are the weighted assigned.

Example 11

Calculate the HM of 2, 3, 4, 5 and 7

Solution

$$\text{HM} = \frac{n}{\sum \dfrac{1}{x}} = \frac{5}{\dfrac{1}{2}+\dfrac{1}{3}+\dfrac{1}{4}+\dfrac{1}{5}+\dfrac{1}{7}}$$

$$= \frac{5}{\dfrac{210+140+105+84+60}{420}} = \frac{5 \times 420}{599} = \mathbf{3.50}$$

Example 12

Calculate HM of 5, 11, 12,16, 7, 9, 15, 13, 10 and 8

Solution

X	1/x	X	1/x
5	0.2000	9	0.1111
11	0.0909	15	0.0667
12	0.0833	13	0.0769
16	0.0625	10	0.1000
7	0.1429	8	0.1250
		Total	1.0593

$$\text{HM} = \frac{\mathbf{n}}{\sum\left(\dfrac{\mathbf{1}}{\mathbf{x}}\right)} = (10/1.0593) = 9.44$$

Merits and Demerits

Merits

1. It is rigidly defined. It has clear cut mathematical formula.
2. It is based on all the items. The magnitude of every item is considered for its computation.
3. It is affected less by extreme items than A.M. or even G.M.
4. It gives lesser weight to larger items and greater weight to lesser items.

5. It can be algebraically manipulated. The H.M. of the combined set can be calculated from the H.M.s and sizes of the sets. For example,

$$HM_{12} = \frac{N_1 + N_2}{\frac{N_1}{HM_1} + \frac{N_2}{HM_2}}$$

6. It is suitable to find the average speed.

Demerits

1. It is neither simple to understand nor easy to calculate.
2. It has less sampling stability than the A.M.
3. Theoretically, it cannot be calculated for open-end data.
4. It cannot be found graphically.
5. It is not defined for qualities. It is not calculated when atleast one item or one mid value is zero or negative.
6. It gives undue weightage to small items and least weightage to largest items. It is not used for analysing business or economic data.

Median

Median is defined as the middle most observation when the observations are arranged in ascending or descending order of magnitude. That means the number of observations preceding median will be equal to the number of observations succeeding it. Median is denoted by M.

Definition for a raw data

For a raw data if there are odd number of observations, there will be only one middle value and it will be the median. That means, if there are n observations arranged in order of their magnitude, the size of (n+1)/2 – th observation will be the median. If there are even number of observations the average of two middle values will be the median. That means, median will be the average of $n/2^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

Definition for a frequency data

For a frequency distribution median is defined as the value of the variable which divides the distribution into two equal parts. The median can be calculated using the following formula.

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

where, l	-	lower limit of median class
Median class	-	the class in which $N/2^{th}$ observation falls
N	-	total frequency
m	-	cumulative frequency up to median class
c	-	class interval of the median class
f	-	frequency of median class

found to lie with in that interval.

Example 13

Find the median height from the following heights (in cms.) of 9 soldiers.
160, 180, 175, 179, 164, 178, 171, 164, 176

Solution

Step 1. Heights are arranged in ascending order:
160, 164, 164, 171, 175, 176, 178, 179, 180.

Step 2. Position of median = $\frac{n+1}{2}$ is calculated. It is $\frac{9+1}{2} = 5$.

Step 3. Median is identified (5^{th} value) M = **175cms.**

It is to be noted that $\frac{n+1}{2}$ may be a fraction, in which case, median is found as follows.

Example 14

Find the median weight from the following weights (in Kgs) of 10 soldiers. 75, 71, 73, 70, 74, 80, 85, 81, 86, 79

Solution

Step 1. Weights are arranged in ascending order:

70, 71, 73, 74, 75, 79, 80, 81, 85, 86

Step 2. Position of median $\frac{n+1}{2} = \frac{10+1}{2} = 5\frac{1}{2}$ is calculated

Step 3. Median is found. It is the mean of the values at 5th and 6th positions and so M = $\frac{75+79}{2}$ = **77Kgs.**

Example 15

Find the median for the following data.

Height in cms	:	160	164	170	173	178	180	182
No. of soldiers	:	1	2	10	22	19	14	2

Solution

Step 1. Heights are arranged in ascending order. Cumulative frequencies (c.f) are found. (They help to know the values at different positions)

Height in cms.	No. of Soldiers	C.f.
160	1	1
164	2	3
170	10	13
173	22	35
178	19	54
180	14	68
182	2	70
Total	70	–

Step 2. Position of median, $\frac{N+1}{2} = \frac{70+1}{2} = 35\frac{1}{2}$ is calculated.

Step 3. Median is identified as the average of the values at the positions 35 and 36. The values are 173 and 178 respectively.

$$\therefore\ M = \frac{173+178}{2} = \mathbf{175.5cm}$$

Example 16

Calculate median for the following data

Class	:	0-5	5-10	10-15	15-20	20-25
f	:	5	10	15	12	8

Solution

Class	f	CF
0-5	5	5
5-10	10	15
10-15	15	30
15-20	12	42
20-25	8	50
Total	50	

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

Median class is 10-15

Here $l = 10,\ N/2 = 50/2 = 25,\ c = 5,\ m = 15,\ f = 15$

$$\therefore M = 10 + \frac{(25-15)5}{15}$$

$$= 10 + \frac{10 \times 5}{15} = 10 + \frac{10}{3} = 10+3.33 = \mathbf{13.33}$$

Example 17

Calculate median for the data given below.

Classes :	0-6	7-13	14-20	21-27	28-34	35-41
f :	8	17	28	15	9	3

Solution:

Class	f	Actual class	CF
0-6	8	– 0.5-6.5	8
7-13	17	6.5-13.5	25
14-20	28	13.5-20.5	53
21-27	15	20.5-27.5	68
28-34	9	27.5-34.5	77
35-41	3	34.5-41.5	80
Total	80		

Median class is 13.5-20.5, l = 13.5, N/2 = 80/2 = 40

c = 7, m = 25, f = 28

$$M = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c = 13.5 + \frac{(40-25)}{28} \times 7$$

$$= 13.5 + \frac{15 \times 7}{28} = 13.5 + \frac{15}{4}$$

= 13.5+3.75

= **17.25**

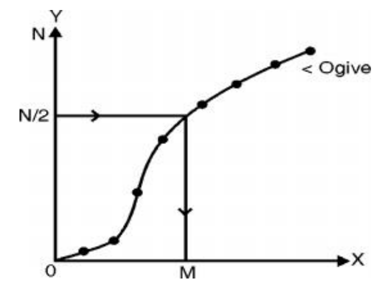
Graphical Determination of Median

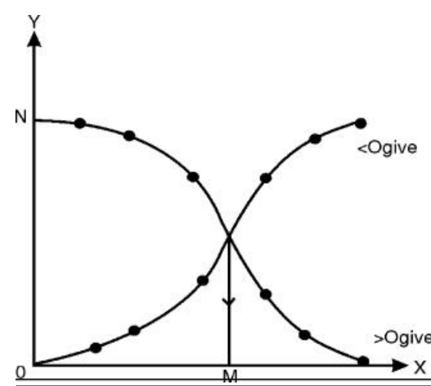
Median can be determined graphically using the following

Steps

1. Draw the less than or more than ogive
2. Locate N/2 on the Y axis.
3. At N/2 draw a perpendicular to the Y axis and extend it to meet the ogive
4. From the point of intersection drop a perpendicular to the X axis
5. The point at which the perpendicular meets the X axis will be the median value.

Median can also be determined by drawing the two ogives, simultaneously. Here drop a perpendicular from the point of intersection to the X axis. This perpendicular will meet at the median value.





Merits and Demerits

Merits

1. It is not unduly affected by extreme items.
2. It is simple to understand and easy to calculate.
3. It can be calculated for open end data
4. It can be determined graphically.
5. It can be used to deal with qualitative data.

Demerits

1. It is not rigidly defined. When there are even number of individual observations, median is approximately taken as the mean of the two middle most observations.
2. It is not based on the magnitude of all the items. It is a positional measure. It is the value of the middle most item.
3. It cannot be algebraically manipulated. For example, the median of the combined set can not be found from the medians and the sizes of the individual sets alone.
4. It is difficult to calculate when there are large number of items which are to be arranged in order of magnitude.
5. It does not have sampling stability. It varies more markedly than A M from sample to sample although all the samples are from one and the same population.
6. Its use is lesser than that of AM.

Mode

Mode is that value of the variable, which occur maximum number of times in a set of observations. Thus, mode is the value of the variable, which occur most frequently. Usually statements like, 'average student', 'average buyer', 'the typical firm', etc. are referring to mode of the phenomena. Mode is denoted by Z or Mo. For a raw data as well as for a discrete frequency distribution we can locate mode by inspection.

For a frequency distribution mode is defined as the value of the variable having the maximum frequency. For a continuous frequency distribution it can be calculated using the formula given below:

$$Z = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

where l : lower limit of modal class

Modal class : Class having the maximum frequency

Δ_1 : difference between the frequency of modal class and that of the premodal class

Δ_2 : difference between frequency of the modal class and that of the post modal class

c : class interval

For applying this formula, the class intervals should be (i) of equal size (ii) in ascending order and (iii) in exclusive form.

Example 18

Determine the mode of

420, 395, 342, 444, 551, 395, 425, 417, 395, 401, 390

Solution

Mode = **395**

Example 19

Determine the mode

Size of shoes	:	3	4	5	6	7	8	9
No of pairs sold	:	10	25	32	38	61	47	34

Solution

Mode = Z = 7

Example 20

Calculate mode for the following data

Classes	:	0-9	10 - 19	20-29	30-39	40-49	50-59
f	:	5	10	17	33	22	13

Solution

Classes	f	Atual class
0-9	5	– 0.5-9.5
10-19	10	9.5-19.5
20-29	17	19.5-29.5
30-39	33	29.5-39.5
40-49	22	39.5-49.5
50-59	13	49.5-59.5

$$Z = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

Modal class is 29.5-39.5

l = 29.5

$\Delta_1 = 33 - 17 = 16$

$\Delta_2 = 33 - 22 = 11,\ c = 10$

$$= 29.5 + \frac{16}{16+11} \times 10$$

= 29.5+5.92 = **35.42**

For a symmetrical or moderately assymmetrical distribution, the empirical relation is

Mean – Mode = 3 (Mean – Median)

This relation can be used for calculating any one measure, if the remaining two are known.

Example 21

In a moderately assymmetrical distribution Mean is 24.6 and Median 25.1. Find the value of mode.

Solution

We have

$$\begin{aligned}
\text{Mean} - \text{Mode} &= 3(\text{Mean} - \text{Median}) \\
24.6 - Z &= 3(24.6 - 25.1) \\
24.6 - Z &= 3(-0.5) = -1.5 \\
Z &= 24.6 + 1.5 = \mathbf{26.1}
\end{aligned}$$

Example 22

In a moderately assymmetrical distribution Mode is 48.4 and Median 41.6. Find the value of Mean

Solution

We have,

$$\begin{aligned}
\text{Mean} - \text{Mode} &= 3(\text{Mean} - \text{Median}) \\
\bar{x} - 48.4 &= 3(\bar{x} - 41.6) \\
\bar{x} - 48.4 &= 3\bar{x} - 124.8 \\
3\bar{x} - \bar{x} &= 124.8 - 48.4 \\
2\bar{x} &= 76.4 \\
\bar{x} &= 76.4 \div 2 = \mathbf{38.2}
\end{aligned}$$

Merits and Demerits

Merits

1. Mode is not unduly affected by extreme items.
2. It is simple to understand and easy to calculate
3. It is the most typical or representative value in the sense that it has the greatest frequency density.
4. It can be calculated for open-end data.
5. It can be determined graphically. It is the x-coordinate of the peak of the frequency curve.
6. It can be found for qualities also. The quality which is observed more often than any other quality is the modal quality.

Demerits

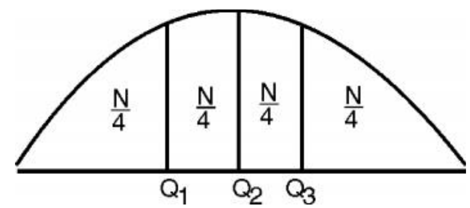
1. It is not rigidly defined.
2. It is not based on all the items. It is a positional value.
3. It cannot be algebraically manipulated. The mode of the combined set cannot be determined as in the case of AM.
4. Many a time, it is difficult to calculate. Sometimes grouping table and frequency analysis table are to be formed.
5. It is less stable than the A.M.
6. Unlike other measures of central tendency, it may not exist for some data. Sometimes there may be two or more modes and so it is said to be ill defined.
7. It has very limited use. Modal wage, modal size of shoe, modal size of family, etc., are determined. Consumer preferences are also dealt with.

Partition Values

We have already noted that the total area under a frequency curve is equal to the total frequency. We can divide the distribution or area under a curve into a number of equal parts choosing some points like median. They are generally called *partition values or quantiles*. The important partition values are *quartiles, deciles and percentiles.*

Quartiles

Quartiles are partition values which divide the distribution or area under a frequency curve into 4 equal parts at 3 points namely Q_1, Q_2, and Q_3 . Q_1 is called *first quartile or lower quartile*, Q_2 is called *second quartile, middle quartile or median* and Q_3 is called *third quartile or upper quartile*. In other words Q_1 is the value of the variable such that the number of observations lying below it, is N/4 and above it is 3N/4. Q_2 is the value of the variable such that the number of observations on either side of it is equal to N/2. And Q_3 is the value of the variable such that the number of observations lying below Q_3 is 3N/4 and above Q_3 is N/4.



Deciles and Percentiles

Deciles are partition values which divide the distribution or area under a frequency curve into 10 equal parts at 9 points namely D_1, D_2, ..., D_9.

Percentiles are partition values which divide the distribution into 100 equal parts at 99 points namely P_1, P_2, P_3, ... P_{99}. Percentile is a very useful measure in education and psychology. Percentile ranks or scores can also be calculated. Kelly's measure of skewness is based on percentiles.

Calculation of Quartiles

The method of locating quartiles is similar to that method used for finding median. Q_1 is the value of the item at $(n + 1)/4^{th}$ position and Q_3 is the value of the item at $3(n + 1) / 4^{th}$ position when actual values are known. In the case of a frequency distribution Q_1 and Q_3 can be calculated as follows.

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)}{f} \times c$$

where l_1 - lower limit of Q_1 class

Q_1 class - the class in which $N/4^{th}$ item falls

m - cumulative frequency up to Q_1 class

c - class interval

f - frequency of Q_1 class

$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m\right)}{f} \times c$$

where l_3 - lower limit of Q_3 class

Q_3 class - the class in which $3N/4^{th}$ item falls

m - cumulative frequency up to Q_3 class

c - class interval

f - frequency of Q_3 class

We can combine these three formulae and can be written as

$$Q_i = l_i + \frac{\left(\frac{iN}{4} - m\right)}{f} \times c, \quad i = 1, 2, 3$$

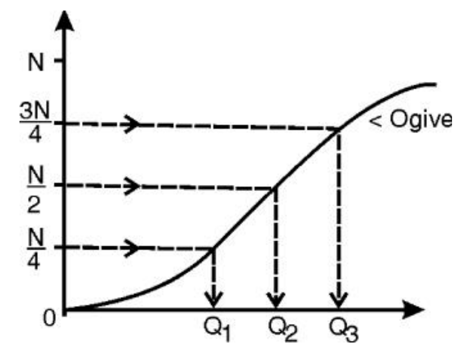
In a similar fashion deciles and percentiles can be calculated as

$$D_i = l_i + \frac{\left(\frac{iN}{10} - m\right)}{f} \times c, \quad i = 1, 2, 3, \ldots. 9$$

$$P_i = l_i + \frac{\left(\frac{iN}{100} - m\right)}{f} \times c, \quad i = 1, 2, 3, \ldots., 99$$

Graphical Determination of Quartiles

Quartiles can be determined graphically by drawing the ogives of the given frequency distribution. So draw the less than ogive of the given data. On the Y axis locate N/4, N/2 and 3N/4. At these points draw perpendiculars to the Y axis and extend it to meet the ogive. From the points of intersection drop perpen-diculars to the X axis. The point corresponding to the CF, N/4 is Q_1 corresponding to the CF N/2 is Q_2 and corres-ponding to the CF 3N/4 is Q_3.



Example 23

Find , Q_1, Q_3, D_2, D_9, P_{16}, P_{65} for the following data. 282, 754, 125, 765, 875, 645, 985, 235, 175, 895, 905, 112 and 155.

Solution

Step 1. Arrange the values in ascending order
112, 125, 155, 175, 235, 282, 645, 754, 765, 875, 895, 905 and 985.

Step 2. Position of Q_1 is $\frac{n+1}{4} = \frac{13+1}{4} = \frac{14}{4} = 3.5$

Similarly positions of Q_3, D_2, D_9, P_{16} and P_{65} are 10.5, 2.8, 12.6, 2.24 and 9. 1 respectively.

Step 3.

$$Q_1 = 155 + 0.5(175 - 155) = \mathbf{165}$$

$$Q_3 = 875 + 0.5(895 - 875) = \mathbf{885}$$

$$D_2 = 125 + 0.8(155 - 125) = \mathbf{149.0}$$

$$D_9 = 905 + 0.6(985 - 905) = \mathbf{953}$$

$$P_{16} = 125 + 0.24(155 - 125) = \mathbf{132.20}$$

$$P_{65} = 765 + 0.1(875 - 765) = \mathbf{776.0}$$

Note

The value of the 12.6-th position (D_9) is obtained as value of 12-th position + 0.6 (value at 13-th position - value at 12-th position)

Example 24

Find Q_1, Q_3, D_4, P_{20} and P_{99} for the data given below.

Mark	:	25	35	40	50	52	53	67	75	80
No of students	:	3	29	32	41	49	54	38	29	27

Marks	No of students	Cumulative frequency
25	3	3
35	29	32
40	32	64
50	41	105
52	49	154
53	54	208
67	38	246
75	29	275
80	27	302

Step 1. The cumulative frequencies of marks given in ascending order are found

Step 2. The positions of Q_1, Q_3, D_4, P_{20} and P_{99} are found. They are

$$\frac{N+1}{4} = \frac{303}{4} = 75.75$$

$$\frac{3(N+1)}{4} = 3 \times \frac{303}{4} = 227.25$$

$$\frac{4(N+1)}{10} = \frac{40 \times 303}{10} = 121.20$$

$$\frac{20(N+1)}{100} = \frac{20 \times 303}{100} = 60.60$$

$$\frac{99(N+1)}{100} = \frac{99 \times 303}{100} = 299.97$$

Step 3. The marks of students at those positions are found

$$Q_1 = 50 + 0.75(50 - 50) = \mathbf{50} \text{ Marks}$$

$$Q_3 = 67 + 0.25(67 - 67) = \mathbf{67} \text{ Marks}$$

$$D_4 = 52 + 0.20(52 - 52) = \mathbf{52} \text{ Marks}$$

$$P_{20} = 40 + 0.60(40 - 40) = \mathbf{40} \text{ Marks}$$

$$P_{99} = 80 + 0.97(80 - 80) = \mathbf{80} \text{ Marks}$$

Note

Refer the above example to know the method of finding the values of the items whose positions are fractions.

Example 25

Calculate quartiles for the following data

Classes	:	30-35	35-40	40-45	45-50	50-55	55-60	60-65
Freq.	:	10	16	18	27	18	8	3

Solution

Class	f	CF
30-35	10	10
35-40	16	26
40-45	18	44
45-50	27	71
50-55	18	89
55-60	8	97
60-65	3	100
Total	100	

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)c}{f}$$

$$= 35 + \frac{(25 - 10)5}{16}$$

$$= 35 + \frac{15 \times 5}{16} = 35 + \frac{75}{16}$$

$$= 35 + 4.68 = \mathbf{39.68}$$

$$Q_2 = l_2 + \frac{\left(\frac{N}{2} - m\right)c}{f}$$

$$= 45 + \frac{(50 - 44)5}{27}$$

$$= 45 + \frac{6 \times 5}{27}$$

$$= 45 + \frac{10}{9} = 45 + 1.11 = \mathbf{46.11}$$

$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m\right)c}{f}$$

$$= 50 + \frac{(75 - 71)5}{18}$$

$$= 50 + \frac{4 \times 5}{18} = 50 + \frac{10}{9} = 50 + 1.11 = \mathbf{51.11}$$

Very Short Answer Questions

17. What is central tendency?
18. Define Median and mode.
19. Define harmonic mean
20. Define partition values
21 State the properties of AM.
22 In a class of boys and girls the mean marks of 10 boys is 38 and the mean marks of 20 girls 45. What is the average mark of the class?
23. Define deciles and percentiles.
24 Find the combined mean from the following data.

	Series x	Series y
Arithmetic mean	12	20
No of items	80	60

Short Essay Questions

25 Define mode. How is it calculated. Point out two
26. Define AM, median and mode and explain their uses
27. Give the formulae used to calculate the mean, median and mode of a frequency distribution and explain the symbols used in them.
28. How will you determine three quartiles graphically from a less than ogive?
29. Three samples of sizes 80, 40 and 30 having means 12.5, 13 and 11 respectively are combined. Find the mean of the combined sample.
30. Explain the advantages and disadvantages of arithmetic mean as an average.
31. For finding out the 'typical' value of a series, what measure of central tendency is appropriate?

32 Explain AM and HM. Which one is better? And Why?
33. Prove that the weighted arithmetic mean of first n natural numbers whose weights are equal to the corresponding number is equal to $(2n+1)/3$
34. Show that GM of a set of positive observation lies between AM & AM.
35. What are the essential requisites of a good measure of central tendency? Compare and contrast the commonly employed measures in terms of these requisites.
36. Discuss the merits and demerits of the various measures of central tendency. Which particular measure is considered the best and why? Illustrate your answer.
37.. What is the difference between simple and weighted average? Explain the circumstances under which the latter should be used in preference to the former.
38. Find the average rate of increase in population which in the first decade has increased 12 percent, in the next by 16 per cent, and in third by 21 percent.
39.. A person travels the first mile at 10 km. per hour, the second mile at 8 km. per hour and the third mile at 6 km. per hour. What is his average speed?

Long Essay Questions

40. Compute the AM, median and mode from the following data

Age last birth day	:	15-19	20-24	25-29	30-34	35-39	40-44
No of persons	:	4	20	38	24	10	

41. Calculate Arithmetic mean, median and mode for the following data.

Age	:	55-60	50-55	45-50	40-45	35-40	30-35	25-30	20-25
No of people	:	7	13	15	20	30	33	28	14

42. Calculate mean, median and mode from the following data

Class	Frequency
Up to 20	52
20-30	161

30-40	254
40-50	167
50-60	78
60-80	64
Over 80	52

43. Calculate mean, median and mode

Central wage in Rs. :	15	20	25	30	35	40	45
No. of wage earners:	3	25	19	16	4	5	6

44. (i) Find the missing frequencies in the following distribution given that N = 100 and median of the distribution is 110.

(ii) Calculate the arithmetic mean of the completed frequency distribution.

Class	**:**	**20-40**	**40-60**	**60-80**	**80-100**	**100-120**
Frequency	**:**	**6**	**9**	**-**	**14**	**20**
Class	**:**	**120-140**	**140-160**	**160-180**	**180-200**	
Frequency	**:**	**15**	**-**	**8**	**7**	

UNIT - II
MEASURES OF DISPERSION

By dispersion we mean *spreading* or *scatteredness* or *variation*. It is clear from the above example that dispersion measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of second order.

Desirable properties of an ideal measure of dispersion

The following are the requisites for an ideal measure of dispersion.

1. It should be rigidly defined and its value should be definite.
2. It should be easy to understand and simple to calculate.
3. It should be based on all observations.
4. It should be capable of further algebraic treatment.
5. It should be least affected by sampling fluctuations.

Methods of Studying Variation

The following measures of variability or dispersion are commonly used.

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation

Here the first two are called positional measures of dispersion. The other two are called calculation measures of deviation.

Absolute and Relative Dispersion

Absolute measures and relative measures are the two kinds of measures of dispersion. The formers are used to assess the variation among a set of values. The latter are used whenever the variability of two or more sets of values are to be compared. Relative measures give pure numbers, which are free from the units of measurements of the data. Even data in different units and with unequal average values can be compared on the basis of relative measures of dispersion. Less is the value of a relative measure, less is the variation of the set and more is the consistency. The terms, stability, homogeneity, uniformity and consistency are used as if they are synonyms.

1. Range

Definition Range is the difference between the greatest (largest) and the smallest of the given values.

In symbols, **Range = L – S** where L is the greatest value and S is the smallest value.

The corresponding relative measure of dispersion is defined as

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

Example 1

The price of a share for a six-day week is fluctuated as follows:

₹156 ₹ 165 ₹ 148 ₹ 151 ₹ 147 ₹ 162

Calculate the Range and its coefficient.

Range = L – S = ₹ 165 – ₹ 147 = ₹ 18

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{165-147}{165+147} = \mathbf{0.0577}$$

Example 2

Calculate coefficient of range from the following data:

Mark:	10-20	20-30	30-40	40-50	50-60
No.of students:	8	10	12	8	4

Solution

$$\text{Coefficient of Range} = \frac{L-S}{L+S} = \frac{60-10}{60+10} = \mathbf{0.7143}$$

Merits and Demerits

Merits

1. It is the simplest to understand and the easiest to calculate.
2. It is used in Statistical Quality Control.

Demerits

1. Its definition does not seem to suit the calculation for data with class intervals. Further, it cannot be calculated for open-end data.
2. It is based on the two extreme items and not on any other item.
3. It does not have sampling stability. Further, it is calculated for samples of small sizes only.
4. It could not be mathematically manipulated further.
5. It is a very rarely used measure. Its scope is limited to very few considerations in Quality Control.

2. Quartile Deviation

Definition

Quartile deviation is half of the difference between the first and the third quartiles.

In symbols, Q.D = $\frac{Q_3 - Q_1}{2}$, Q.D is the abbreviation.

Among the quartiles Q_1, Q_2 and Q_3, the range is $Q_3 - Q_1$.

ie., inter-quartile range is $Q_3 - Q_1$ and Q.D which is $\frac{Q_3 - Q_1}{2}$ is the semi inter-quartile range.

Coefficient of Quartile Deviation = $\frac{\mathbf{Q_3 - Q_1}}{\mathbf{Q_3 + Q_1}}$

This is also called quartile coefficient of dispersion.

Example 3

Find the Quartile Deviation for the following:

391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution

Before finding Q.D., Q_1 and Q_3 are found from the values in ascending order:

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488

Position of Q_1 is $\frac{n+1}{4} = \frac{10+1}{4}$ = 2.75

Q_1 = 391 + 0.75 (407 – 391) = 403

Position of Q_3 is $\frac{3(n+1)}{4} = 3 \times 2.75$ = 8.25

Q_3 = 777+0.25 (1490 – 777) = 955.25

QD = $\frac{Q_3 - Q_1}{2} = \frac{955.25 - 403.00}{2}$ = **276.125**

Example 5

Calculate Quartile deviation for the following data. Also calculate quartile coefficient of dispersion.

Class:	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
f :	6	18	25	50	37	30	24	10

Solution

Classes	f	CF
20-30	6	6
30-40	18	24
40-50	25	49
50-60	50	99
60-70	37	136
70-80	30	166
80-90	24	190
90-100	10	200

$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m\right)}{f}c$ $\frac{N}{4} = \frac{200}{4} = 50$

$= 50 + \frac{(50-49)}{50}10$ $l_1 = 50, c = 10$

$= 50 + \frac{1 \times 10}{50} = 50 + \frac{1}{5}$ m = 49, f = 50

= 50 + 0.2 = **50.2**

$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m\right)}{f}c = 70 + \frac{(150-136)}{30}10$

$= 70 + \frac{14 \times 10}{30} = 70 + 4.67$ = **74.67**

$$QD = \frac{Q_3 - Q_1}{2} = \frac{74.67 - 50.20}{2} = \frac{24.47}{2} = \mathbf{12.23}$$

Quartile coefficient of dispersion

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{74.67 - 50.20}{74.67 - 50.20} = \frac{24.47}{124.87} = \mathbf{0.196}$$

Merits and Demerits

Merits

1. It is rigidly defined.
2. It is easy to understand and simple to calculate.
3. It is not unduly affected by extreme values.
4. It can be calculated for open-end distributions.

Demerits

1. It is not based on all observations
2. It is not capable of further algebraic treatment
3. It is much affected by fluctuations of sampling.

Mean Deviation

The Mean Deviation is defined as the Arithmetic mean of the absolute value of the deviations of observations from some origin, say mean or median or mode.

Thus for a raw data

$$\textbf{M.D about Mean} = \frac{\sum |\boldsymbol{x} - \bar{\boldsymbol{x}}|}{\boldsymbol{n}}$$

where $|x - \bar{x}|$ stands for the absolute deviation of x from $\bar{x}$ and is read as modulus of $(x - \bar{x})$ or mod $(x - \bar{x})$.

Instead of taking deviation from mean, if we are using median we get the mean deviation about median.

$$\therefore \text{ M.D. about Median} = \frac{\sum |\boldsymbol{x} - M|}{\boldsymbol{n}}$$

For a frequency data, MD about Mean is given by

$$(MD)\bar{x} = \frac{\sum f\,|x - \bar{x}|}{N};\; N = \sum f$$

$$\text{MD about Median (MD)} = \frac{\sum f\,|x - M|}{N}$$

Note

Whenever nothing is mentioned about the measure of Central tendency from which deviations are to be considered, deviations are to be taken from the mean and the required MD is MD about mean.

(i) Coefficient of MD (about mean) = $\dfrac{\textit{MD about mean}}{\textit{Mean}}$

(ii) Coefficient of MD (about median) = $\dfrac{\textit{MD about median}}{\textit{Median}}$

Example 6

Calculate MD about Mean of 8, 24, 12, 16, 10, 20

Solution

x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$
8	– 7	7
24	9	9
12	– 3	3
16	1	1
10	– 5	5
20	5	5
90		30

Example 8

Calculate MD about Mean and the coefficient of MD

Classes:	0-10	10-20	20-30	30-40	40-50
f :	5	15	17	11	2

Solution

Class	f	x	fx	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$
0-10	5	5	25	– 18	18	90
10-20	15	15	225	– 8	8	120
20-30	17	25	425	2	2	34
30-40	11	35	385	12	12	132
40-50	2	45	90	22	22	44
Total	50		1150			420

$$\bar{x} = \frac{\sum fx}{N} = \frac{1150}{50} = 23$$

$$(ND)_{\bar{x}} = \frac{\sum f\lvert x-\bar{x}\rvert}{N} = \frac{420}{50} = \mathbf{8.4}$$

$$\text{Coefficient of MD} = \frac{MD\ about\ mean}{Mean} = \frac{8.4}{23} = \mathbf{0.3652}$$

Merits and Demerits

Merits

1. It is rigidly defined
2. It is easy to calculate and simple to understand
3. It is based on all observations.
4. It is not much affected by the extreme values of items.
5. It is stable.

Demerits

1. It is mathematically illogical to ignore the algebraic signs of deviations.
2. No further algebraic manipulation is possible.
3. It gives more weight to large deviations than smaller ones.

Standard Deviation

The standard deviation is the most useful and the most popular measure of dispersion. The deviation of the observations from the AM are considered and then each squared. The sum of squares is divided by the number of observations. The square root of this value is known as the standard deviation. *Thus Standard deviation (SD) is defined as the square root of the AM of the squares of the deviations of observations from AM*. It is denoted by 'σ' (sigma). We can calculate SD using the following formula.

So for a raw data, if $x_1, x_2, x_3 \dots x_n$ are n observations

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

For a frequency data, if $x_1, x_2, x_3 \dots x_n$ are n observations or middle values of n classes with the corresponding frequencies $f_1, f_2, \dots f_n$

then, $$\text{SD} = \sigma = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}}$$

The square of the SD is known as 'Variance' and is denoted as s^2 or SD is the positive square root of variance.

Simplified formula for SD

For a raw data, we have

$$\pm^2 = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{1}{n}\sum\left(x^2 - 2x\,\bar{x} + \bar{x}^2\right)$$

$$= \frac{1}{n}\sum x^2 - 2\bar{x}\,\frac{1}{n}\sum x + \bar{x}^2\,\frac{1}{n}\sum 1$$

$$= \frac{\sum x^2}{n} - 2\bar{x}\cdot\bar{x} + \bar{x}^2$$

$$= \frac{\sum x^2}{n} - \bar{x}^2$$

$$\backslash \quad s = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

In a similar way, for a frequency data

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

Short Cut Method

For a raw data, $s = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ where d = x - A

For a frequency data, $s = c \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

where $d = \frac{x-A}{c}$, A - assumed mean, c - class interval.

The relative measure of dispersion based on SD or coefficient of SD is given by

Coefficient of SD = $\frac{\boldsymbol{SD}}{\boldsymbol{AM}} = \frac{\pm}{x}$

Importance of Standard Deviation

Standard deviation is always associated with the mean. It gives satisfactory information about the effectiveness of mean as a representative of the data. More is the value of the standard deviation less is the concentration of the observations about the mean and vice versa. Whenever the standard deviation is small mean is accepted as a good average.

According to the definition of standard deviation, it can never be negative. When all the observations are equal standad deviation is zero. Therefore a small value of s suggests that the observations are very close to each other and a big value of s suggests that the observations are widely different from each other.

Properties of Standard Deviation

1. Standard deviation is not affected by change of origin.

Proof

Let $x_1, x_2, \ldots, x_n$ be a set of n observations.

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

Choose $y_i = x_i + c$ for i = 1, 2, 3... n

Then $\bar{y} = \bar{x} + c$

$$\therefore \quad y_i - \bar{y} = x_i - \bar{x}$$

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

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

$$\text{ie., } \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

ie., $s_y = s_x$

Hence the proof

2. Standard deviation is affected by change of scale.

Proof

Let $x_1, x_2, \ldots, x_n$ be a set of n observations.

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

Choose $y_i = c\,x_i + d$, i = 1, 2, 3... n and c and d are constants. This fulfils the idea of changing the scale of the original values.

Now $$\bar{y} = c\,\bar{x} + d$$

$$\therefore \quad y_i - \bar{y} = c\,(x_i - \bar{x})$$

$$\sum (y_i - \bar{y})^2 = c^2 \sum (x_i - \bar{x})^2$$

$$\frac{1}{n}\sum (y_i - \bar{y})^2 = c^2\,\frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\text{ie., } \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2} = c\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

ie., $s_y = .\ ´\ s_x$

SD of y values = c ´ SD of x values

Hence the proof.

Note

If there are k groups then the S.D. of the k groups combined is given by the formula.

$$(n_1 + n_2 + \ldots + n_k)\pm^2 = n_1 \pm_1^2 + n_2 \pm_2^2 + \ldots + n_k \pm_k^2$$
$$+ \; + n_1 d_1^2 + n_2 d_2^2 + \ldots + n_k d_k^2$$

Coefficient of Variation

Coefficient of variation (CV) is the most important relative measure of dispersion and is defined by the formula.

$$\text{Coefficient of Variation} = \frac{\text{Standard deviation}}{\text{Arithmetic mean}} \times 100$$

$$\text{CV} = \frac{SD}{AM} \times 100 = \frac{\pm}{\bar{x}} \times 100$$

CV is thus the ratio of the SD to the mean, expressed as a percentage. According to Karl Pearson, Coefficient of variation is the percentage variation in the mean.

Coefficient of Variation is the widely used and most popular relative measure. The group which has less C.V is said to be more consistent or more uniform or more stable. More coefficient of variation indicates greater variability or less consistency or less uniformity or less stability.

Example 9

Calculate SD of 23, 25, 28, 31, 38, 40, 46

Solution

x	$x-\bar{x}$	$(x-\bar{x})^2$
23	– 10	100
25	– 8	64
28	– 5	25
31	– 2	4
38	5	25
40	7	49
46	13	169
231		436

$\bar{x}$ =231/7=33

SD = $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$ = **7.89**

Example 10

Calculate SD of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Solution

x	1	2	3	4	5	6	7	8	9	10
x^2	1	4	9	16	25	36	49	64	81	100

$\sum x$ = 55

$\sum x^2$ = 385

$$\bar{x} = \frac{\sum x}{n} = \frac{55}{10} = 5.5$$

$$\text{SD} = \pm = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{385}{10} - 5.5^2}$$

$$= \sqrt{38.5 - 30.25} = \sqrt{8.25} = \mathbf{2.87}$$

Calculate SD of 42, 48, 50, 62, 65

Solution

x	$d = x - 50$	d^2
42	– 8	64
48	– 2	4
50	0	0
62	12	144
65	15	225
Total	17	437

SD = $\sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ = **8.70**

Example 12

Calculate SD of the following data

Size (x) :	10	12	14	16	18
Frequency:	2	4	10	3	1

Solution

x	**f**	**fx**	**fx²**
10	2	20	200
12	4	48	576
14	10	140	1960
16	3	48	768
18	1	18	324
Total	20	274	3828

$$\pm = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{3828}{20} - \left(\frac{247}{20}\right)^2}$$

$$= \sqrt{191.4 - (13.7)^2} = \sqrt{191.40 - 187.69} = \sqrt{3.71} = \mathbf{1.92}$$

Example 13

Calculate SD of the following data

Classes :	0-4	4-8	8-12	12-16	16-20
f :	3	8	17	10	2

Solution

Classes	f	x	fx	fx^2
0-4	3	2	6	12
4-8	8	6	48	288
8-12	17	10	170	1700
12-16	10	14	140	1960
16-20	2	18	36	648
Total	40		400	4608

$$\pm = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{4608}{40} - \left(\frac{400}{40}\right)^2}$$

$$= \sqrt{115.2 - 100} = \sqrt{15.2} = \mathbf{3.89}$$

Example 14

Calculate mean, SD and CV for the following data

Classes :	0-6	6-12	12-18	18-24	24-30
f :	5	12	30	10	3

Solution

Class	f	x	$d = \frac{x-15}{6}$	fd	fd^2
0-6	5	3	– 2	– 10	20
6-12	12	9	– 1	– 12	12
12-18	15	15	0	0	0
18-24	10	21	1	10	10
24-30	3	27	2	6	12
Total	60			– 6	54

$$\bar{x} = A + \frac{\sum fd}{N} \times c = 15 + \frac{-6}{60} \times 6$$

$$= 15 - \frac{6}{10} = 15 - 0.6 = \mathbf{14.4}$$

$$\pm = c \times \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = 6 \times \sqrt{\frac{54}{60} - \left(\frac{-6}{60}\right)^2}$$

$$= 6 \times \sqrt{0.90 - 0.01} = 6 \times \sqrt{0.89}$$

$$= 6 \times 0.9434 = \mathbf{5.66}$$

$$CV = \frac{SD}{AM} \times 100 = \frac{5.66}{14.4} \times 100 = \mathbf{39.30\%}$$

Merits and Demerits

Merits

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.

7. If the mean deviation of a distribution is 20.20, the standard deviation of the distribution is:

a. 15.15 b. 25.25

c. 30.30 d. none of the above

8. The mean of a series is 10 and its coefficient of variation is 40 percent, the variance of the series is:

a. 4 b. 8 c. 12 d. none of the above

9. Which measure of dispersion can be calculated in case of open end intervals?

a. range b. standard deviation
c. coefficient of variation d. quartile deviation

Very Short Answer Questions

10. What are the uses of standard deviation?
11. Why measures of dispersion are called averages of second order?
12. For the numbers 3 and 5 show that SD = (1/2) Range.
13. Define CV and state its use.
14. State the desirable properties of a measure of dispersion
15. Define Quartile deviation.
16. Give the empirical relation connecting QD, MD and SD.

Short Essay questions

17. Define coefficient of variation. What is its relevance in economic studies?

18. What is a relative measure of dispersion? Distinguish between absolute and relative measure of dispersion.

19. Calculate coefficient of variation for the following distribution.

x	: 0	1	2	3	4	5	6
f	: 1	4	13	21	16	7	3

20. For the following data compute standard deviation,

x	: 10	20	30	40	50	60
f	: 3	5	7	20	8	7

21. Calculate median and quartile deviation for the following data

x	: 60	62	64	66	68	70	72
f	: 12	16	18	20	15	13	9

22. Calculate SD for the following data

Class interval:	0-5	5-10	10-15	15-20	20-25	25-30
Frequency :	4	8	14	6	3	I

Long Essay Questions

23. Compute coefficient of variation from the data given below.

Marks Less than:	10	20	30	40	50	60	70	80	90	100
No. of students :	5	13	25	48	65	80	92	97	99	100

24. Calculate the standard deviation of the following series.

More than :	0	10	20	30	40	50	60	70
Frequency :	100	90	75	50	25	15	5	0

25. The mean and the standard deviation of a group of 50 observations were calculated to be 70 and 10 respectively, It was later discovered that an observation 17 was wrongly-recorded as 70. Find the mean and the standard deviation (i) if the incorrect observation is omitted (ii) if the incorrect observation is replaced by the correct value.

11. Write short notes on

a. Method of least squares

b. Curve fitting

c. Normal equations.

Long essay questions

12. Fit a straight line by the method of least squares to the following data.

x:	0	1	2	3	4
y:	1	1.8	3.3	4.5	6.3

13. Fit a straight line y = *a* + bx to the following data.

x:	1	2	3	4	6	8
y:	2.4	3	3.6	4	5	6

14. Fit a straight line y = *a*x + b to the following data.

x:	1	2	3	4	5	6	7
y:	80	90	92	83	94	99	92

15. Fit a parabola $y = a + bx + x^2$ to the following data:

x:	0	1	2	3	4
y:	1	1.8	1.3	2.5	6.3

16. Fit a curve of the form $y = ax + bx^2$ for the data given below.

x:	1	2	3	4	5
y:	1.8	5.1	8.9	14.1	19.8

UNIT - III

CORRELATION AND REGRESSION

In the earlier chapters we have discussed the characteristics and shapes of distributions of a single variable, eg, mean, S.D. and skewness of the distributions of variables such as income, height, weight, etc. We shall now study two (or more) variables simultaneously and try to find the quantitative relationship between them. For example, the relationship between two variables like (1) income and expenditure (2) height and weight, (3) rainfall and yield of crops, (4) price and demand, etc. will be examined here. The methods of expressing the relationship between two variables are due mainly to Francis Galton and Karl Pearson.

Correlation

Correlation is a statistical measure for finding out degree (or strength) of association between two (or more) variables. By 'association' we mean the tendency of the variables to move together. Two variables X and Y are so related that movements (or variations) in one, say X, tend to be accompanied by the corresponding movements (or variations) in the other Y, then X and Y are said to be correlated. The movements may be in the same direction (i.e. either both X, Y increase or both of them decrease) or in the opposite directions (ie., one, say X, increases and the other Y decreases). Correlation is said to be positive or negative according as these movements are in the same or in the opposite directions. If Y is unaffected by any change in X, then X and Y are said to be uncorrelated.

In the words L.R. Conner:

If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other, then they are said to be correlated."

Correlation may be linear or non-linear. If the amount of variation in X bears a constant ratio to the corresponding amount of variation in Y, then correlation between X and Y is said to be linear. Otherwise it is non-linear. Correlation coefficient (r) measures the degree of linear relationship, (i.e., linear correlation) between two variables.

Determination of Correlation

Correlation between two variables may be determined by any one of the following methods:

1. Scatter Diagram
2. Co-variance Method or Karl Pearson's Method
3. Rank Method

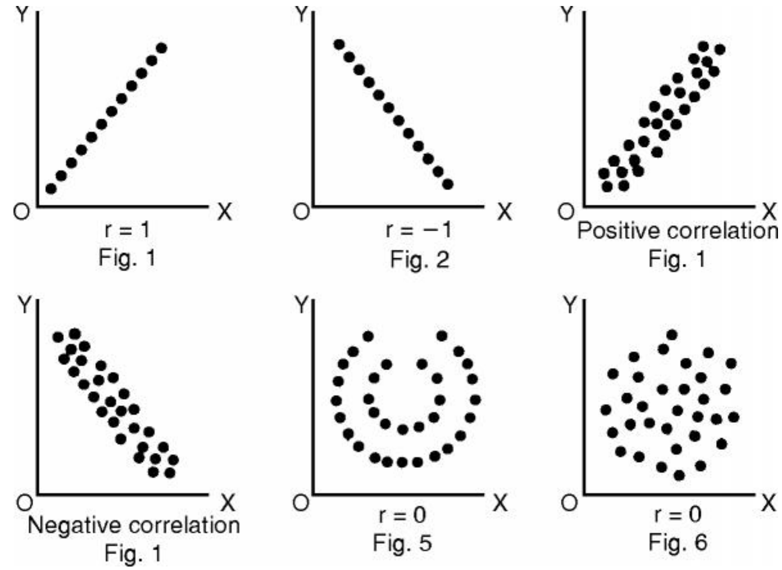
Scatter Diagram

The existence of correlation can be shown graphically by means of a *scatter diagram.* Statistical data relating to simultaneous movements (or variations) of two variables can be graphically represented-by points. One of the two variables, say X, is shown along the horizontal axis OX and the other variable Y along the vertical axis OY. All the pairs of values of X and Y are now shown by points (or dots) on the graph paper. This diagrammatic representation of bivariate data is known as scatter diagram.

The scatter diagram of these points and also the direction of the scatter reveals the nature and strength of correlation between the two variables. The following are some scatter diagrams showing different types of correlation between two variables.

In Fig. 1 and 3, the movements (or variations) of the two variables are in the same direction and the scatter diagram shows a linear path. In this case, correlation is positive or direct.

In Fig. 2 and 4, the movements of the two variables are in opposite directions and the scatter shows a linear path. In this case correlation is negative or indirect.



In Fig. 5 and 6 points (or dots) instead of showing any linear path lie around a curve or form a swarm. In this case correlation is very small and we can take r = 0.

In Fig. 1 and 2, all the points lie on a straight line. In these cases correlation is perfect and r = +1 or – 1 according as the correlation is positive or negative.

Karl Pearson's Correlation Coefficient

We have remarked in the earlier section that a scatter diagram gives us only a rough idea of how the two variables, say x and y, are related. We cannot draw defensible conclusions by merely examining data from the scatter diagram. In other words, we cannot simply look at a scatter diagram variables. On the other hand, neither can we conclude that the correlation at all. We need a quantity (represented by a number), which is a measure of the extent to which x and y are related. The quantity that is used for this purpose is known as the Co-efficient of Correlation, usually denoted by r_{xy} or r. The co-efficient of correlation r_{xy} measures the degree (or extent) of relationship between the two variables x and y and is given by the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n \pm_x \pm_y} \quad .... \ (1)$$

where X_t and Y_i $(i = 1, 2,.., n)$ are the two sets of values of x and y respectively and $\overline{x}$, $\overline{y}$, $\pm_x$, $\pm_y$ are respectively the corresponding means and standard deviations so that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\ , \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\pm_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\,\Sigma X_i^2 - \overline{X}^2$$

$$\text{and } \pm_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\,\Sigma Y_i^2 - \overline{Y}^2$$

The above definition of the correlation co-efficient was given by Karl Pearson in 1890 and is called *Karl Pearson's Correlation Co-efficient* after his name.

Definition

If (X_1, Y_1), (X_2, Y_2) .... (X_n, Y_n) be n pairs of observations on two variables X and Y, then the covariance of X and Y, written as cov (X,Y) is defined by

$$\text{Cov }(X,Y) = \frac{1}{n}\Sigma\,(X_i - \overline{X})(Y_i - \overline{Y})$$

Covariance indicates the joint variations between the two variables.

So the correlation coefficient or the coefficient of correlation (r) between X and Y is defined by

$$r = \frac{\mathbf{Cov\,(X, Y)}}{\pm_x \pm_y}$$

where $\pm_x$, $\pm_y$ are standard deviations of X and Y respectively.

The formula for the Correlation Coefficient r may be written in different forms.

i. If $\quad x_i = X - \overline{X} \text{ and } y_i = Y - \overline{Y}$

then $$r = \frac{\Sigma\, \mathbf{x_i\, y_i}}{n \pm_x \pm_y} \quad (1)$$

$$\therefore \text{ from (1), r } = \frac{\frac{1}{n}\,\Sigma x_i\, y_i}{\sqrt{\frac{\Sigma x_i^2}{n}} \times \sqrt{\frac{\Sigma y_i^2}{n}}} = \frac{\Sigma x_i\, y_i}{\sqrt{\Sigma x_i^2} \times \sqrt{\Sigma y_i^2}}$$

ii. We have

$$\begin{aligned}
\text{Cov (X, Y)} &= \frac{1}{n}\,\Sigma(X_i - \overline{X})\,(Y_i - \overline{Y}) \\
&= \frac{1}{n}\,\Sigma(X_iY_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\,\overline{Y}) \\
&= \frac{\Sigma X_iY_i}{n} - \overline{Y}\frac{\Sigma X_i}{n} - \overline{X}\frac{\Sigma Y_i}{n} + \frac{n\overline{X}\ \overline{Y}}{n} \\
&= \frac{\Sigma X_iY_i}{n} - \overline{X}\ \overline{Y} - \overline{X}\ \overline{Y} + \overline{X}\ \overline{Y} \\
&= \frac{\Sigma X_iY_i}{n} - \overline{X}\ \overline{Y} = \frac{\Sigma X_iY_i}{n} - \left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma Y_i}{n}\right)
\end{aligned}$$

and conclude that since more than half of the points appear to be nearly in a straight line, there is a positive or negative correlation between the

Now, $\quad r \quad = \quad \dfrac{Cov\,(X,\,Y)}{\pm_x\ \pm_y}$

$$= \frac{\dfrac{\Sigma X_i Y_i}{n} - \left(\dfrac{\Sigma X_i}{n}\right)\left(\dfrac{\Sigma Y_i}{n}\right)}{\sqrt{\dfrac{\Sigma X_i^2}{n} - \left(\dfrac{\Sigma X_i}{n}\right)^2} \times \sqrt{\dfrac{\Sigma Y_i^2}{n} - \left(\dfrac{\Sigma Y_i}{n}\right)^2}} \qquad \ldots(2)$$

iii. By multiplying each term of (2) by n^2, we have

$$\mathbf{r} = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{n\,\Sigma X_i^2 - (\Sigma X_i)^2} \times \sqrt{n\,\Sigma Y_i^2 - (\Sigma Y_i)^2}}$$

Theorem

The correlation coefficient is independent (not affected by) of the change of origin and scale of measurement.

Proof

Let $(x_1, y_1), (x_2, y_2)$ .... (x_n, y_n) be a set of n pairs of observations.

$$r_{xy} = \frac{\dfrac{1}{n}\Sigma(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\dfrac{1}{n}\Sigma(x_i - \bar{x})^2}\ \sqrt{\dfrac{1}{n}\Sigma(y_i - \bar{y})^2}} \qquad \ldots(1)$$

Let us transform x_i to u_i and y_i to v_i by the rules,

$$u_i = \frac{x_i - x_0}{c_1} \ \ and\ \ v_i = \frac{y_i - y_0}{c_2} \qquad \ldots(2)$$

where x_0, y_0, c_1, c_2 are arbitrary constants.

From (2), we have

$$x_i = c_1 u_i + x_0 \quad \text{and} \quad y_i = c_2 v_i + y_0$$

$$\bar{x} = x_0 + c_1\,\bar{u} \quad \text{and} \quad \bar{y} = y_0 + c_2\,\bar{v}$$

where $\bar{u}$ and $\bar{v}$ are the means u_i^s and v_i^s respectively.

$$x_i - \bar{x} = c_1(u_i - \bar{u}) \quad \text{and} \quad y_i - \bar{y} = c_2(v_i - \bar{v})$$

Substituting these values in (1), we get

$$r_{xy} = \frac{\dfrac{1}{n}\sum_{i=1}^{n} c_1(u_i - \bar{u})c_2(v_i - \bar{v})}{\sqrt{\dfrac{1}{n}\sum_{i=1}^{n} c_1^2\,(u_i - \bar{u})^2}\ \sqrt{\dfrac{1}{n}\sum_{i=1}^{n} c_2^2\,(v_i - \bar{v})^2}}$$

$$= \frac{\dfrac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2}\ \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

$$= \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{n\,\pm_u \pm_v} = r_{uv}$$

Here, we observe that if we change the origin and choose a new scale, the correlation co-efficient remains unchanged. Hence the proof.

Here, r_{uv} can be further simplified as

$$r_{xy} = \frac{Cov(u,\,v)}{\pm_u\ \pm_v}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} u_i v_i - \bar{u}\,\bar{v}}{\sqrt{\frac{1}{n}\Sigma u_i^2 - \bar{u}^2}\sqrt{\frac{1}{n}\Sigma v_i^2 - \bar{v}^2}}$$

$$= \frac{n\,\Sigma u_i v_i - \Sigma u_i\,\Sigma v_i}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i)^2}\sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

Limits of Correlation Co-efficient

We shall now find the limits of the correlation coefficient between two variables and show that it lies between -1 and $+1$.

ie., $-1 \le r_{xy} < +1$

Proof

Let (x_1, y_1), (x_2, y_2) .... (x_n, y_n) be the given pairs of observations.

Then $$r_{xy} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}}$$

We put

$$X_i = x_i - \bar{x}, \qquad Y_i = y_i - \bar{y}$$

$$\pm_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \quad \text{. . . (1)}$$

Similarly $$\pm_y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \quad \text{...(2)}$$

and $$r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i}{n \pm_x \pm_y} \quad \text{...(3)}$$

Now we have

$$\sum_{i=1}^{n}\left(\frac{X_i}{\pm_x} \pm \frac{Y_i}{\pm_y}\right)^2 = \frac{\sum_{i=1}^{n} X_i^2}{\pm_x^2} + \frac{\sum_{i=1}^{n} Y_i^2}{\pm_y^2} + \frac{2\sum_{i=1}^{n} X_i Y_i}{\pm_x \pm_y}$$

$$= \frac{n \pm_x^2}{\pm_x^2} + \frac{n \pm_y^2}{\pm_y^2} \pm 2nr_{xy} \quad \text{using (1), (2), (3).}$$

$$= 2n \pm 2n\,r_{xy} = 2n(1 \pm r_{xy})$$

Left hand side of the above identity is the sum of the squares of n numbers and hence it is positive or zero.

Hence, $1 \pm r_{xy} \ge 0$ or, $r_{xy} \le 1$ and $r_{xy} \ge -1$

or $-1 \le r_{xy} \le +1$

ie., the correlation co-efficient lies between – 1 and + 1. Hence the proof.

Note:

If r_{xy} = 1, we say that there is perfect positive correlation between x and y.

If r_{xy} = – 1, we say that there is perfect negative correlation between x and y.

If r_{xy} = 0, we say that there is no correlation between the two variables, i.e., the two variables are uncorrelated.

If r_{xy} > 0, we say that the correlation between x and y is positive (direct).

If r_{xy} < 0, we say that the correlation between x and y is negative (indirect).

X	Y	x = X – $\overline{X}$	y = Y – $\overline{Y}$	x^2	y^2	xy
1	6	– 3	– 4	9	16	12
2	8	– 2	– 2	4	4	4
3	11	– 1	1	1	1	– 1
4	9	0	– 1	0	1	0
5	12	1	2	1	4	2
6	10	2	0	4	0	0
7	14	3	4	9	16	12
28	70			28	42	29

$$\overline{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4 \quad \text{and} \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{70}{7} = 10$$

Karl Pearson's coefficient of correlation (r) is given by

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

$$= \frac{29}{\sqrt{28}\sqrt{42}} = \mathbf{0.8457}$$

Example 9

Karl Pearson's coefficient of correlation between two variables X and Y is 0.28 their covariance is +7.6. If the variance of X is 9, find the standard deviation of Y-series.

Solution

Karl Pearson's coefficient of correlation r is given by

$$r = \frac{\text{cov}(X, Y)}{\pm_x \pm_y}$$

Here r =0.28, Cov (X, Y) = 7.6 and $\pm_x^2 = 9$; $\pm_x = 3$.

Using (1) $\quad 0.28 = \dfrac{7.6}{3\pm_y}$

or, $0.84\pm_y = 7.6, \quad or \quad \pm_y = \dfrac{7.6}{0.84} = \dfrac{760}{84}$

= **9.048**

Example 10

Calculate Pearson's coefficient of correlation between advertisement cost and sales as per the data given below:

Advt cost in '000 Rs:	39	65	62	90	82	75	25	98	36	78
Sales in lakh Rs:	47	53	58	86	62	68	60	91	51	84

Solution

Karl Pearson's coefficient of correlation (r) is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}} \quad \text{where } x = X - \overline{X} \quad \text{and} \quad y = Y - \overline{Y}$$

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
39	47	– 26	– 19	676	361	494
65	53	0	– 13	0	169	0
62	58	– 3	– 8	9	64	24
90	86	25	20	625	400	500
82	62	17	– 4	289	16	– 68
75	68	10	2	100	4	20
25	60	– 40	– 6	1600	36	240
98	91	33	25	1089	625	825

Example 8

Find the coefficient of correlation from the following data:

X :	1	2	3	4	5	6	7
Y :	6	8	11	9	12	10	14

36	51	– 29	– 15	841	225	435
78	84	13	18	169	324	234
650	660	0	0	5398	2224	2704

$$\overline{X} = \frac{\Sigma X}{n} = \frac{650}{10} = 65; \quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{660}{10} = 66$$

$$r = \frac{2704}{\sqrt{5398} \times \sqrt{2224}} = \mathbf{0.78}$$

Example 11

Calculate Pearson's coefficient of correlation from the following taking 100 and 50 as the assumed average of X and Y respectively:

X:	104	111	104	114	118	117	105	108	106	100	104	105
Y:	57	55	47	45	45	50	64	63	66	62	69	61

Solution

X	Y	u = X – 100	v = Y – 50	u^2	v^2	uv
104	57	4	7	16	49	28
111	55	11	5	121	25	55
104	47	4	– 3	16	9	– 12
114	45	14	– 5	196	25	– 70
118	45	18	– 5	324	25	– 90
117	50	17	0	289	0	0
105	64	5	14	25	196	70

$$\Sigma XY = 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8)$$

108	63	8	13	64	169	104
106	66	6	16	36	256	96
100	62	0	12	0	144	0
104	69	4	19	16	361	76
105	61	5	11	25	121	55
–	–	96	84	1128	1380	312

$$r = \frac{n\,\Sigma u_i v_i - \Sigma u_i\,\Sigma v_i}{\sqrt{n\,\Sigma u_i^2 - (\Sigma u_i)^2}\,\sqrt{n\,\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{12 \times 312 - 96 \times 84}{\sqrt{12 \times 1128 - (96)^2}\,\sqrt{12 \times 1380 - (84)^2}}$$

$$= \mathbf{-0.67}$$

Example 12

A computer while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results: $n = 25, \Sigma X = 125, \Sigma Y = 100, \Sigma X^2 = 650, \Sigma Y^2 = 460$ and $\Sigma XY = 508$. It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6), while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be 2/3.

Solution

When the two incorrect pairs of observations are replaced by the correctpairs, the revised results for the whole series are:

ΣX = 125 – (Sum of two incorrect values of X) + (Sum of two correct values of X)

= 125 – (6 + 8) + (8 + 6) = 125

Similarly

ΣY = 100 – (14 + 6) + (12 + 8) = 100

ΣX^2 = 650 – $(6^2 + 8^2) + (8^2 + 6^2)$ = 650

ΣY^2 = 460 – $(14^2 + 6^2) + (12^2 + 8^2)$

= 460 – 232 + 208 = 436 and

$$= \ 508 - 132 + 144 = 520 \ ;$$

Correct value of the correlation co efficient is

$$r = \frac{n\,\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\,\Sigma X^2 - (\Sigma X^2)}\ \sqrt{n\,\Sigma Y^2 - (\Sigma Y^2)}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - 125^2}\ \sqrt{25 \times 436 - 100^2}}$$

= **2/3**

Rank Correlation Coefficient

Simple correlation coefficient (or product-moment correlation coefficient) is based on the magnitudes of the variables. But in many situations it is not possible to find the magnitude of the variable at all. For example, we cannot measure beauty or intelligence quantitatively. In this case, it is possible to rank the individuals in some order. Rank correlation is based on the rank or the order and not on the magnitude of the variable. It is more suitable if the individuals (or variables) can be arranged in order of merit or proficiency. If the ranks assigned to individuals range from 1 to n, then the Karl Pearson's correlation coefficient between two series of ranks is called Rank correlation coefficient. Edward Spearman's formula for Rank correlation coefficient (R) is given by.

$$R = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)} \ \ or \ \ 1 - \frac{6\,\Sigma d^2}{(n^3 - n)}$$

where d is the difference between the ranks of the two series and n is the number of individuals in each series.

Derivation of Spearman's Formula for Rank Correlation Coefficient

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

Proof:

Let $(x_1, y_1), (x_2, y_2), \ldots. (x_n, y_n)$ be the ranks of n individuals in two characters (or series) Edward Spearman's Rank correlation coefficient R is the product-moment correlation coefficient between these ranks and, therefore, we can write.

$$R = \frac{Cov(x, y)}{\pm_x \pm_y} \qquad \ldots(1)$$

where cov (x, y) $= \dfrac{\Sigma\{(x_i - \bar{x})(y_i - \bar{y})\}}{n}$

But the ranks of n individuals are the natural numbers 1, 2,.... n arranged in some order depending on the qualities of the individuals.

$\therefore$ $x_1, x_2, \ldots, x_n$ are the numbers 1, 2... n in some order.

$$\therefore \Sigma x = 1 + 2 + \ldots. + n = \frac{n(n+1)}{2} \text{ and}$$

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

$$\bar{x} = \frac{\Sigma x}{n} = \frac{n+1}{2}, \ \frac{\Sigma x^2}{n} = \frac{(n+1)(2n+1)}{6}$$

$$\therefore \pm_x^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

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

similarly,

$$\bar{y} = \frac{n+1}{2} \ \text{ and } \ \pm_y^2 = \frac{n^2 - 1}{12}$$

Let $d_i = x_i - y_i$; then $d_i = (x_i - \bar{x}) - (y_i - \bar{y})$ $[\because \bar{x} = \bar{y}]$

Calculate the rank correlation coefficient.

$$\therefore \frac{\Sigma d_i^2}{n} = \frac{\Sigma\{(x_i-\bar{x})-(y_i-\bar{y})\}^2}{n}$$

$$= \frac{\Sigma(x_i-\bar{x})^2}{n} + \frac{\Sigma(y_i-\bar{y})^2}{n} - \frac{2\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$= \pm_X^2 + \pm_Y^2 - 2\,\text{cov}\,(x,y)$$

or, 2cov (x, y) =

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

or, cov (x, y) $= \frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}$

Hence, from (1), we get

$$R = \left(\frac{n^2-1}{12} - \frac{\Sigma d_i^2}{2n}\right) \bigg/ \left(\frac{n^2-1}{12}\right)$$

$$= \mathbf{1} \text{ J } \frac{\mathbf{6}\ \text{p}\ \boldsymbol{d}^{\mathbf{2}}}{\boldsymbol{n}(\boldsymbol{n}^{\mathbf{2}} \text{ J } \mathbf{1})} \quad \text{[omitting i]}$$

Example 13

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Maths.	78	36	98	25	75	82	90	62	65	69
Marks in Stat.	84	51	91	60	68	62	86	58	53	47

Solution

In Mathematics, Student with Roll No. 3 gets the highest mark 98 and is ranked 1; Roll No. 7 securing 90 marks has rank 2 and so on. Similarly, we can find the ranks of students in statistics.

Roll No.	Mathematics Marks	Rank (x)	Statistics Marks	Rank (y)	Rank Diff. d = x – y	2
1	78	4	84	3	1	1
2	36	9	51	9	0	0
3	98	1	91	1	0	0
4	25	10	60	6	4	16
5	75	5	68	4	1	1
6	82	3	62	5	2	4
7	90	2	86	2	0	0
8	62	7	58	7	0	0
9	65	6	53	8	– 2	4
10	39	8	47	10	– 2	4
Total	–	–	–	–	–	$30 = \Sigma d^2$

Applying Edward Spearman's formula:

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

$$= 1 - \frac{6 \times 30}{10(10^2-1)} = 1 - \frac{18}{99}$$

$$= 1 - \frac{2}{11} = \frac{9}{11} = \mathbf{0.82}$$

Regression

In some situations, one may need to know the probable value of one variable corresponding to certain value of another variable. This is possible using the mathematical relation between the two variables. Scatter diagram, explained above helps to ascertain the nature of relationship such as linear (straight line), second degree polynomial (parabola), etc. Discussion in this book is restricted to linear relation between two variables.

During study of hereditary characteristics, Sir Francis Galton found that the heights of different groups of sons had the tendency to *regress*, that is to *go back* towards the overall average height of all groups of fathers. He called the lines of the average relationship as the lines of the regression. It is also referred to as the estimating equations because based on the value of one variable one can predict or estimate the value of the other variable.

Suppose we are given n pairs of values (x_1, y_1) (x_2, y_2), ... (x_n, y_n) of two variables x and y. If we fit a straight line to this data by taking x as independent variable and y as dependent variable, then the straight line obtained is called the *regression line of y on x*. Its slope is called the *regression coefficient of y on x*. Similarly, if we fit a straight line to the data by taking y as independent variable and x as dependent variable, the line obtained is the *regression line of x on y;* the reciprocal of its slope is called the *regression coefficient of x on y.*

Equation for regression lines

Let $$y = a + bx \quad .... (1)$$

be the equation of the regression line of y on x, where *a* and b are determined by solving the normal equations obtained by the principle of least squares.

$$\Sigma y_i = na + b\,\Sigma x_i \quad .... (2)$$

$$\Sigma x_i y_i = a\,\Sigma x_i + b\,\Sigma x_i^2 \quad .... (3)$$

Divide the equation (2) by n, we get

$$\frac{1}{n}\Sigma y_i = a + \frac{1}{n} b \Sigma x_i$$

or $$\bar{y} = a + b\,\bar{x} \quad .... (4)$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y series. Substituting for *a* from (4) in (1), we get the equation,

$$y - \bar{y} = b(x - \bar{x}) \quad .... (5)$$

Solving the equations (2) and (3) for b after eliminating '*a*' we get the value of b as

$$b = \frac{n\,\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\,\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\frac{\Sigma x_i y_i}{n} - \bar{x}\,\bar{y}}{\frac{\Sigma x_i^2}{n} - \bar{x}^2}, \text{ dividing each term by } n^2$$

$$= \frac{Cov(x, y)}{\pm_x^2} = \frac{P_{xy}}{\pm_x^2}$$

Substituting b in (5), we get the regression equation of y on x as

$$\boldsymbol{y - \bar{y} = \frac{P_{xy}}{\pm_x^2}(x \text{ J } \bar{x})} \quad .... (6)$$

Similarly, when x is depending on y, the regression equation of x on y is obtained as

$$x - \bar{x} = \frac{P_{xy}}{\pm_y^2}(y - \bar{y}) \quad .... (7)$$

Let us denote $\frac{P_{xy}}{\pm_x^2}$ as b_{yx} and $\frac{P_{xy}}{\pm_x^2}$ as b_{xy}

Thus $b_{yx} = \frac{P_{xy}}{\pm_x^2}$ as $b_{xy} = \frac{P_{xy}}{\pm_y^2}$

Here b_{yx} is called the regression coefficient of y on x and b_{xy} is called the regression coefficient of x on y.

So we can rewrite the regression equation of y on x as

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

and the regression equation of x on y as

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

Some remarks

1. The slope of the regression line of y on x is $b_{xy} = \dfrac{r \pm_y}{\pm_x}$ and the slope of the regression line of x on y is the reciprocal of b_{xy} which is $\dfrac{\pm_y}{r \pm_x}$.

2. Since $b_{yx} = r(\pm_y / \pm_x)$ and $\pm_x$ and $\pm_y$ are positive, it follows that r has the same sign as that of b_{yx}.

3. Since $b_{xy} = r(\pm_x / \pm_y)$ we readly find that $(b_{yx})\,(b_{xy}) = r^2$. Since $r^2 > 0$. It follows that b_{xy} has the same sign as that of b_{yx}. Thus, r, b_{xy} and b_{yx} always have the same signs. Also $|r| = \sqrt{(b_{yx})(b_{xy})}$. That is, |r| is the geometric mean of b_{xy} and b_{yx}. Since |r| < 1 it follows that $b_{yx} > 1$ whenever $b_{xy} < 1$ and vice-versa.

4. Since the arithmetic mean is always greater than the geometric mean for any two numbers, we have $\frac{1}{2}(b_{yx} + b_{xy}) > \sqrt{b_{yx} \times b_{xy}} = |r|$.

   Thus, the arithmetic mean of b_{xy} and b_{yx} is always greater than the coefficient of correlation.

5. The two lines of regression always pass through the point $(\bar{x}, \bar{y})$.

6. The regression equation of y on x is need for estimating or predicting the value of y for a given value of x and the regression equation of x on y is used for estimating or predicting x for a specified value of y.

SOLVED PROBLEMS

Example 21

Calculate the coefficient of correlation for the following ages of husbands and wives.

Age of husband (x):	23	27	28	29	30	31	33	35	36	39
Age of wife (y):	18	22	23	24	25	26	28	29	30	32

Solution

We have, $\bar{x} = \dfrac{1}{n}\Sigma x_i = \dfrac{311}{10} = 31.1$

$$\bar{y} = \frac{1}{n}\Sigma y_i = \frac{257}{10} = 25.7$$

We prepare the following table.

X_i	$x_i = X_i - \bar{x}$	x_i^2	Y_i	$y_i = Y_i - \bar{y}$	y_i^2	$x_i y_i$
23	– 8.1	65.61	18	– 7.7	52.29	62.37
27	– 4.1	16.81	22	– 3.7	13.69	15.17
28	– 3.1	9.61	23	– 2.7	7.29	8.37
29	– 2.1	4.41	24	– 1.7	2.89	3.57
30	– 1.1	1.21	25	– 0.7	0.49	0.77
31	– 0.1	0.01	26	0.3	0.09	– 0.03
33	1.9	3.61	28	2.3	5.29	4.37
35	3.9	15.21	29	3.3	10.89	12.87
36	4.9	24.01	30	4.3	18.49	12.07
39	7.9	62.41	32	6.3	39.69	49.77
		202.90			158.10	178.30

Now, $r = \dfrac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}} = \dfrac{178.30}{\sqrt{202.90} \times \sqrt{158.10}}$ = **0.9955**

Example 22

Calculate the coefficient of correlation for the following data.

x:	6	2	10	4	8
y:	9	11	5	8	7

Solution

Here we prepare the following table

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

$$r = \frac{n\,\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\,\Sigma X^2 - (\Sigma X)^2}\,\sqrt{n\,\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5 \times 214 - 30 \times 40}{\sqrt{5 \times 220 - 30^2}\,\sqrt{5 \times 340 - 40^2}}$$

$$= \frac{-130}{\sqrt{200}\,\sqrt{100}} = \mathbf{0.919}$$

Example 23

Find the correlation coefficient between X and Y given

x:	10	16	13	12	15	17	14
y:	20	33	25	27	26	30	30

Solution

Here we prepare the following table

X	Y	$u_i = X - 14$	$v_i = Y - 25$	u_i^2	v_i^2	$u_i\,v_i$
10	20	– 4	– 5	16	25	20
16	33	2	8	4	64	16
13	25	– 1	0	1	0	0
12	27	– 2	2	4	4	– 4
15	26	1	1	1	1	1
17	30	3	5	9	25	15
14	30	0	5	0	25	0
		– 1	16	35	144	48

$$r_{xy} = r_{uv} = \frac{n\,\Sigma u_i\,v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\,\Sigma u_i^2 - (\Sigma u_i)^2}\,\sqrt{n\,\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{7 \times 48 - (-1) \times 16}{\sqrt{7 \times 35 - (-1)^2}\,\sqrt{7 \times 144 - 16^2}}$$

$$= \frac{336 + 16}{\sqrt{245 - 1}\,\sqrt{1008 - 256}}$$

$$= \frac{352}{\sqrt{244}\,\sqrt{752}} = \mathbf{0.82}$$

Example 24

Calculate the rank correlation coefficient from the following data specifying the ranks of 7 students in two subjects.

Rank in the first subject :	1	2	3	4	5	6	7
Rank in the second subject :	4	3	1	2	6	5	7

Solution

Here n = 7. Let x and y denote respectively the ranks in the first and second subjects. We prepare the following table.

x_i	y_i	$d_i = x_i - y_i$	d_i^2
1	4	– 3	9
2	3	– 1	1
3	1	2	4
4	2	2	4
5	6	– 1	1
6	5	1	1
7	7	0	0
			20

The Spearman's rank correlation coefficient is

$$R = 1 - \frac{6\,\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 20}{7 \times (7^2-1)} = 0.643$$

Example 25

Find the rank correlation coefficient between marks in two subjects A and B scored by 10 students

A:	88	72	95	60	35	46	52	58	30	67
B:	65	90	86	72	30	54	38	43	48	75

Solution

The following table is prepared.

A	B	Ranks in A	Ranks in B	d_i	d_i^2
88	65	2	5	– 3	9
72	90	3	1	2	4
95	86	1	2	– 1	1
60	72	5	4	1	1
35	30	9	10	– 1	1
46	54	8	6	2	4
52	38	7	9	– 2	4
58	43	6	8	– 2	4
30	48	10	7	3	9
67	75	4	3	1	1

$$R = 1 - \frac{6\,\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 38}{10 \times (10^2-1)}$$

$$= 1 - 0.2303 = \mathbf{0.7697}$$

Example 26

The coefficient of rank correlation of marks obtained by 10 students in two subjects was computed as 0.5. It was later discovered that the difference in marks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution

Here given R = 0.5, and n = 10.

Then we have,

0.5 =

or = 82.5

Deleting the wrong item from this and adding the correct item to it we obtain corrected

= 82.5 3^2 + 72 = 122.5.

Consequently, the correct coefficient of rank correlation is

R = = **0.2576**

Example 27

The following are the data on the average height of the plants and weight of yield per plot recorded from 10 plots of rice crop.

Height (X) (cms)	:	28	26	32	31	37	29	36	34	39	40
Yield (Y) (kg)	:	75	74	82	81	90	80	88	85	92	95

Find (i) correlation coefficient between X and Y (ii) the regression coefficient and hence write down regression equation of y on x and that of x on y (iii) probable value of the yield of a plot having an average plant height of 98 cms.

Solution

Here we prepare the following table.

X	Y	$u_i = X - 34$	$v_i = Y - 80$	u_i^2	v_i^2	$u_i v_i$
28	75	– 6	– 5	36	25	30
26	74	– 8	– 6	64	36	48
32	82	– 2	2	4	4	– 4
31	81	– 3	1	9	1	– 3
37	90	4	10	16	100	40
29	80	– 5	0	25	0	0
36	88	2	8	4	64	16
34	85	0	5	0	25	0
39	92	5	12	25	144	60
40	95	6	15	36	225	90
		– 7	42	219	624	277

i. $$r_{xy} = r_{uv} = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}}$$

$$= \frac{10 \times 277 - (-7) \times 42}{\sqrt{10 \times 219 - (-7)^2}\sqrt{10 \times 624 - (42)^2}}$$

$$= \frac{3064}{46.271 \times 66.903} = \mathbf{0.989}$$

ii. The regression coefficient of y on x is

$$b_{yx} = \frac{n\sum u_i v_i - (\sum u_i \sum v_i)}{n\sum u_i^2 - (\sum u_i)^2}$$

$$= \frac{3064}{2140.99} = 1.431$$

The regression coefficient of x on y is

$$b_{xy} = \frac{n\sum u_i v_i - (\sum u_i \sum v_i)}{n\sum v_i^2 - (\sum v_i)^2}$$

$$= \frac{3064}{4476.01} = 0.684$$

The regression equation of y on x is $\qquad \bar{x} = A + \frac{\sum u_i}{n}$

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad\qquad = 34 + \frac{-7}{10} = 33.3$$

ie., $\quad y - 84.2 = 1.431(x - 33.3) \qquad \bar{y} = B + \frac{\sum v_i}{n}$

ie., $\quad$ **y = 1.431x – 36.55** $\qquad = 80 + \frac{42}{10} = 84.2$

The regression equation of x on y is

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

i e . , $\quad x - 33.3 = 0.684(y - 84.2)$

ie., $\quad$ **x = 0.684y – 24.29**

iii.To estimate the yield (y), the regression equation of y on x is

$$y = 1.431x - 36.55$$

when x = 98, y = $1.431 \times 98 - 36.55$ = **103.69kg**

Example 27

For the regression lines $4x - 5y + 33 = 0$ and $20x - 9y = 107$, find (a) the mean values of x and y, (b) the coefficient of correlation between x and y, and (c) the variance of y given that the variance of x is 9.

Solution

Since the lines of regression pass through $(\bar{x}, \bar{y})$ we have

$$4\bar{x} - 5\bar{y} + 33 = 0$$

$$20\bar{x} - 9\bar{y} - 107 = 0$$

Solving these equations, we get the mean values of x and y as $\bar{x} = 13$, $\bar{y} = 17$. We rewrite the given equations respectively as

$$y = \frac{4}{5}x + \frac{33}{5}, \; x = \frac{9}{20}y + \frac{107}{20} \text{ so that } b_{yx} = \frac{4}{5}, \; b_{xy} = \frac{9}{20}$$

Therefore, the coefficient of correlation between x and y is

$$r = \sqrt{(b_{xy})(b_{yx})} = \mathbf{0.6}$$

Here positive sign is taken since both b_{xy} and b_{yx} are positive.

Since $r\frac{\pm_y}{\pm_x} = b_{yx} = \frac{4}{5}$, *and* $\pm_x^2 = 9$ (given), we get

$$\pm_y = \frac{4\pm_x}{5r} = \frac{4 \times 3}{5 \times 0.6} = 4$$

Thus, the variance of y is $\pm_y^2$ = **16**.