LESSON 1
NATURE, SCOPE
AND LIMITATIONS OF STATISTICS

Introduction

The term "statistics" is used in two senses : first in plural sense meaning a collection of numerical facts or estimates—the figure themselves. It is in this sense that the public usually think of statistics, e.g., figures relating to population, profits of different units in an industry etc. Secondly, as a singular noun, the term'statistics' denotes the various methods adopted for the collection, analysis and interpretation of the factsnumerically represented. In singular sense, the term 'statistics' is better described as statistical methods. In our study of the subject, we shall be more concerned with the second meaning of the word 'statistics'.

Definition

Statistics has been defined differently by different authors and each author has assigned new limits to the field which should be included in its scope. We can do no better than give selected definitions of statistics by some authors and then come to the conclusion about the scope of the subject. A.L. Bowley defines, **"Statistics may be called the science of counting".** At another place he defines, **"Statistics may be called the science of averages".** Both these definitions are narrow and throw light only on one aspect of Statistics. According to King, **"The science of statistics is the method of judging collective, natural or social, phenomenon from the results obtained from the analysis or enumeration or collection of estimates".**

Many a time counting is not possible and estimates are required to be made. Therefore, Boddington defines it as "the science of estimates and probabilities". But this definition also does not cover the entire scope of statistics. The statistical methods are methods for the collection, analysis and interpretation of numerical data and form a basis for the analysis and comparison of the observed phenomena. In the words of Croxton &Cowden, "Statistics may be defined as the collection, presentation, analysis and interpretation of numericaldata". Horace Secrist has given an exhaustive definition of the term satistics in the plural sense. According to him:

"By statistics we mean aggregates of facts affected to a marked extent by a multiplicity of causesnumerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a pre-determined purpose and placed in relation to each other".

This definition makes it quite clear that as numerical statement of facts, 'statistic' should possess thefollowing characteristics.

1. **Statistics are aggregate of facts**

A single age of 20 or 30 years is not statistics, a series of ages are. Similarly, a single figure relating to production, sales, birth, death etc., would not be statistics although aggregates of such figures would bestatistics because of their comparability and relationship.

2. Statistics are affected to a marked extent by a multiplicity of causes
A number of causes affect statistics in a particular field of enquiry, e.g., in production statistics are affected by climate, soil, fertility, availability of raw materials and methods of quick transport.

3. Statistics are numerically expressed, enumrated or estimated
The subject of statistics is concerned essentially with facts expressed in numerical form—with theirquantitative details but not qualitative descriptions. Therefore, facts indicated by terms such as 'good', 'poor' are not statistics unless a numerical equivalent, is assigned to each expression. Also this may either beenumerated or estimated, where actual enumeration is either not possible or is very difficult.
4. Statistics are numerated or estimated according to reasonable standard of accuracy
Personal bias and prejudices of the enumeration should not enter into the counting or estimation of figures, otherwise conclusions from the figures would not be accurate. The figures should be counted or estimated according to reasonable standards of accuracy. Absolute accuracy is neither necessary nor sometimes possible in social sciences. But whatever standard of accuracy is once adopted, should be used throughout the process of collection or estimation.
5. Statistics should be collected in a systematic manner for a predetermined purpose
The statistical methods to be applied on the purpose of enquiry since figures are always collected with some purpose. If there is no predetermined purpose, all the efforts in collecting the figures may prove to be wasteful. The purpose of a series of ages of husbands and wives may be to find whether young husbands have young wives and the old husbands have old wives.
6. Statistics should be capable of being placed in relation to each other
The collected figure should be comparable and well-connected in the same department of inquiry. Ages of husbands are to be compared only with the corresponding ages of wives, and not with, say, heights of trees.
Functions of Statistics

The functions of statistics may be enumerated as follows :
(i) To present facts in a definite form : Without a statistical study our ideas are likely to be vague, indefinite and hazy, but figures helps as to represent things in

their true perspective. For example, the statement that some students out of 1,400 who had appeared, for a certain examination, were declared successful would not give as much information as the one that 300 students out of 400 who took the examination were declared successful.

(ii) To simplify unwieldy and complex data : It is not easy to treat large numbers and hence theyare simplified either by taking a few figures to serve as a representative sample or by taking average to give a bird's eye view of the large masses. For example, complex data may be simplified by presenting them in the form of a table, graph or diagram, or representing it through an average etc.

(iii) To use it as a technique for making comparisons : The significance of certain figures can be better appreciated when they are compared with others of the same type. The comparison between two different groups is best represented by certain statistical methods, such as average, coefficients, rates, ratios, etc.

(iv) To enlarge individual experience : An individual's knowledge is limited to what he can observe and see; and that is a very small part of the social organism. His knowledge is extended n various ways by studying certain conclusions and results, the basis of which are numerical investigations. For example, we all have general impression that the cost of living has increased. But to know to what extent the increase has occurred, and how far the rise in prices has affected different income groups, it would be necessary to ascertain the rise in prices of articles consumed by them.

(v) To provide guidance in the formulation of policies : The purpose of statistics is to enable correct decisions, whether they are taken by a businessman or Government. In fact statistics is a great servant of business in management, governance and development. Sampling methods are employed in industry in tacking the problem of standardisation of products. Big business houses maintain a separate department for statistical intelligence, the work of which is to collect, compare and coordinate figures for formulating future policies of the firm regarding production and sales.

(vi) To enable measurement of the magnitude of a phenomenon : But for the development of the statistical science, it would not be possible to estimate the population of a country or to know the quantity of wheat, rice and other agricultural commodities produced in the country during any year.

Importance of Statistics

These days statistical methods are applicable everywhere. There is no field of work in which statistical methods are not applied. According to A L. Bowley, 'A knowledge of statistics is like a knowledge of foreign languages or of Algebra, it may prove of use at any time under any circumstances". The importance of the statistical science is increasing in almost all spheres of knowledge, e g., astronomy, biology, meteorology, demography, economics and mathematics. Economic planning without statistics is bound to be baseless. Statistics serve in administration, and facilitate the work of formulation of new policies. Financial institutions and investors utilise statistical data to summaries the past experience. Statistics are also helpful to an auditor, when he uses sampling techniques or test checking to audit the accounts of his client.

LIMITATIONS OF STATISTICS

The scope of the science of statistic is restricted by certain limitations :
1. The use of statistics is limited numerical studies: Statistical methods cannot be applied to study the nature of all type of phenomena. Statistics deal with only such phenomena as are capable of being quantitatively measured and numerically expressed. For, example, the health, poverty and intelligence of a group of individuals, cannot be quantitatively measured, and thus are not suitable subjects for statistical study.

2. Statistical methods deal with population or aggregate of individuals rather than with individuals. When we say that the average height of an Indian is 1 metre 80 centimetres, it shows the height not of an individual but as found by the study of all individuals.

3. Statistical relies on estimates and approximations : Statistical laws are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus the statistical inferences are uncertain.

4. Statistical results might lead to fallacious conclusions by deliberate manipulation of figures and unscientific handling. This is so because statistical results are represented by figures, which are liable to be manipulated. Also the data placed in the hands of an expert may lead to fallacious results. The figures may be stated without their context or may be applied to a fact other than the one to which they really relate. An interesting example is a survey made some years ago which reported that 33% of all the girl students at John Hopkins University had married

University teachers. Whereas the University had only three girls student at that time and one of them married to a teacher.

Distrust of Statistics

Due to limitations of statistics an attitude of distrust towards it has been developed. There are some people who place statistics in the category of lying and maintain that, "there are three degrees of comparison in lying-lies, dammed lies and statistics". But this attitude is not correct. The person who is handling statistics may be a liar or inexperienced. But that would be the fault not of statistics but of the person handling them. The person using statistics should not take them at their face value. He should check the result from an independent source. Also only experts should handle the statistics otherwise they may be misused. It may be noted that the distrust of statistics is due more to insufficiency of knowledge regarding the nature, limitations and uses of statistics then to any fundamental inadequacy in the science of statistics. Medicines are meant for curing people, but if they are unscientifically handle by quacks, they may prove fatal to the patient. In both the cases, the medicine is the same; but its usefulness or harmfulness depends upon the man who handles it. We cannot blame medicine for such a result. Similarly, if a child cuts his finger with a sharp knife, it is not a knife that is to be blamed, but the person who kept the knife at a place that the child could reach it. These examples help us in emphasising that if statistical facts are misused by some people it would be wrong to blame the statistics as such. It is the people who are to be blamed. In fact statistics are like clay which can be moulded in any way.

Collection of data

For studying a problem statistically first of all, the data relevant thereto must be collected. The numerical facts constitute the raw material of the statistical process. The interpretation of the ultimate conclusion and the decisions depend upon the accuracy with which the data are collected. Unless the data are collected with sufficient care and are as accurate as is necessary for the purposes of the inquiry, the result obtained cannot be expected to be valid or reliable. Before starting the collection of the data, it is necessary to know the sources from which the data are to be collected.

Primary and Secondary Sources

The original compiler of the data is the primary source. For example, the office of the Registrar General will be the primary source of the decennial population census figures.

A secondary source is the one that furnishes the data that were originally compiled

by someone else. If the population census figures issued by the office of the Registrar-General are published in the Indian year Book, this publication will be the secondary source of the population data. The source of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary source is known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data.

Choice between Primary and Secondary Data

An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter can be relied upon only by examining the following factors :—
(i) source from which they have been obtained;
(ii) their true significance;
(iii) completeness and
(iv) method to collection.
In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :
(i) Nature and scope of enquiry.
(ii) Availability of time and money.
(iii) Degree of accuracy required and
(iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.
However, it may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :
1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation

According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the

investigator is closely connected with the collection of data, it is bound to be more accurate. Thus, for example, if an inquiry is to be conducted into the family budgets and giving conditions of industrial labour, the investigation himself live in the industrial area as one of the industrial workers, mix with other residents and make patience and careful personal observation regarding how they spend, work and live.

2. Indirect Personal Observation

According to this method, the investigator interviews several persons who are either directly or indirectly in possession of the information sought to be collected. It may be distinguished form the first method in which information is collected directly from the persons who are involved in the inquiry. In the case of indirect personal observation, the persons from whom the information is being collected are known as witnesses or informants. However it should be ascertained that the informants really passes the knowledge and they are not prejudiced in favour of or against a particular view point.

This method is adopted in the following situations:

1. Where the information to be collected is of a complete nature.
2. When investigation has to be made over a wide area.
3. Where the persons involved in the inquiry would be reluctant to part with the information.

This method is generally adopted by enquiry committee or commissions appointed by government.

3. Schedules to be filled in by the informants

Under this method properly drawn up schedules or blank forms are distributed among the persons from whom the necessary figure are to be obtained. The informants would fill in the forms and return them to the officer incharge of investigation. The Government of India issued slips for the special enumeration of scientific and technical personnel at the time of census. These slips are good examples of schedules to be filled in by the informants.

The merit of this method is its simplicity and lesser degree of trouble and pain for the investigator. Its greatest drawback is that the informants may not send back the schedules duly filled in.

4. Information from Correspondents

Under this method certain correspondent are appointed in different parts of the field of enquiry, who submit their reports to the Central Office in their own manner. For example, estimates of agricultural wages may be periodically furnished to the Government by village school teachers.

The local correspondents being on the spot of the enquiry are capable of giving reliable information.

But it is not always advisable to place much reliance on correspondents, who have often got their own personal prejudices. However, by this method, a rough and approximate estimate is obtained at a very low cost. This method is also adopted by various departments of the government in such cases where regular information is to be collected from a wide area.

Questionnaire incharge of Enumerations

A questionnaire is a list of questions directly or indirectly connected with the work of the enquiry. The answers to these questions would provide all the information sought. The questionnaire is put in the charge of trained investigators whose duty is to go to all persons or selected persons connected with the enquiry. This method is usually adopted in case of large inquiries. The method of collecting data is relatively cheap. Also the information obtained is that of good quality. The main drawback of this method is that the enumerator (i.e., investigator in charge of the questionnaire) may be a biased one and may not enter the answer given by the information. Where there are many enumerators, they may interpret various terms in questionnaire according to their whims. To that extent the information supplied may be either inaccurate or inadequate or not comparable. This drawback can be removed to a great extent by training the investigators before the enquiry begins. The meaning of different questions may be explained to them so that they do not interpret them according to their whims.

Drafting the Questionnaire

The success of questionnaire method of collecting information depends on the proper drafting of the questionnaire. It is a highly specialized job and requires great deal of skill and experience. However, the following general principle may be helpful in framing a questionnaire:

1. The number of the questions should be kept to the minimum fifteen to twenty five may be a fair number.
2. The questions must be arranged in a logical order so that a natural and spontaneous reply to each is induced.
3. The questions should be short, simple and easy to understand and they should convey one meaning.
4. As far as possible, quotation of a personal and pecuniary nature should not be asked.
5. As far as possible the questions should be such that they can be answered briefly in ‘Yes’ or ‘No’, or in terms of numbers, place, date, etc.
6. The questionnaire should provide necessary instructions to the Informants. For instance, if there is a question on weight. It should be specified as to whether

weight is to be indicated in lbs or kilograms.
7. Questions should be objective type and capable of tabulation.

Specimen Questionnaire

We are giving below a specimen questionnaire of Expenditure Habits or Students residing in college Hostels.

Name of Student ..... Class .....
State and District of origin ..... Age .....

1. How much amount do you get from your father/guardian p.m. ?
2. Do you get some scholarship? If so, state the amount per month.
3. Is there any other source from which you get money regularly? (e.g. mother, brother or uncle).

4. How much do you spend monthly on the following items:

	Rs.
College Tuition Fee	.....
Hostel Food Expenses	.....
Other hostel fees	.....
Clothing	.....
Entertainment	.....
Smoking	.....
Miscellaneous	.....
Total	

5. Do you smoke? If so what is the daily expenditure on it?
6. Any other item on which you spend money ?

Sources of Secondary Data

There are number of sources from which secondary data may be obtained. They may be classified as follow. :

1. Published sources, and

2. Unpublished sources.

1. Published Sources
The various sources of published data are :
1. Reports and official publications of-
(a) International bodies such as the International Monetary Fund, International

Finance Corporation, and United Nations Organisation.
(b) Central and State Governments- such as the Report of the Patel Committee, etc.
2. Semi Official Publication. Various local bodies such as Municipal Corporation, and Districts Boards.
3. Private Publication of—
(a) Trade and professional bodies such as the Federation of India, Chamber of Commerce and Institute of Chartered Accountants of India.
(b) Financial and Economic Journals such as "Commerce", 'Capital' etc.
(c) Annual Reports of Joint Stock Companies.
(d) Publication brought out by research agendas, research scholars, etc.

2. Unpublished Sources

There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars, etc., such source can also be used where necessary.

Census and Sampling Techniques of Collection of Data

There are two important techniques of Data collection, (i) Census enquiry implies complete enumeration of each unit of the universe, (ii) In a sample survey, only a small part of the group, is considered, which is taken as representative. For example the population census in India implies the counting of each and every human being within the country. In practice sometimes it is not possible to examine every item in the population. Also many a time it is possible to obtain sufficiently accurate results by studying only a part of the "population". For example, if the marks obtained in statistics by 10 students in an examination are selected at random, say out of 100, then the average marks obtained by 10 students will be reasonably representative of the average marks obtained by all the 100 students. In such a case, the populations will be the marks of the entire group of 100 students and that of 10 students will be a sample.

LESSON 2
CONSTRUCTION OF FREQUENCY DISTRIBUTION AND GRAPHICAL PRESENTATION

What is frequency distribution

Collected and classified data are presented in a form of frequency distribution. Frequency distribution is simply a table in which the data are grouped into classes on the basis of common characteristics and the number of cases which fall in each class are recorded. It shows the frequency of occurrence of differentvalues of a single variable. A frequency distribution is constructed to satisfy three objectives :
(i) to facilitate the analysis of data,

(ii) to estimate frequencies of the unknown population distribution from the distribution of sample data, and

(iii) to facilitate the computation of various statistical measures.

Frequency distribution can be of two types :

1. Univariate Frequency Distribution.

2. Bivariate Frequency Distribution.

In this lesson, we shall understand the Univariate frequency distribution. Univariate distribution incorporatesdifferent values of one variable only whereas the Bivariate frequency distribution incorporates the values of two variables. The Univariate frequency distribution is further classified into three categories:

(i) Series of individual observations,

(ii) Discrete frequency distribution, and

(iii) Continuous frequency distribution.

Series of individual observations, is a simple listing of items of each observation. If marks of 14 students in statistics of a class are given individually, it will form a series of individual observations.

Marks obtained in Statistics :

Roll Nos.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Marks:	**60**	**71**	**80**	**41**	**81**	**41**	**85**	**35**	**98**	**52**	**50**	**91**	**30**	**88**

Marks in Ascending Order	Marks in Descending Order
30	98
35	91
41	88
41	85
50	81
52	80
60	71
71	60
80	52
81	50
85	41
88	41
91	35
98	30

Discrete Frequency Distribution: In a discrete series, the data are presented in such a way that exact measurements of units are indicated. In a discrete frequency distribution, we count the number of times each value of the variable in data given to you. This is facilitated through the technique of tally bars.
In the first column, we write all values of the variable. In the second column, a vertical bar called tally bar against the variable, we write a particular value has occurred four times, for the fifth occurrence, we put a cross tally mark (/) on the four tally bars to make a block of 5. The technique of putting cross tally bars at every fifth repetition facilitates the counting of the number of occurrences of the value. After putting tally bars for all the values in the data; we count the number of times each value is repeated and write it against the corresponding value of the variable in the third column entitled frequency. This type of representation of the data is called discrete frequency distribution.

We are given marks of 42 students:
55 51 57 40 26 43 46 41 46 48 33 40 26 40 40 41
43 53 45 53 33 50 40 33 40 26 53 59 33 39 55 48
15 26 43 59 51 39 15 45 26 15

We can construct a discrete frequency distribution from the above given marks.
Marks of 42 Students

Marks	Tally Bars	Frequency
15					3
26		5
33						4
39				2
40		5
41				2
43					3
45				2
46				2
48				2
50			1
51				2
53					3
55					3
57			1
59				2

Total 42

The presentation of the data in the form of a discrete frequency distribution is better than arranging but it does not condense the data as needed and is quite difficult to grasp and comprehend. This distribution is quite simple in case the values of the variable are repeated otherwise there will be hardly any condensation.

Continuous Frequency Distribution; If the identity of the units about a particular information collected, is neither relevant nor is the order in which the observations occur, then the first step of condensation is to classify the data into different classes by dividing the entire group of values of the variable into a suitable number of groups and then recording the number of observations in each group. Thus, we divide the total range of values of the variable (marks of 42 students) i.e. 59-15 = 44 into groups of 10 each, then we shall get (42/10) 5 groups and the distribution of marks is displayed by the following frequency distribution:

Marks of 42 Students

Marks (×)	Tally Bars	Number of Students (f)
15 – 25	|||	3
25 – 35	||||	9
35 – 45	||	12
45 – 55	||	12
55 – 65			6
	Total	**42**

The various groups into which the values of a variable are classified are known classes, the length of the class interval (10) is called the width of the class. Two values, specifying the class. are called the class limits. The presentation of the data into continuous classes with the corresponding frequencies is known as continuous frequency distribution. There are two methods of classifying the data according to class intervals :
(i) exclusive method, and
(ii) inclusive method
In an exclusive method, the class intervals are fixed in such a manner that upper limit of one class becomes the lower limit of the following class. Moreover, an item equal to the upper limit of a class would be excluded from that class and included in the next class. The following data are classified on this basis.

Income (Rs.)	No. of Persons
200 – 250	50
250 – 300	100
300 – 350	70
350 – 400	130
400 – 50	50
450 – 500	100

Total 500

It is clear from the example that the exclusive method ensures continuity of the data in as much as the upper limit of one class is the lower limit of the next class. Therefore, 50 persons have their incomes between 200 to 249.99 and a person whose income is 250 shall be included in the next class of 250 – 300. According to the inclusive method, an item equal to upper limit of a class is included in that class
itself. The following table demonstrates this method.

Income (Rs.)	**No.of Persons**
200 – 249	50
250 – 299	100
300 – 349	70
350 – 399	130
400 – 149	50
450 – 499	100

Total 500

Hence in the class 200 – 249, we include persons whose income is between Rs. 200 and Rs. 249.

Principles for Constructing Frequency Distributions

Inspite of the great importance of classification in statistical analysis, no hard and fast rules are laid down for it. A statistician uses his discretion for classifying a frequency distribution and sound experience, wisdom, skill and aptness for an appropriate classification of the data. However, the following guidelines must be considered to construct a frequency distribution:

1. **Type of classes:** The classes should be clearly defined and should not lead to any ambiguity.They should be exhaustive and mutually exclusive so that any value of variable corresponds to only class.
2. **Number of classes:** The choice about the number of classes in which a given frequency distribution should he divided depends upon the following

things;

(i) The total frequency which means the total number of observations in the distribution.

(ii) The nature of the data which means the size or magnitude of the values of the variable.

(iii) The desired accuracy.

(iv) The convenience regarding computation of the various descriptive measures of the frequency distribution such as means, variance etc.

The number of classes should not be too small or too large. If the classes are few, the classification becomes very broad and rough which might obscure some important features and characteristics of the data. The accuracy of the results decreases as the number of classes becomes smaller. On the other hand, too many classes will result in a few frequencies in each class. This will give an irregular pattern of frequencies in different classes thus makes the frequency distribution irregular. Moreover a large number of classes will render the distribution too unwieldy to handle. The computational work for further processing of the data will become quite tedious and time consuming without any proportionate gain in the accuracy of the results.

Hence a balance should be maintained between the loss of information in the first case and irregularity of frequency distribution in the second case, to arrive at a suitable number of classes. Normally, the number of classes should not be less than 5 and more than 20. Prof. Sturges has given a formula:

$$\mathbf{k = 1 + 3.322 \log n}$$

where k refers to the number of classes and n refers to total frequencies or number of observations.

The value of k is rounded to the next higher integer :

If n = 100 k = 1 + 3.322 log 100 = 1 + 6.644 = 8
If n = 10,000 k = 1 + 3.22 log 10,000 = 1 + 13.288 = 14
However, this rule should be applied when the number of observations are not very small.

Further, the number or class intervals should be such that they give uniform and unimodal distribution which means that the frequencies in the given classes increase and decrease steadily and there are no sudden jumps. The number of classes should be an integer preferably 5 or multiples of 5, 10, 15, 20, 25 etc. which are convenient for numerical computations.

3. Size of Class Intervals : Because the size of the class interval is inversely proportional to the number of classes in a given distribution, the choice about the size of the class interval will depend upon the sound subjective judgment of the statistician. An approximate value of the magnitude of the class interval say i can be calculated with the help of Sturge's Rule : where i slands for class magnitude or interval, Range refers to the difference between the largest and smallest value of the distribution, and n refers to total number of observations. If we are given the following information; n = 400, Largest item = 1300 and Smallest item = 340. then, = = Another rule to determine the size of class interval is that the length of the class interval should not he greater than of the estimated population standard deviation. If 6 is the estimate of population standard deviation then the length of class interval is given by: i £ 6/4, The size of class intervals should he taken as 5 or multiples of 5, 10, 15 or 20 for easy computations of various statistical measures of the frequency distribution, class intervals should be so fixed that each class has a convenient mid-point around which all the observations in that class cluster. It means that the entire frequency of the class is concentrated at the mid value of the class. It is always desirable to take the class intervals of equal or uniform magnitude throughout the frequency distribution.

4. Class Boundaries: If in a grouped frequency distribution there are gaps between the upper limit of any class and lower limit of the succeeding class (as in case of inclusive type of classification), there is a need to convert the data into a continuous distribution by applying a correction factor for continuity for determining new classes of exclusive type. The lower and upper class limits of new exclusive type classes are called class boundaries. If d is the gap between the upper limit of any class and lower limit of succeeding class, the class boundaries for any class are given by:

d/2 is called the correction factor. Let us consider the following example to understand :

Marks	**Class Boundaries**
20 – 24	(20 – 0.5, 24 + 0.5) i.e., 19.5 – 24.5
25 – 29	(25 – 0.5,29 + 0.5) i.e., 24.5 – 29.5
30 – 34	(30 – 0.5,34 + 0.5) i.e., 29.5 – 34.5
55 – 39	(35 – 0.5,39 + 0.5) i.e., 34.5 – 39.5
40 – 44	(40 – 0.5,44 + 0.5) i.e., 39.5 – 44.5

Correction factor = =

---

5. Mid-value or Class Mark: The mid value or class mark is the value of a variable which is exactly at the middle of the class. The mid-value of any class is obtained by dividing the sum of the upper and lower class limits by 2. **Mid value of a class = 1/2 [Lower class limit + Upper class limit]** The class limits should be selected in such a manner that the observations in any class are evenly distributed throughout the class interval so that the actual average of the observations in any class is very close to the mid-value of the class.
6. Open End Classes : The classification is termed as open end classification if the lower limit of the first class or the upper limit of the last class or both are not specified and such classes in which one of the limits is missing are called open end classes. For example, the classes like the marks less than 20 or age above 60 years. As far as possible open end classes should be avoided because in such classes the mid-value cannot be accurately obtained. But if the open end classes are inevitable then it is customary to estimate the class mark or mid-value for the first class with reference to the succeeding class. In other words, we assume that the magnitude of the first class is same as that of the second class.

Example : Construct a frequency distribution from the following data by inclusive method taking 4 as the class interval:

10	17	15	22	11	16	19	24	29	18
25	26	32	14	17	20	23	27	30	12
15	18	24	36	18	15	21	28	33	38
34	13	10	16	20	22	29	19	23	31

Solution : Because the minimum value of the variable is 10 which is a very convenient figure for taking the lower limit of the first class and the magnitude of the class interval is given to be 4, the classes for preparing frequency distribution by the Inclusive method will be 10 – 13, 14 – 17, 18 – 21, 22 – 25,..... 38 – 41.

Frequency Distribution

Class Interval	Tally Bars	Frequency (f)
10 – 13		**5**
14 – 17	|||	**8**
18 – 21	|||	**8**
22 – 25	||	**7**
26 – 29		**5**
30 – 33	||||	**4**
34 – 37	||	**2**
38 – 41			**1**

Example : Prepare a statistical table from the following :

Weekly wages (Rs.) of 100 workers of Factory A

88	23	27	28	86	96	94	93	86	99
82	24	24	55	88	99	55	86	82	36
96	39	26	54	87	100	56	84	83	46
102	48	27	26	29	100	59	83	84	48
104	46	30	29	40	101	60	89	46	49
106	33	36	30	40	103	70	90	49	50
104	36	37	40	40	106	72	94	50	60
24	39	49	46	66	107	76	96	46	67
26	78	50	44	43	46	79	99	36	68
29	67	56	99	93	48	80	102	32	51

Solution : The lowest value is 23 and the highest 106. The difference between the lowest and highest value is 83. If we take a class interval of 10. nine classes would be made. The first class should be taken as 20 – 30 instead of 23 – 33 as per the guidelines of classification.

Frequency Distribution of the Wages of 100 Workers

Wages (Rs.)	Tally Bars	Frequency (f)
29 – 30	||||	13
30 – 40			11
40 – 50	|||	18
50 – 60		10
60 – 70			6
70 – 80		5
80 – 90	||||	14
90 – 100	||	12
100 – 110			11
	Total	**100**

Graphs of Frequency Distributions

The guiding principles for the graphic representation of the frequency distributions are same as for the diagrammatic and graphic representation of other types of data. The information contained in a frequency distribution can be shown in graphs which reveals the important characteristics and relationships that are not easily discernible on a simple examination of the frequency tables. The most commonly used graphs for charting a frequency distribution are :

1. Histogram
2. Frequency polygon
3. Smoothed frequency curves
4. Ogives or cumulative frequency curves.

1. Histogram

The term ‘histogram’ must not be confused with the term ‘historigram’ which relates to time charts. Histogram is the best way of presenting graphically a simple

frequency distribution. The statistical meaning of histogram is that it is a graph that represents the class frequencies in a frequency distribution by vertical adjacent rectangles.

While constructing histogram the variable is always taken on the X-axis and the corresponding classinterval. The distance for each rectangle on the X-axis shall remain the same in case the class-intervals are uniform throughout; if they are different the width of the rectangles shall also change proportionately. The Yaxis represents the frequencies of each class which constitute the height of its rectangle. We get a series of rectangles each having a class interval distance as its width and the frequency distance as its height. The area of the histogram represents the total frequency.

The histogram should be clearly distinguished from a bar diagram. A bar diagram is one-dimensional where the length of the bar is important and not the width, a histogram is two-dimensional where both the length and width are important. However, a histogram can be misleading if the distribution has unequal class intervals and suitable adjustments in frequencies are not made.

The technique of constructing histogram is explained for :

(i) distributions having equal class-intervals, and

(ii) distributions having unequal class-intervals.

Example : Draw a histogram from the following data :

Classes	**Frequency**
0 – 10	5
10 – 20	11
20 – 30	19
30 – 40	21
40 – 50	16
50 – 60	10
60 – 70	8
70 – 80	6
80 – 90	3
90 – 100	1

Solution :

When class-intervals are unequal the frequencies must be adjusted before constructing a histogram. We take that class which has the lowest class-interval and adjust the frequencies of classes accordingly. If one class interval is twice as wide as the one having the lowest class-interval we divide the height of its rectangle by two, if it is three times more we divide it by three etc. the heights will be proportional to the ratios of the frequencies to the width of the classes.
Example : Represent the following data on a histogram.
Average monthly income of 1035 employees in a construction industry is given below :

Monthly Income (Rs.)	**No. of Workers**
600 – 700	25
700 – 800	100
800 – 900	150
900 – 1000	200
1000 – 1200	140
1200 – 1400	80
1400 – 1500	50
1500 – 1800	30
1800 or more	20

Solution : Histogram showing monthly incomes of workers :
When mid point are given, we ascertain the upper and lower limits of each class and then construct the histogram in the same manner.
Example : Draw a histogram of the following distribution :

Life of Electric Lamps (hours)	Firm A	Firm B	Frequency
10	10	102	87
10	30	130	105
10	50	48	226
10	70	360	230
10	90	18	352

Solution : Since we are given the mid points, we should ascertain the class limits. To calculate the class limits of various classes, take difference of two consecutive mid-points and divide the difference by 2, then add and subtract the value obtained from each mid-point to calculate lower and higher class-limits.

3. Frequency Polygon

This is a graph of frequency distribution which has more than four sides. It is particularly effective in comparing two or more frequency distributions. There are two ways of constructing a frequency polygon.

1)We may draw a histogram of the given data and then join by straight line the mid-points of the upper horizontal side of each rectangle with the adjacent ones. The figure so formed shall be frequency polygon. Both the ends of the polygon should be extended to the base line in order to make the area under frequency polygons equal to the area under Histogram.

2)Another method of constructing frequency polygon is to take the mid-points of the various classintervals and then plot the frequency corresponding to each point and join all these points by straight lines. The figure obtained by both the methods would be identical.

Frequency polygon has an advantage over the histogram. The frequency polygons of several distributions can be drawn on the same axis, which makes comparisons possible whereas histogram cannot be used in the same way. To compare histograms we need to draw them on separate graphs.

3)Smoothed Frequency Curve

A smoothed frequency curve can be drawn through the various points of the polygon. The curve is drawn by free hand in such a manner that the area included under the curve is approximately the same as that of the polygon. The object of drawing a smoothed curve is to eliminate all accidental variations which exists in the original data, while smoothening, the top of the curve would overtop the highest point of polygon particularly when the magnitude of the class interval is large. The curve should look as regular as possible and all sudden turns should be avoided. The extent of smoothening would depend upon the nature of the data. For drawing smoothed frequency curve it is necessary to first draw the polygon and then smoothen it. We must keep in mind the following points to smoothen a frequency graph:

(i) Only frequency distribution based on samples should be smoothened. Only continuous series should be smoothened.

(ii) The total area under the curve should be equal to the area under the histogram or polygon.

The diagram given below will illustrate the point :

4. Cumulative Frequency Curves or Ogives

We have discussed the charting of simple distributions where each frequency refers to the measurement of the class-interval against which it is placed. Sometimes it becomes necessary to know the number of items whose values are greater or less than a certain amount. We may, for example, be interested in knowing the number of students whose weight is less than 65 Ibs. or more than say 15.5 Ibs. To get this information, it is necessary to change the form of frequency distribution from a simple to a cumulative distribution. In a cumulative frequency distribution, the frequency of each class is made to include the frequencies of all the lower or all the upper classes depending upon the manner in which cumulation is done. The graph of such a distribution is called a cumulative frequency curve or an Ogive. There are two method of constructing ogives, namely:
(i) less than method, and

(ii) more than method

(iii) In less than method, we start with the upper limit of each class and go on adding the frequencies.When these frequencies are plotted we get a rising curve.In more than method, we start with the lower limit of each class and we subtract the frequency of each class from total frequencies. When these frequencies are plotted, we get a declining curve. This example would illustrate both types of ogives.

Example : Draw ogives by both the methods from the following data.

Distribution of weights of the students of a college (Ibs.)

Weights	**No. of Students**
90.5 – 100.5	5
100.5 – 110.5	34
110.5 – 120.5	139
120.5 – 130.5	300
130.5 – 140.5	367
140.5 – 150.5	319
150.5 – 160.5	205
160.5 – 170.5	76
170.5 – 180.5	43
180.5 – 190.5	16
190.5 – 200.5	3
200.5 – 210.5	4
210.5 – 220.5	3
220.5 – 230.5	1

Solution : First of all we shall find out the cumulative frequencies of the given data by less than method.

Less than (Weights)	**Cumulative Frequency**
100.5	5
110.5	39
120.5	178

130.5	478
140.5	845
150.5	1164
160.5	1369
170.5	1445
180.5	1488
190.5	1504
200.5	1507
210.5	1511
220.5	1514
230.5	1515

Plot these frequencies and weights on a graph paper. The curve formed is called an Ogive Now we calculate the cumulative frequencies of the given data by more than method.

More than (Weights)	**Cumulative Frequencies**
90.5	1515
100.5	1510
110.5	1476
120.5	1337
130.5	1037
140.5	670
150.5	351
160.5	146
170.5	70
180.5	27

190.5	11
200.5	8
210.5	4
220.5	1

By plotting these frequencies on a graph paper, we will get a declining curve which will be our cumulative frequency curve or Ogive by more than method.

Although the graphs are a powerful and effective method of presenting statistical data, they are not under all circumstances and for all purposes complete substitutes for tabular and other forms of presentation. The specialist in this field is one who recognizes not only the advantages but also the limitations of these techniques. He knows when to use and when not to use these methods and from his experience and expertise is able to select the most appropriate method for every purpose. **Example :**Draw an ogive by less than method and determine the number of companies earning profits between Rs. 45 crores and Rs. 75 crores :

Profits (Rs. crores)	**No. of Companies**	**Profits (Rs. crores)**	**No. Of Companies**
10—20	8	60—70	10
20—30	12	70—80	7
30—40	20	80—90	3
40—50	24	90—100	1
50—6.0	15		

Solution :

OGIVE BY LESS THAN METHOD

Profits (Rs. crores)	**No.of Companies**
Less than 20	8
Less than 30	20

Less than 40	40
Less than 50	64
Less than 60	79
Less than 70	89
Less than 80	96
Less than 90	99
Less than 100	100

It is clear from the graph that the number of companies getting profits less than Rs.75 crores is 92 and the number of companies getting profits less than Rs. 45 crores is 51. Hence the number of companies getting profits between Rs. 45 crores and Rs. 75 crores is 92 – 51 = 41.

Example : The following distribution is with regard to weight in grams of mangoes of a given variety. If mangoes of weight less than 443 grams be considered unsuitable for foreign market, what is the percentage of total mangoes suitable for it? Assume the given frequency distribution to be typical of the variety:

Weight in gms.	**No. of mangoes**	**Weight in gms.**	**No. of mangoes**
410 – 119	10	450 – 159	45
420 – 429	20	460 – 469	18
430 – 139	42	470 – 179	7
440 – 449	54		

Draw an ogive of 'more than' type of the above data and deduce how many mangoes will be more than 443 grams.
Solution : Mangoes weighting more than 443 gms. are suitable for foreign market. Number of mangoes weighting more than 443 gms. lies in the last four classes. Number of mangoes weighing between 444 and 449 grams would be Total number of mangoes weighing more than 443 gms. = 32.4 + 45 + 18 + 7 = 102.4
Percentage of mangoes =

Therefore, the percentage of the total mangoes suitable for foreign market is 52.25.

OGIVE BY MORE THAN METHOD

Weight more than (gms.)	**No. of Mangoes**
410	196
420	186
430	166
440	124
450	70
460	25
470	7

From the graph it can be seen that there are 103 mangoes whose weight will be more than 443 gms. and are suitable for foreign market.

DIAGRAM

Statistical data can be presented by means of frequency tables, graphs and diagrams. In this lesson, so far we have discussed the graphical presentation. Now we shall take up the study of diagrams. There are many variety of diagrams but here we are concerned with the following types only :

(i) Bar diagrams

(ii) Rectangles, squares and circles

Bar Diagram

A bar diagram may be simple or component or multiple. A simple bar diagram is used to represent only one variable. Length of the bars is proportional to the magnitude to be represented. But when we are interested in showing various parts of a whole, then we construct component or composite bar diagrams. Whenever comparisons of more than one variable is to be made at the same time, then multiple bar chart, which groups two or more bar charts together, is made use of. We shall now illustrate these by examples.

Example 1 : The following table gives the average approximate yield of rice in Ibs, per acre in various countries of the world in 2000–05.

Country	Yield in lbs. per acre
India	728
Siam	943
U.S.A.	1469
Italy	2903
Egypt	2153
Japan	2276

Indicate this by a suitable diagram

Solution :

In the above example, bars have been erected vertically. Also bars may be erected horizontally.

Example 2 : Draw a suitable diagram for the following date of expenditure of an average working class family,

Item of Expenditure	Percentage of Total Expenditure
Food	65
Clothing	10
Housing	12
Fuel and lighting	5
Miscellaneous	8

Solution : This is a case of percentage bar diagram as per cent of total expenditure is given.

Example 3 : Represent the following data with a suitable diagram.

Year	Men	Women	Children	Total
1990	180	110	100	390
1995	200	140	125	465
2000	250	200	150	600

Solution : This is case of a component or composite bar diagram. In addition to the number of men, women and children employed the total number of labour force for the three years is obvious.

Example 4 : The following table gives the number of companies at work in India for a few years. Represent the data by a suitable diagram.

Year	Public companies	Private companies	Total
2000	5000	20,000	25,000
2001	4000	16,000	20,000
2002	6000	18,000	24,000
2003	7000	21,000	28,000
2004	5000	15,000	20,000

Solution : The data can be shown with the help of a component bar diagram for each year. Also it can be shown with the help of multiple bar diagram which is drawn below. From the above diagram it is clear that comparison between the number of private companies and public companies is very sharp as the data is placed side by side. But as compared to a component bar diagram, no idea can be formed about the total number of companies at work.

Circles or Pie Diagrams

When circles are drawn to represent its idea equivalent to the figures, they are said to form piediagrams or circle diagrams. In case of circles the square roots of magnitudes are proportional to the radius. Suppose we are given the following figures 144, 81, 64, 16, 9 For the purpose of showing the data with the help of

circles, we shall find out the square roots of all the values. We get 12, 9, 8, 4 and 3. Now we shall use these values as radii of the different circles. It may be noted that in this case, bar diagram would not show the comparison and also it would be difficult to draw as there is a wide gap between the smallest and the highest value of the variate.

Sub-divided Pie-diagrams

Sub-divided pie-diagrams are used when comparison of the component parts is done with another and the total. The total value is equated to 360° and then the angles corresponding to component parts are calculated. Let us take an example.
Example : A rupee spent on "Khadi" is distributed as follows :

Farmer	20 Paise
Carder and spinner	35
Weaver	25
Washerman, dyer and printer	10
Administrative Agency	10
Total	100

Present the data in the form of a pie-diagram.

Solution : The angles subtended at the centre would be calculated as follows :

Expenditure	**Paise**	**Angle**
Farmer	20	72
Carder and spinner	35	126
Weaver	25	90
Washerman, dyer and printer	10	36
Administrative Agency	10	36
Total	100	360°

A sub-divided circle is drawn with the angles of 72°, 126°, 90°, 36° and 36° for the various items of expenditure. The above data could also be presented by a percentage bar diagram as there is not much difference between the smaller and the highest values. It is simple and easier to draw a bar diagram in this case.

Choice of a Suitable Diagram

The choice of diagram out of several ones in a given situation is a ticklish problem. The choice primarily depends upon two factors, (i) the nature of the data; and (ii) the type of people for whom the diagram is needed. On the nature of the data would depend whether to use one dimensional, two dimensional or three dimensional diagram, and if it is one dimensional, whether to adopt the simple bar or sub-divided bar, multiple bar or some other type.
While selecting the diagram the type of the people for whom the diagram is intended must also be considered. For example, for drawing attention of an uneducated mass, pictograms and cartograms are more effective.
There are different types of bars and the appropriate type of bar chart can be divided on the following basis :
(a) Simple bar charts should be used where changes in totals are required to be conveyed.
(b) Components bar charts are more useful where changes in totals as well as in the size of component figures (absolute ones) are required to be displayed.
(c) Percentage composition bar charts are better suited where changes in the relative size of components figures are to be exhibited.
(d) Multiple bar charts should be used where changes in the absolute values of the components figures are to be emphasised and the overall total is of no importance.
However, multiple and component bar charts should be used only when there are not more than three or four components as a large number of components make the bar charts too complex to enable worthwhile visual impression to be gained. When a large number of components have to be shown a pie chart is more suitable.
Occassionally, circles are used to represent size. But it is difficult to compare them and they should not be used when it is possible to use bars. This is because it is easier to compare the lengths of lines or bars than to compare areas or volume.

LESSON 3
MEASURES OF CENTRAL TENDENCY

What is Central Tendency

One of the important objectives of statistics is to find out various numerical values which explains the inherent characteristics of a frequency distribution. The first of such measures is averages. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which represent the entire distribution. The inherent inability of the human mind to remember a large body of numerical data compels us to few constants that will describe the data. Averages provide us the gist and give a bird's eye view of the huge mass of unwieldy numerical data. Averages are the typical values around which other items of the distribution congregate. This value lie between the two extreme observations of the distribution and give us an idea about the concentration of the values in the central part of the distribution. They are called the measures of central tendency.

Averages are also called measures of location since they enable us to locate the position or place of the distribution in question. Averages are statistical constants which enables us to comprehend in a single value the significance of the whole group. According to Croxlon and Cowden, an average value is a single value within the range of the data that is used to represent all the values in that series. Since an average is some where within the range of data, it is sometimes called a measure of central value. An average is the most typical representative item of the group to which it belongs and which is capable of revealing all important characteristics of that group or distribution.

What are the Objects of Central Tendency

The most important object of calculating an average or measuring central tendency is to determine a single figure which may be used to represent a whole series involving magnitudes of the same variable.

Second object is that an average represents the empire data, it facilitates comparison within one group or between groups of data. Thus, the performance of

the members of a group can be compared with the average performance of different groups.

Third object is that an average helps in computing various other statistical measures such as dispersion,skewness, kurtosis etc. Essential of a Good Average An average represents the statistical data and it is used for purposes of comparison, it must possess the following properties.

1. It must be rigidly defined and not left to the mere estimation of the observer. If the definition is rigid, the computed value of the average obtained by different persons shall be similar.

2. The average must be based upon all values given in the distribution. If the item is not based on all value it might not be representative of the entire group of data.
3. It should be easily understood. The average should possess simple and obvious properties. It should be too abstract for the common people.
4. It should be capable of being calculated with reasonable care and rapidity.

5. It should be stable and unaffected by sampling fluctuations.
6. It should be capable of further algebraic manipulation.
Different methods of measuring "Central Tendency" provide us with different kinds of averages. The following are the main types of averages that are commonly used:
1. **Mean**
(i) Arithmetic mean
(ii) Weighted mean
(iii) Geometric mean
(iv) Harmonic mean
2. **Median**
3. **Mode**
Arithmetic Mean: The arithmetic mean of a series is the quotient obtained by dividing the sum of the values by the number of items. In algebraic language, if X1, X2, X3 ..... Xn are the n values of a variate X.
Then the Arithmetic Mean is defined by the following formula:
=
=

Example : The following are the monthly salaries (Rs.) of ten employees in an office. Calculate the mean salary of the employees: 250, 275, 265, 280, 400, 490, 670, 890, 1100, 1250.

Solution : =

= Rs. 587

Short-cut Method: Direct method is suitable where the number of items is moderate and the figures are small sizes and integers. But if the number of items is large and/or the values of the variate are big, then the process of adding together all the values may be a lengthy process. To overcome this difficulty of computations, a short-cut method may be used. Short cut method of computation is based on an important characteristic of the arithmetic mean, that is, the algebraic sum of the deviations of a series of individual observation from their mean is always equal to zero. Thus deviations of the various values of the variate from an assumed mean computed and the sum is divided by the number of items. The quotient obtained is added to the assumed mean lo find the arithmetic mean.

Symbolically, = . where A is assumed mean and dx are deviations = (X – A).

We can solve the previous example by short-cut method.

Computation of Arithmetic Mean

Serial Number	Salary (Rupees) X	Deviations from assumed mean where dx = (X – A), A = 400
1.	250	–150
2.	275	–125
3.	265	–135
4.	280	–120
5.	400	0
6.	490	+90
7.	670	+270
8.	890	+490

9.	1100	+ 700
10.	1250	+ 850

N = 10 $\sum dx = 1870$

By substituting the values in the formula, we get
=
Computation of Arithmetic Mean in Discrete series. In discrete series, arithmetic mean may be computed by both direct and short cut methods. The formula according to direct method is:
=

where the variable values X1 X2, ... Xn, have frequencies f1, f2, ...fn and N = $\sum f$.

Example : The following table gives the distribution of 100 accidents during seven days of the week in a given month. During a particular month there were 5 Fridays and Saturdays and only four each of other days. Calculate the average number of accidents per day.

Days :	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
Number of accidents :	20	22	10	9	11	8	20	= 100

Solution : Calculation of Number of Accidents per Day

Day	No. of Accidents X	No. of Days in Month f	Total Accidents fX

Sunday	20	4	80
Monday	22	4	88

Tuesday	10	4	40
Wednesday	9	4	36
Thursday	11	4	44
Friday	8	5	40
Saturday	20	5	100
	100	N = 30	∑fX = 428

= = 14.27 = 14 accidents per day

The formula for computation of arithmetic mean according to the short cut method is = where A is assumed mean, dx = (X – A) and N = ∑f.
We can solve the previous example by short-cut method as given below :

Calculation of Average Accidents per Day

Day	X	dx = X – A (where A = 10)	f	fdx
Sunday	20	+ 10	4	+ 40
Monday	22	+ 12	4	+ 48
Tuesday	10	+ 0	4	+ 0
Wednesday	9	– 1	4	– 4
Thursday	11	+ 1	4	+ 4
Friday	8	– 2	5	– 10
Saturday	20	+ 10	5	+ 50
			30	+ 128

= = = 14 accidents per day

Calculation of arithmetic mean for Continuous Series: The arithmetic mean can be computed both by direct and short-cut method. In addition, a coding method or step deviation method is also applied for simplification of calculations. In any case, it is necessary to find out the mid-values of the various classes in the frequency distribution before arithmetic mean of the frequency distribution can be computed. Once the mid-points of various classes are found out, then the process of the calculation of arithmetic mean is same as in the case of discrete series. In case of direct method, the formula to be used:

= , when m = mid points of various classes and N = total frequency In the short-cut method, the following formula is applied:
= where dx = (m – A) and N = ∑f

The short-cut method can further be simplified in practice and is named coding method. The deviations from the assumed mean are divided by a common factor to reduce their size. The sum of the products of the deviations and frequencies is multiplied by this common factor and then it is divided by the total frequency and added to the assumed mean. Symbolically
= where and i = common factor

Example : Following is the frequency distribution of marks obtained by 50 students in a test of Statistics:

Marks	Number of Students

---

0 – 10	4
10 – 20	6
20 – 30	20
30 – 40	10
40 – 50	7
50 – 60	3

Calculate arithmetic mean by:
(i) direct method.
(ii) short-cut method, and
(iii) coding method

Solution : Calculation of Arithmetic Mean

X	f	m	fm	dx = (m – A) (where A = 25)	fdx where i = 10		fd'x
0 – 10	4	5	20	– 20	– 2	– 80	– 8
10 – 20	6	15	90	– 10	– 1	– 60	– 6
20 – 30	20	25	500	0	0	0	0
30 – 40	10	35	350	+10	+1	100	+ 10
40 – 50	7	45	315	+20	+2	140	+ 14
50 – 60	3	55	165	+30	+3	90	+ 9
	N = 50		∑fm = 1440		∑fdx = 190		∑fd'x = 19

Direct Method :
= = marks
Short-cut Method :
= marks
Coding Method :
= marks
We can observe that answer of average marks i.e. 28.8 is identical by all methods.

Mathematical Properties of the Arithmetic Mean

(i) The sum of the deviation of a given set of individual observations from the arithmetic mean is always zero. Symbolically. = 0. It is due to this property that the arithmetic mean is characterised as the centre of gravity i.e., the sum of positive deviations from the mean is equal to the sum of negative deviations.

(ii) The sum of squares of deviations of a set of observations is the minimum when deviations are taken from the arithmetic average. Symbolically, = smaller than $\sum$ (X – any other value)2. We can verify the above properties with the help of the following data:

Values X	Deviations from		Deviations from Assumed Mean	
3	– 6	36	– 7	49
5	– 4	16	– 5	25
10	1	1	0	0
12	3	9	2	4
15	6	36	5	25
Total = 45	0	98	– 5	103

= where A (assumed mean) = 10

(iii) If each value of a variable X is increased or decreased or multiplied by a constant k, the arithmetic mean also increases or decreases or multiplies by the same constant.

(iv) If we are given the arithmetic mean and number of items of two or more groups, we can compute the combined average of these groups by apply the following formula :

=

where refers to combined average of two groups.

refers to arithmetic mean of first group.
refers to arithmetic mean of second group.
N1 refers to number of items of first group, and N2 refers to number of items of second group
We can understand the property with the help of the following examples.
Example : The average marks of 25 male students in a section is 61 and average marks of 35 female students in the same section is 58. Find combined average marks of 60 students.
Solution : We are given the following information.

61, N1 = 25, = 58, N2 = 35
Apply =

Example : The mean wage of 100 workers in a factory, running two shifts of 60 and 40 workers respectively is Rs.38. The mean wage of 60 workers in morning shift is Rs.40. Find the mean wage of 40 workers working in the evening shift.
Solution : We are given the following information
= 40, N1 = 60, = ?, N2 = 40, = 38, and N = 100
Apply =
38 = or 3800 = 2400 +
=

Example : The mean age of a combined group of men and women is 30 years. If the mean age of the groupof men is 32 and that of women group is 27. find out the percentage of men and women in the group.
Solution : Let us take group of men as first group and women as second group. Therefore. = 32 years. =27 years, and = 30 years. In the problem, we are not given the number of men and women. We can assume
N1 + N2 = 100 and therefore. N1 = 100 – N2
Apply =
30 = (Substitute N1 = 100 – N2)
30 × 100 = 32(100 – N2) + 27N2 or 5N2 = 200
N2 = 200/5 – 40%
N1 = (100 – N2) = (100 – 40) = 60%
Therefore, the percentage of men in the group is 60 and that of women is 40.

(v) Because =
» ∑X = N.

If we replace each item in the series by the mean, the sum of these substitutions will be equal to the sum of the individual items. This property is used to find out the aggregate values and corrected averages. We can understand the property with the help of an example.
Example : Mean of 100 observations is found to be 44. If at the time of computation two items are wrongly taken as 30 and 27 in place of 3 and 72. Find the corrected average.
Solution : =

∑X = N. = 100 × 44 = 4400
Corrected ∑X = ∑X + correct items – wrong items = 4400 + 3 + 72 – 30 – 27 = 4418
Corrected average =

Calculation of Arithmetic mean for Open-End Classes
Open-end classes are those in which lower limit of the first class and the upper limit of the last class are not defined. In these series, we can not calculate mean unless we make an assumption about the unknown limits. The assumption depends upon the class-interval following the first class and preceding the last class. For example:

Marks	No. of Students
Below 15	4
15 – 30	6
30 – 15	12
45 – 60	8
Above 60	7

In this example, because all defined class-intervals are same, the assumption would be that the first and last class shall have same class-interval of 15 and hence the lower limit of the first class shall be zero and upper limit of last class shall be 75. Hence first class would be 0 – 15 and the last class 60 – 75.

What happens in this case?

Marks	No. of Students
Below 10	4
10 – 30	7
30 – 60	10
60 – 100	8
Above 100	4

In this problem because the class interval is 20 in the second class, 30 in the third, 40 in the fourth class and so on. The class interval is increasing by 10. Therefore the appropriate assumption in this case would be that the lower limit of the first class is zero and the upper limit of the last class is 150. In case of other open-end class distributions the first class limit should be fixed on the basis of succeeding class interval and the last class limit should be fixed on the basis of preceding class interval.

If the class intervals are of varying width, an effort should be made to avoid calculating mean and mode. It is advisable to calculate median.

Weighted Mean

In the computation of arithmetic mean, we give equal importance to each item in the series. Raja Toy Shop sell : Toy Cars at Rs. 3 each; Toy Locomotives at Rs. 5 each; Toy Aeroplane at Rs. 7 each; and Toy Double Decker at Rs. 9 each. What shall be the average price of the toys sold ? If the shop sells 4 toys one of each kind.

(Mean Price) =

In this case the importance of each toy is equal as one toy of each variety has been sold. While computing the arithmetic mean this fact has been taken care of including the price of each toy once only.

But if the shop sells 100 toys, 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four toys to the dealer is not equal as a source of earning revenue. In fact their respective importance is equal to the number of units of each toy sold, i.e., the importance of Toy car is 50; the importance of Locomotive is 25; the importance of Aeroplane is 15; and the importance of Double Decker is 10.

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. These quantities are called as ‘weights’ in statistical language. Weight is represented by symbol W and SW represents the sum of weights.
While determining the average price of toy sold these weights are of great importance and are taken into account to compute weighted mean.
=
where, W1, W2, W3, W4 are weights and X1, X2, X3, X4 represents the price of 4 varieties of toy.
Hence by substituting the values of W1, W2, W3, W4 and X1, X2, X3, X4, we get
=
=
The table given below demonstrates the procedure of computing the weighted Mean.
Weighted Arithmetic mean of Toys by the Raja Shop.

Toy	Price per toy (Rs.) X	Number Sold W	Price x Weight WX
Car	3	50	150
Locomotive	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		$\sum W = 100$	$\sum WX = 470$

Example: The table below shows the number of skilled and unskilled workers in two localities along with their average hourly wages.

	Ram Nagar		Shyam Nagar	
Worker Category	Number	Wages (per hour)	Number	Wages (per hour)
Skilled	150	1.80	350	1.75
Unskilled	850	1.30	650	1.25

Determine the average hourly wage in each locality. Also give reasons why the results show that the average hourly wage in Shyam Nagar exceed the average hourly wage in Ram Nagar even though in Shyam Nagar the average hourly wages of both categories of workers is lower. It is required to compute weighted mean.
Solution :

	Ram Nagar			Shyam Nagar		
	X	W	WX	X	W	WX
Skilled	1.80	150	270	1.75	350	612.50
Unskilled	1.30	850	1105	1.25	650	812.50
Total		1000	1375		1000	1425

It may be noted that weights are more evenly assigned to the different categories of workers in Shyam Nagar than in Ram Nagar.
Geometric Mean :
In general, if we have n numbers (none of them being zero), then the GM. is defined as

G.M. =

In case of a discrete series, if x1, x2,..... xn occur f1, f2, ... fn times respectively and N is

the
total frequency (i.e. N = f1 + f2.....fn), then
G.M. =

For convenience, use of logarithms is made extensively to calculate the nth root. In terms of logarithms
G.M. =
= , where AL refers to antilog.
and in case of continuous series, G.M. =
Example: Calculate G.M. of the following data : 2, 4, 8

Solution: G.M. =
In terms of logarithms, the question can be solved as follows :
log 2 = 0.3010, log 4 = 0.6021, and log 8 = 9.9031
Apply the formula :
G.M. =

Example : Calculate geometric mean of the following data :

x	5	6	7	8	9	10	11
f	2	4	7	10	9	6	2

Solution : Calculation of G.M.

x	log x	f	f log x
5	0.6990	2	1.3980
6	0.7782	4	3.1128
7	0.8451	7	5.9157
8	0.9031	10	9.0310
9	0.9542	9	8.5878
10	1.0000	6	6.0000
11	1.0414	2	2.0828

N = 40 $\sum f \log x = 36.1281$

G.M. =

Example : Calculate G.M. from the following data :

X	f
9.5 – 14.5	10
14.5 – 19.5	15
19.5 – 24.5	17
24.5 – 29.5	25
29.5 – 34.5	18
34.5 – 39.5	12
39.5 – 44.5	8

Solution: Calculation of G.M.

X	m	log m	f	f log m
9.5 – 14.5	12	1.0792	10	10.7920
14.5 – 19.5	17	1.2304	15	18.4560
19.5 – 24.5	22	1.3424	17	22.8208
24.5 – 29.5	27	1.4314	25	35.7850
29.5 – 34.5	32	1.5051	18	27.0918
34.5 – 39.5	37	1.5682	12	18.8184
39.5 – 14.5	42	1.6232	8	12.9850

$$N = 105 \quad \sum f \log m = 146.7410$$

Specific uses of G.M. : The geometric Mean has certain specific uses, some of them are :
(i) It is used in the construction of index numbers.
(ii) It is also helpful in finding out the compound rates of change such as the rate of growth of population in a country.
(iii) It is suitable where the data are expressed in terms of rates, ratios and percentage.
(iv) It is quite useful in computing the average rates of depreciation or appreciation.
(v) It is most suitable when large weights are to be assigned to small items and small weights to large items.

Example : The gross national product of a country was Rs. 1.000 crores 10 years earlier. It is Rs. 2,000 crores now. Calculate the rate of growth in G.N.P.
Solution: In this case compound interest formula will be used for computing the average annual per cent increase of growth.

$Pn = Po (l + r)n$
where Pn = principal sum (or any other variate) at the end of the period.
Po = principal sum in the beginning of the period.
r = rate of increase or decrease.
n = number of years.
It may be noted that the above formula can also be written in the following form :

r =
Substituting the values given in the formula, we have
r =
=
Hence, the rate of growth in GNP is 7.18%.

Example : The price of commodity increased by 5 per cent from 2001 to 2002, 8 percent from 2002 to 2003 and 77 per cent from 2003 to 2004. The average increase from 2001 to 2004 is quoted at 26 per cent and not 30 per cent. Explain

this slatement and verify the arithmetic.
Solution : Taking Pn as the price at the end of the period, Po as the price in the beginning, we can substitute the values of Pn and Po in the compound interest formula. Taking Po = 100; Pn = 200.72

Pn = Po (l + r)n

200.72 = 100 (1 + r)3

or (l + r)3 = or l + r =
r = – 1 = 1.260 – 1 = 0.260 = 26%

Thus increase is not average of (5 + 8 + 77)/3 = 30 percent. It is 26% as found out by G.M.

Weighted G.M.: The weighted GM. is calculated with the help of the following formula :
G.M. =
=
=
Example : Find out weighted G.M. from the following data :

Group Index Number Weights

Food 352 48
Fuel 220 10
Cloth 230 8
House Rent 160 12
Misc. 190 15

Solution : Calculation of Weighted GM.

Group Index Number (x) Weights (W) Log x w log x

Food	352	48	2.5465	122.2320
Fuel	220	10	2.3424	23.4240
Cloth	230	8	2.3617	17.8936
House Rent	160	12	2.2041	26.4492
Misc.	190	15	2.2788	34.1820
		93		225.1808

G.M. =

Example: A machine depreciates at the rate of 35.5% per annum in the first year, at the rate of 22.5% per annum in the second year, and at the rate of 9.5% per annum in the third year, each percentage being computed on the actual value. What is the average rate of depreciation?

Solution: Average rate of depreciation can be calculated by taking GM.

Year	X (value is taking 100 as base)	log X
I	100 – 35.5 = 64.5	1.8096
II	100 – 22.5 = 77.5	1.8893
III	100 – 9.5 = 90.5	1.9566
		$\sum \log X = 5.6555$

Apply G.M. =

Average rate of depreciation = 100-76.77 = 23.33%.

Example : The arithmetic mean and geometric mean of two values are 10 and 8 respectively. Find the values.
Solution : If two values are taken as a and b, then
and

or a + b = 20, ab = 64
then a – b =
Now, we have a + b = 20, ...(i)
Solving for a and b, we get a = 4 and b = 16. ...(ii)

Harmonic Mean : The harmonic mean is defined as the reciprocals of the average of reciprocals of items in a series. Symbolically,
H. M. =
In case of a discrete series,
H. M. =
and in case of a continuous series,
H. M. =
It may be noted that none of the values of the variable should be zero.
Example: Calculate harmonic mean from the following data: 5, 15, 25, 35 and 45.
Solution :

X
5 0.20
15 0.067
25 0.040
35 0.029
45 0.022
N = 5

H.M. =

Example : From the following data compute the value of the harmonic mean :
x : 5 15 25 35 45
f : 5 15 10 15 5
Solution : Calculation of Harmonic Mean

X f

---

5 5 0.200 1.000
15 15 0.067 1.005
25 10 0.040 0.400
35 15 0.29 0.435
45 5 0.022 0.110

$\sum f = 50$

H.M. =

Example : Calculate harmonic mean from the following distribution :

x	f
0 – 10	5
10 – 20	15
20 – 30	10
30 – 40	15
40 – 50	5

Solution : First of all, we shall find out mid points of the various classes. They are 5, 15, 25, 35 and 45.

Then we will calculate the H.M. by applying the following formula :

H.M. =

Calculation of Harmonic Mean

---

x (Mid Points) f

---

5 5 0.200 1.000
15 15 0.067 1.005
25 10 0.040 0.400

35	15	0.29	0.435
45	5	0.022	0.110

$\sum f = 50$

H.M. =

Application of Harmonic Mean to special cases: Like Geometric means, the harmonic mean is also applicable to certain special types of problems. Some of them are:

(i) If in averaging time rates, distance is constant, then H.M. is to be calculated.
Example: A man travels 480 km a day. On the first day he travels for 12 hours @ 40 km. per hour and second day for 10 hours @ 48 km. per hour. On the third day he travels for 1.5 hours @ 32 km. per hour. Find his average speed.
Solution: We shall use the harmonic mean,
H.M. =
The arithmetic mean would be
(ii) If, in averaging the price data, the prices are expressed as "quantity per rupee". Then harmonic mean should be applied.
Example : A man purchased one kilo of cabbage from each of four places at the rate of 20 kg. 16 kg. 12 kg., and 10 kg. per rupees respectively. On the average how many kilos of cabbages he has purchased per rupee.
Solution : H.M. =

POSITIONAL AVERAGES

Median

The median is that value of the variable which divides the group in two equal parts. One part comprising the values greater than and the other all values less than median. Median of a distribution may be defined as that value of the variable which exceeds and is exceeded by the same number of observation. It is the value such that the number of observations above it is equal to the number of observations below it.
Thus we know that the arithmetic mean is based on all items of the distribution, the

median is positional average, that is, it depends upon the position occupied by a value in the frequency distribution.

When the items of a series are arranged in ascending or descending order of magnitude the value of the middle item in the series is known as median in the case of individual observation. Symbolically.

Median = size of th item

It the number of items is even, then there is no value exactly in the middle of the series. In such a situation the median is arbitrarily taken to be halfway between the two middle items. Symbolically.

Median =

Example : Find the median of the following series:

Solution : Computation of Median

(i) (ii)

Serial No. X Serial No. X

1 3 1 5

2 4 2 5

3 4 3 7

4 5 4 9

5 6 5 11

6 8 6 12

7 8 7 15

8 8 8 28

9 10

N = 9 N – 8

Far (i) series Median = size of th item = size of the th item = size of 5th item = 6

For (ii) series Medium = size of th item = size of the th item

= =

Location of Median in Discrete series: In a discrete series, medium is computed in the following manner:

(i) Arrange the given variable data in ascending or descending order,

(ii) Find cumulative frequencies.

(iii) Apply Med. = size of th item

(iv) Locate median according to the size i.e., variable corresponding to the size or for next cumulative frequency.

Example: Following are the number of rooms in the houses of a particular locality. Find median of the data:
No. of rooms: 3 4 5 6 7 8
No of houses: 38 654 311 42 12 2

Solution: Computation of Median

No. of Rooms X	No. of Houses f	cumulative Frequency Cf
3	38	38
4	654	692
5	311	1003
6	42	1045
7	12	1057
8	2	1059

Median = size of th item = size of th item = 530 th item.
Median lies in the cumulative frequency of 692 and the value corresponding to this is 4
Therefore, Median = 4 rooms.
In a continuous series, median is computed in the following manner:
(i) Arrange the given variable data in ascending or descending order.
(ii) If inclusive series is given, it must he converted into exclusive series to find real class interval
(iii) Find cumulative frequencies.
(iv) Apply Median = size of th item to ascertain median class.
(v) Apply formula of interpolation to ascertain the value of median.

Median = l1 + or Median = l2 –
where, l1 refers to lower limit of median class,
l2 refers to higher limit of median class,
cfo refers cumulative frequency of previous to median class,
f refers to frequency of median class,
Example: The following table gives you the distribution of marks secured by some students in an examination:

Marks	No. of Students
0—20	42
21—30	38
31—40	120
41—50	84
51— 60	48
61—70	36
71—80	31

Find the median marks.

Solution: Calculation of Median Marks

Marks (x)	No. of Students (f)	cf
0 – 20	42	42
21 – 30	38	80
31 – 40	120	200
41 – 50	84	284
51 – 60	48	332
61 – 70	36	368
71 – 80	31	399

Median = size of th item = size of th item = 199.5 th item.
which lies in (31 – 40) group, therefore the median class is 30.5 – 40.5.
Applying the formula of interpolation.
Median = l1 +
= 30.5 +

Related Positional Measures: The median divides the series into two equal parts. Similarly there are certain other measures which divide the series into certain equal parts, there are first quartile, third quartile, deciles, percentiles etc. If the items are arranged in ascending or descending order of magnitude, Qt is that value which covers l/4th of the total number of items. Similarly, if the total number of items are divided into ten equal parts, then, there shall be nine deciles.
Symbolically,

First decile (Q1) = size of th item

Third quartile (Q3) = size of th item

First decile (D1) = size of th item

Sixth decile (D6) = size of th item

First percentile (P1) = size of th item

Once values of the items are found out, then formulae of interpolation are applied for ascertaining the value of Q1, Q2, D1, D4, P40 etc.

Example: Calculate Q1, Q3, D2, and P5, from following data:

Marks :	Below 10	10 – 20	20 – 40	40 – 60	60 – 80	above 80
No. of Students:	8	10	22	25	10	5

Solution: Calculation of Positional Values

Marks	No. of Students (f)	C.f.
Below 10	8	8
10 – 20	10	18

20 – 40	22	40
40 – 60	25	65
60 – 80	10	75
Above 80	5	80

N = 80

Q1 = size of th item = = 20th item
Hence Qt lies in the class 20 – 40, apply
Q1 = where l1 = 20, Cfo = 18, f = 22 and i = (l2 – l1) = 20
By substituting the values, we get
Q1 =
Similarly, we can calculate
Q3 = size of th item = th item = 60th item
Hence Q3 lies in the class 40 – 60, apply
Q3 = where l1 = 40, , Cfo = 40, f = 25, i = 20
\ Q3 =
D2 = size of th item = 16th item. Hence D2 lies in the class 10 – 20.
D2 = where
D2 =
P5 = size of th item = th item = 4th item. Hence P5 lies in the class 0 – 10.
P5 = where l1 = 0, Cfo = 0, f = 8, i = 10
P5 =

Calculation of Missing Frequencies:
Example: In the frequency distribution of 100 families given below: the number of families corresponding to expenditure groups 20 – 40 and 60 – 80 are missing from the table. However the median is known to be 50. Find out the missing frequencies.

Expenditure:	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
No. of families:	14	?	27	?	15

Solution: We shall assume the missing frequencies for the classes 20—40 to be x and 60—80 to y

---

Expenditure (Rs.)	No. of Families	C.f.
0 – 20	14	14
20 – 40	x	14 + x
40 – 60	27	14 + 27 + x
60 – 80	y	41 + x + y
80 – 100	15	41 + 15 + x + y

N = 100 = 56 + x + y

From the table, we have N = SF = 56 + x + y = 100

x + y = 100 – 56 + 44

Median is given as 50 which lies in the class 40 – 60, which becomes the median class, By using the median formula we get:

Median =

50 = or 50 =

or 50 – 40 = or 50 – 40 =

or 10 × 27 = 720 – 20x or 270 = 720 – 20x

20x = 720 – 270

x =

By substitution the value of x in the equation,

x + y = 44

We get, 22.5 + y = 44

y = 44 – 22.5 = 21.5

Hence frequency for the 20 – 40 is 22.5 and 60 – 80 is 21.5

Mode

Mode is that value of the variable which occurs or repeats itself maximum number of item. The mode is most " fashionable" size in the sense that it is the most common and typical and is defined by Zizek as "the value occurring most

frequently in series of items and around which the other items are distributed most densely." In the words of Croxton and Cowden, the mode of a distribution is the value at the point where the items tend to be most heavily concentrated. According to A.M. Tuttle, Mode is the value which has the greater frequency density in its immediate neighbourhood. In the case of individual observations, the mode is that value which is repeated the maximum number of times in the series. The value of mode can be denoted by the alphabet z also.

Example : Calculate mode from the following data:

Sr. Number :	1	2	3	4	5	6	7	8	9	10
Marks obtained :	10	27	24	12	27	27	20	18	15	30

Solution :

Marks	No. of students	
10	1	
12	1	
15	1	
18	1	
20	1	
24	1	
27	3	Mode is 27 marks
30	1	

Calculation of Mode in Discrete series. In discrete series, it is quite often determined by inspection.We can understand with the help of an example:

X	1	2	3	4	5	6	7
f	4	5	13	6	12	8	6

By inspection, the modal size is 3 as it has the maximum frequency. But this test of greatest frequency is not fool proof as it is not the frequency of a single class, but

also the frequencies of the neighbour classes that decide the mode. In such cases, we shall be using the method of Grouping and Analysis table.

Size of shoe	1	2	3	4	5	6	7
Frequency	4	5	13	6	12	8	6

Solution : By inspection, the mode is 3, but the size of mode may be 5. This is so because the neighboring frequencies of size 5 are greater than the neighbouring frequencies of size 3. This effect of neighbouring frequencies is seen with the help of grouping and analysis table technique.

When there exist two groups of frequencies with equal magnitude, then we should consider either both or omit both while analysing the sizes of items.

Analysis Table

Column	Size of Items with Maximum Frequency
1	3
2	5, 6
3	1, 2, 3, 4, 5
4	4, 5, 6
5	5, 6, 7
6	3, 4, 5

Item 5 occurs maximum number of times, therefore, mode is 5. We can note that by inspection we had determined 3 to be the mode.
Determination of mode in continuous series: In the continuous series, the determination of mode requires one additional step. Once the modal class is determined by inspection or with the help of grouping technique, then the following formula of interpolation is applied:
Mode = or Mode =
l1 = lower limit of the class, where mode lies,

l2 = upper limit of the class, where mode lies,
f 0 = frequency of the class preceding the modal class.
f 1 = frequency of the class, where mode lies.
f 2 = frequency of the class succeeding the modal class.
Example: Calculate mode from the following frequency distribution:

Variable	Frequency
0 – 10	5
10 – 20	10
20 – 30	15
30 – 40	14
40 – 50	10
50 – 60	5
60 – 70	3

Analysis Table

Column	Size of Item with Maximum Frequency
1	20 – 30
2	20 – 30, 30 – 40
3	10 – 20, 20 – 30
4	0 – 10, 10 – 20, 30 – 40
5	10 – 20, 20 – 30, 30 – 40
6	20 – 30, 30 – 40, 40 – 50

Modal group is 20 – 30 because it has occurred 6 times. Applying the formula of interpolation,
Mode =
=
Calculation of mode where it is ill defined. The above formula is not applied where there are many modal values in a series or distribution. For instance there may be two or more than two items having the maximum frequency. In these cases, the

series will be known as bimodal or multimodal series. The mode is said to be ill-defined and in such cases the following formula is applied.
Mode = 3 Median – 2 Mean.
Example: Calculate mode of the following frequency data:

Variate	Value Frequency
10 – 20	5
20 – 30	9
30 – 40	13
40 – 50	21
50 – 60	20
60 – 70	15
70 – 80	8
80 – 90	3

Solution : First of all, ascertain the modal group with the help of process of grouping.

Analysis Table

Column	Size of Item with Maximum Frequency
1	40 – 50
2	50 – 60, 60 – 70
3	40 – 50, 50 – 60
4	40 – 50, 50 – 60, 60 – 70
5	20 – 30, 30 – 40, 40 – 50, 50 – 60, 60 – 70, 70 – 80
6	30 – 40, 40 – 50, 50 – 60

There are two groups which occur equal number of items. They are 40 – 50 and 50 – 60. Therefore, we will apply the following formula:
Mode = 3 median – 2 mean and for this purpose the values of mean and median are required to be computed.
Calculation of Mean and Median

Variate X	Frequency f	Mid Values m	d'x	fd'x	Cf	
10 – 20	5	15	– 3	– 15	5	
20 – 30	9	25	– 2	– 18	14	
30 – 40	13	35	–1	– 13	27	
40—50	21	45	0	0	48	Median is the
50—60	20	55	+ 1	+ 20	68	value of
60—70	15	65	+ 2	+ 30	83	item which lies
70—80	8	75	+ 3	+ 24	91	in (40 – 50) group
80—90	3	85	+ 4	+ 12	94	
	N = 94			∑fd' = + 40		

= Med. =

= =

Mode = 3 median – 2 mean

= 3 (49.5)-2 (49.2)= 148.5-98.4 = 50.1

Determination of mode by curve fitting : Mode can also be computed by curve fitting. The following steps are to be taken;

(i) Draw a histogram of the data.

(ii) Draw the lines diagonally inside the modal class rectangle, starting from each upper corner of the rectangle to the upper corner of the adjacent rectangle.

(iii) Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis.

The abscissa of the point at which the perpendicular line meets is the value of the mode.

Example : Construct a histogram for the following distribution and, determine the mode graphically:

X : 0 – 10 10 – 20 20 – 30 30 – 40 40 – 50
f : 5 8 15 12 7
Verify the result with the help of interpolation.
Solution :
Mode =
=
Example: Calculate mode from the following data:

Marks No. of Students

Below 10 4
‘’ 20 6
‘’ 30 24
‘’ 40 46
‘’ 50 67
‘’ 60 86
‘’ 70 96
‘’ 80 99
‘’ 90 100

Solution :

Since we are given the cumulative frequency distribution of marks, first we shall convert it into the normal frequency distribution:

Marks Frequencies

---

0 – 10 4
10 – 20 6 – 4 = 2
20 – 30 24 – 6 = 18
30 – 40 46 – 24 = 22
40 – 50 67 – 46 = 21

50 – 60	86 – 67 = 19
60 – 70	96 – 86 = 10
70 – 80	99 – 96 = 3
80 – 90	100 – 99 = 1

It is evident from the table that the distribution is irregular and maximum chances are that the distribution would be having more than one mode. You can verify by applying the grouping and analysing table.

The formula to calculate the value of mode in cases of bio-modal distributions is :

Mode = 3 median – 2 mean.

Computation of Mean and Median:

Marks	Mid-value (X)	Frequency (f)	Cf	(dx)	fdx
0 – 10	5	4	4	– 4	– 16
10 – 20	15	2	6	– 3	– 6
20 – 30	25	18	24	– 2	– 36
30 – 40	35	22	46	– 1	– 22
40—50	45	21	67	0	0
50—60	55	19	86	1	19
60—70	65	10	96	2	20
70—80	75	3	99	3	9
80—90	85	1	100	4	4
		$\sum f = 100$			$\sum fdx = -28$

Mean =

Median = size of item = = 50th item

Because 50 is smaller to 67 in C.f. column. Median class is 40 – 50

Median =

Median =
Apply, Mode = 3 median – 2 mean

Mode = 3 × 41.9 – 2 × 42.2 = 125.7 – 84.6 = 41.3
Example : Median and mode of the wage distribution are known to be Rs. 33.5 and 34 respectively. Find the missing values.

Wages (Rs.)	No. of Workers
0 – 10	4
10 – 20	16
20 – 30	?
30 – 40	?
40 – 50	?
50 – 60	6
60 – 70	4
Total = 230	

Solution : We assume the missing frequencies as 20 – 30 as x, 30 – 40 as y, and 40 – 50 as 230 – (4 + 16 + x + y + 6 + 4) = 200 – x – y.
We now proceed further to compute missing frequencies:

Wages (Rs.), X	No. of workers f	Cumulative frequencies C.f.
0 – 10	4	4
10 – 20	16	20
20 – 30	x	20 + x
30 – 40	y	20 + x + y

40 – 50	200 – x – y	220
50 – 60	6	226
60 – 70	4	230

N = 230

Apply, Median =
33.5 =
y(33.5 – 30) = (115 – 20 – x)10
3.5y = 1150 – 200 – 10x
10x + 3.5y = 950 ...(i)
Apply, Mode =
34 =
4(3y – 200) = 10(y – x)
10x + 2y = 800 ...(ii)
Subtract equation (ii) from equation (i),
1.5y = 150, y =
Substitute the value of y = 100 in equation (i), we get

10x + 3.5(100) = 950
10x = 950 – 350
x = 600/10 = 60
\Third missing frequency = 200 – x – y = 200 – 60 – 100 = 40.

LESSON 4
MEASURES OF DISPERSION

Why dispersion?
Measures of central tendency, Mean, Median, Mode, etc., indicate the central position of a series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be explained by certain other measures, known as 'Measures of Dispersion' or Variation.
We can understand variation with the help of the following example :

Series 1	Series 11	Series III
10	2	10
10	8	12
10	20	8
$\sum X = 30$	30	30

In all three series, the value of arithmetic mean is 10. On the basis of this average, we can say that the series are alike. If we carefully examine the composition of three series, we find the following differences:
(i) In case of 1st series, three items are equal; but in 2nd and 3rd series, the items are unequal and do not follow any specific order.
(ii) The magnitude of deviation, item-wise, is different for the 1st, 2nd and 3rd series. But all these deviations cannot be ascertained if the value of simple mean is taken into consideration.
(iii) In these three series, it is quite possible that the value of arithmetic mean is 10; but the value of median may differ from each other. This can be understood as follows ;

I	II	III
10	2	8
10 Median	8 Median	10 Median
10	20	12

The value of Median' in 1st series is 10, in 2nd series = 8 and in 3rd series = 10. Therefore, the value of the Mean and Median are not identical.
(iv) Even though the average remains the same, the nature and extent of the distribution of the size of the items may vary. In other words, the structure of the frequency distributions may differ even (though their means are identical.

What is Dispersion?
Simplest meaning that can be attached to the word 'dispersion' is a lack of uniformity in the sizes or quantities of the items of a group or series. According to Reiglemen, "Dispersion is the extent to which the magnitudes or quantities of the items differ, the degree of diversity." The word dispersion may also be used to indicate the spread of the data.
In all these definitions, we can find the basic property of dispersion as a value that indicates the extent to which all other values are dispersed about the central value in a particular distribution.

Properties of a good measure of Dispersion
There are certain pre-requisites for a good measure of dispersion:
1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each individual item of the distribution.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should not be unduly affected by the extreme items.

Types of Dispersion
The measures of dispersion can be either 'absolute' or "relative". Absolute measures of dispersion are expressed in the same units in which the original data are expressed. For example, if the series is expressed as Marks of the students in a

particular subject; the absolute dispersion will provide the value in Marks. The only difficulty is that if two or more series are expressed in different units, the series cannot be compared on the basis of dispersion.
'Relative' or 'Coefficient' of dispersion is the ratio or the percentage of a measure of absolute dispersion to an appropriate average. The basic advantage of this measure is that two or more series can be compared with each other despite the fact they are expressed in different units. Theoretically, 'Absolute measure' of dispersion is better. But from a practical point of view, relative or coefficient of dispersion is considered better as it is used to make comparison between series.

Methods of Dispersion

Methods of studying dispersion are divided into two types :
(i) Mathematical Methods: We can study the 'degree' and 'extent' of variation by these methods. In this category, commonly used measures of dispersion are :
(a) Range
(b) Quartile Deviation
(c) Average Deviation
(d) Standard deviation and coefficient of variation.
(ii) Graphic Methods: Where we want to study only the extent of variation, whether it is higher or lesser a Lorenz-curve is used.

Mathematical Methods

(a) Range

It is the simplest method of studying dispersion. Range is the difference between the smallest value and the largest value of a series. While computing range, we do not take into account frequencies of different groups.
Formula: Absolute Range = L – S
Coefficient of Range =
where, L represents largest value in a distribution
S represents smallest value in a distribution
We can understand the computation of range with the help of examples of different series,
(i) Raw Data: Marks out of 50 in a subject of 12 students, in a class are given as follows:
12, 18, 20, 12, 16, 14, 30, 32, 28, 12, 12 and 35.

In the example, the maximum or the highest marks obtained by a candidate is '35' and the lowest marks obtained by a candidate is '12'. Therefore, we can calculate range;

L = 35 and S = 12

Absolute Range = L – S = 35 – 12 = 23 marks

Coefficient of Range =

(ii) Discrete Series

Marks of the Students in Statistics (out of 50) (X)	No. of students (f)
Smallest 10	4
12	10
18	16
Largest 20	15
	Total = 45

Absolute Range = 20 – 10 = 10 marks

Coefficient of Range =

(iii) Continuous Series

X	Frequencies
10 – 15	4
S = 10 15 – 20	10

L = 30 20 – 25 26
25 – 30 8

Absolute Range = L – S = 30 – 10 = 20 marks
Coefficient of Range =
Range is a simplest method of studying dispersion. It takes lesser time to compute the 'absolute' and 'relative' range. Range does not take into account all the values of a series, i.e. it considers only the extreme items and middle items are not given any importance. Therefore, Range cannot tell us anything about the character of the distribution. Range cannot be computed in the case of "open ends' distribution i.e., a distribution where the lower limit of the first group and upper limit of the higher group is not given.
The concept of range is useful in the field of quality control and to study the variations in the prices of the shares etc.

(b) Quartile Deviations (Q.D.)
The concept of 'Quartile Deviation does take into account only the values of the 'Upper quartile (Q3) and the 'Lower quartile' (Q1). Quartile Deviation is also called 'inter-quartile range'. It is a better method when we are interested in knowing the range within which certain proportion of the items fall.
'Quartile Deviation' can be obtained as :
(i) Inter-quartile range = Q3 – Q1
(ii) Semi-quartile range =
(iii) Coefficient of Quartile Deviation =

Calculation of Inter-quartile Range, semi-quartile Range and Coefficient of Quartile Deviation in case of Raw Data
Suppose the values of X are : 20, 12, 18, 25, 32, 10
In case of quartile-deviation, it is necessary to calculate the values of Q1 and Q3 by arranging the given data in ascending of descending order.
Therefore, the arranged data are (in ascending order):
X = 10, 12, 18, 20, 25, 32
No. of items = 6
Q1 = the value of item = = 1.75th item

= the value of 1st item + 0.75 (value of 2nd item – value of 1st item)
= 10 + 0.75 (12 – 10) = 10 + 0.75(2) = 10 + 1.50 = 11.50
Q3 = the value of item =
= the value of 3(7/4)th item = the value of 5.25th item
= 25 + 0.25 (32 – 25) = 25 + 0.25 (7) = 26.075

Therefore,
(i) Inter-quartile range = Q3 – Q1 = 26.75 – 11.50 = 15.25
(ii) Semi-quartile range =
(iii) Coefficient of Quartile Deviation =

Calculation of Inter-quartile Range, semi-quartile Range and Coefficient of Quartile Deviation in discrete series
Suppose a series consists of the salaries (Rs.) and number of the workers in a factory:

Salaries (Rs.)	No. of workers
60	4
100	20
120	21
140	16
160	9

In the problem, we will first compute the values of Q3 and Q1

Salaries (Rs.)	No. of workers	Cumulative frequencies
(x)	(f)	(c.f.)

60	4	4
100	20	24 – Q1 lies in this cumulative
120	21	45 frequency
140	16	61 – Q3 lies in this cumulative
160	9	70 frequency

N = Sf = 70

Calculation of Q1 : Calculation of Q3 :

Q1 = size of th item Q3 = size of th item

= size of th item = 17.75 = size of th item = 53.25th item

17.75 lies in the cumulative frequency 24, 53.25 lies in the cumulative frequency 61 which

which is corresponding to the value Rs. 100 is corresponding to Rs. 140

Q1 = Rs. 100 Q3 = Rs. 140

(i) Inter-quartie range = Q3 – Q1 = Rs. 140 – Rs. 100 = Rs. 40

(ii) Semi-quartie range =

(iii) Coefficient of Quartile Deviation =

Calculation of Inter-quartile range, semi-quartile range and Coefficient of Quartile Deviation in case of continuous series

We are given the following data :

Salaries (Rs.)	No. of Workers
10 – 20	4
20 – 30	6
30 – 10	10
40 – 50	5

Total = 25
In this example, the values of Q3 and Q1 are obtained as follows:

Salaries (Rs.) No. of workers Cumulative frequencies
(x) (f) (c.f.)
10 – 20 4 4
20 – 30 6 10
30 – 40 10 20
40 – 50 5 25
N = 25

Q1 =
Therefore, . It lies in the cumulative frequency 10, which is corresponding to class 20 – 30.
Therefore, Q1 group is 20 – 30.
where, l1 = 20, f = 6, i = 10, and cfo = 4
Q1 =
Q3 =
Therefore, = 18.75, which lies in the cumulative frequency 20, which is corresponding to class 30 –40, Therefore Q3 group is 30 – 40.
where, l1 = 30, i = 10, cf0 = 10, and f = 10
Q3 = = Rs. 38.75
Therefore :
(i) Inter-quartile range = Q3 – Ql = Rs. 38.75 – Rs. 23.75 = Rs.15.00
(iii) Semi-quartile range =
(iii) Coefficient of Quartile Deviation =

Advantages of Quartile Deviation
Some of the important advantages are :
(i) It is easy to calculate. We are required simply to find the values of Q1 and Q3 and then apply the formula of absolute and coefficient of quartic deviation.
(ii) It has better results than range method. While calculating range, we consider only the extreme values that make dispersion erratic, in the case of quartile deviation, we take into account middle 50% items.

(iii) The quartile deviation is not affected by the extreme items.

Disadvantages

(i) It is completely dependent on the central items. If these values are irregular and abnormal the result is bound to be affected.

(ii) All the items of the frequency distribution are not given equal importance in finding the values of Q1 and Q3.

(iii) Because it does not take into account all the items of the series, considered to be inaccurate.

Similarly, sometimes we calculate percentile range, say, 90th and 10th percentile as it gives slightly better measure of dispersion in certain cases.

(i) Absolute percentile range = P90 – P10.

(ii) Coefficient of percentile range =

This method of calculating dispersion can be applied generally in case of open end series where the importance of extreme values are not considered.

(c) Average Deviation

Average deviation is defined as a value which is obtained by taking the average of the deviations of various items from a measure of central tendency Mean or Median or Mode, ignoring negative signs. Generally, the measure of central tendency from which the deviations arc taken, is specified in the problem. If nothing is mentioned regarding the measure of central tendency specified than deviations are

taken from median because the sum of the deviations (after ignoring negative signs) is minimum.

Computation in case of raw data

(i) Absolute Average Deviation about Mean or Median or Mode=

where: N = Number of observations,

|d| = deviations taken from Mean or Median or Mode ignoring signs.

(ii) Coefficient of A.D. =

Steps to Compute Average Deviation :

(i) Calculate the value of Mean or Median or Mode

(ii) Take deviations from the given measure of central-tendency and they are shown as d.

(iii) Ignore the negative signs of the deviation that can be shown as \d\ and add them to find S|d|.

(iv) Apply the formula to get Average Deviation about Mean or Median or Mode.
Example : Suppose the values are 5, 5, 10, 15, 20. We want to calculate Average Deviation and Coefficient of Average Deviation about Mean or Median or Mode.
Solution : Average Deviation about mean (Absolute and Coefficient).

(x)	Deviation from mean d	Deviations after ignoring signs | d |
5	– 6	6 =
5	– 6	6 where N = 5. SX = 55
10	+ 1	1
15	+ 4	4
20	+ 9	9
∑X = 55		∑| d | = 26

Average Deviation about Mean =
Coefficient of Average Deviation about mean =
Average Deviation (Absolute and Coefficient) about Median

Average Deviation about Mode =
Coefficient of Average Deviation about Mode =
Average deviation in case of discrete and continuous series
Average Deviation about Mean or Median or Mode =
where N = No. of items
|d| = deviations from Mean or Median or Mode after ignoring signs.
Coefficient of A.D. about Mean or Median or Mode =
Example: Suppose we want to calculate coefficient of Average Deviation about Mean from the following discrete series:

X	Frequency
10	5
15	10
20	15
25	10
30	5

Solution: First of all, we shall calculate the value of arithmetic Mean,
Calculation of Arithmetic Mean

Coefficient of Average Deviation about Mean =
Average Deviation about Mean =
In case we want to calculate coefficient of Average Deviation about Median from the following data:

Class Interval	Frequency
10 – 14	5
15 – 19	10
20 – 24	15
25 – 29	10
30 – 34	5

First of all we shall calculate the value of Median but it is necessary to find the 'real limits' of the given class-intervals. This is possible by subtracting 0.5 from all the lower-limits and add 0.5 to all the upper limits of the given classes. Hence, the real limits shall be : 9.5 – 14.5, 14.5 – 19.5, 19.5 – 24.5, 24.5 – 29.5 and 29.5 – 34.5
Calculation of Median

Advantages of Average Deviations

1. Average deviation takes into account all the items of a series and hence, it provides sufficiently representative results.
2. It simplifies calculations since all signs of the deviations are taken as positive.

3. Average Deviation may be calculated either by taking deviations from Mean or Median or Mode.
4. Average Deviation is not affected by extreme items.

5. It is easy to calculate and understand.
6. Average deviation is used to make healthy comparisons.

Disadvantages of Average Deviations

1. It is illogical and mathematically unsound to assume all negative signs as positive signs.
2. Because the method is not mathematically sound, the results obtained by this method are not reliable.
3. This method is unsuitable for making comparisons either of the series or structure of the series.

This method is more effective during the reports presented to the general public or to groups who are not familiar with statistical methods.

(d) Standard Deviation

The standard deviation, which is shown by greek letter s (read as sigma) is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson has a practical significance because it is free from all defects, which exists in a range, quartile deviation or average deviation.

Standard deviation is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation. The square of standard deviation i.e., s2 is called 'variance'.

Calculation of standard deviation in case of raw data

There are four ways of calculating standard deviation for raw data:

(i) When actual values are considered;
(ii) When deviations are taken from actual mean;
(iii) When deviations are taken from assumed mean; and
(iv) When 'step deviations' are taken from assumed mean.

(i) When the actual values are considered:

σ = where, N = Number of the items,

or σ2 = X = Given values of the series,

= Arithmetic mean of the series

We can also write the formula as follows :

σ = where, =

Steps to calculate σ
(i) Compute simple mean of the given values,
(ii) Square the given values and aggregate them
(iii) Apply the formula to find the value of standard deviation
Example: Suppose the values are given 2, 4, 6, 8, 10. We want to apply the formula

σ =
Solution: We are required to calculate the values of N, , SX2. They are calculated as follows :

X	X2
2	4
4	16
6	36
8	64
10	100

N = 5 ∑X2 = 220

σ =
Variance (σ)2 =
=

(ii) When the deviations are taken from actual mean
σ = where, N = no. of items and x = (X –)
Steps to Calculate σ
(i) Compute the deviations of given values from actual mean i.e., (X –) and represent them by x.
(ii) Square these deviations and aggegate them
(iii) Use the formula, σ =
Example : We are given values as 2, 4, 6, 8, 10. We want to find out standard deviation.

X	(X –) = x	x2
2	2 – 6 = – 4	(– 4)2 = 16
4	4 – 6 = – 2	(– 2)2 = 4
6	6 –6 = 0	= 0
8	8 – 6 = + 2	(2)2 = 4

10	10 – 6 = + 4	(4)2 = 16
N = 5		∑x2 = 40

(iii) When the deviations are taken from assumed mean

$\sigma =$

where, N = no. of items,

dx = deviations from assumed mean i.e., (X – A).

A = assumed mean

Steps to Calculate :

(i) We consider any value as assumed mean. The value may be given in the series or may not be given in the series.

(ii) We take deviations from the assumed value i.e., (X – A), to obtain dx for the series and aggregate them to find ∑dx.

(iii) We square these deviations to obtain dx2 and aggregate them to find ∑dx2.

(iv) Apply the formula given above to find standard deviation.

Example : Suppose the values are given as 2, 4, 6, 8 and 10. We can obtain the standard deviation as:

X	dx = (X – A)	dx2
2	– 2 = (2 – 4)	4
assumed mean (A) 4	0 = (4 – 4)	0
6	+ 2 = (6 – 4)	4
8	+ 4 = (8 – 4)	16
10	+ 6 = (10 – 4)	36
N = 5	∑dx = 10	∑dx2 = 60

(iv) When step deviations are taken from assumed mean

$\sigma =$

where, i = common factor, N = number of item, dx (Step-deviations) =

Steps to Calculate :
(i) We consider any value as assumed mean from the given values or from outside.
(ii) We take deviation from the assumed mean i.e. (X – A).
(iii) We divide the deviations obtained in step (ii) with a common factor to find step deviations and represent them as dx and aggregate them to obtain ∑dx.
(iv) We square the step deviations to obtain dx2 and aggregate them to find ∑dx2.
Example : We continue with the same example to understand the computation of Standard Deviation.

X d = (X – A) dx = and i = 2 dx2
2 – 2 1 1
A = 4 0 0 0
6 + 2 1 1
8 + 4 2 4
10 + 6 3 9

N = 5 Sdx = 5 Sdx2 = 15

s = where N = 5, i = 2, dx = 5, and Sdx2 = 15
s =
Note :We can notice an important point that the standard deviation value is identical by four methods. Therefore any of the four formulae can be applied to find the value of standard deviation. But the suitability of a formula depends on the magnitude of items in a question.
Coefficient of Standard-deviation =
In the above given example, s = 2.828 and = 6
Therefore, coefficient of standard deviation =

Coefficient of Variation or C. V.
=
Generally, coefficient of variation is used to compare two or more series. If coefficient of variation (C.V.) is more for one series as compared to the other, there will be more variations in that series, lesser stability or consistency in its composition. If coefficient of variation is lesser as compared to other series, it will be more stable or consistent. Moreover that series is always better where coefficient of variation or coefficient of standard deviation is lesser.

Example : Suppose we want to compare two firms where the salaries of the employees are given as follows:

	Firm A	FirmB
No. of workers	100	100
Mean salary (Rs.)	100	80
Standard-deviation (Rs.)	40	45

Solution : We can compare these firms either with the help of coefficient of standard deviation or coefficient of variation. If we use coefficient of variation, then we shall apply the formula :

Firm A
C.V. =
= 100, σ = 40.

Firm B
C.V. =
= 80, σ = 45

Because the coefficient of variation is lesser for firm A than firm B, therefore, firm A is less variable and more stable.

Calculation of standard-deviation in discrete and continuous series

We use the same formula for calculating standard deviation for a discrete series and a continuous series. The only difference is that in a discrete series, values and frequencies are given whereas in a continuous series, class-intervals and frequencies are given. When the mid-points of these class-intervals are obtained, a continuous series takes shape of a discrete series. X denotes values in a discrete series and mid points in a continuous series.

When the deviations are taken from actual mean

We use the same formula for calculating standard deviation for a continuous series

σ =

where N = Number of items

f = Frequencies corresponding to different values or class-intervals.

x = Deviations from actual mean (X –).

X = Values in a discrete series and mid-points in a continuous series.

Step to calculate σ

(i) Compute the arithmetic mean by applying the required formula.

(ii) Take deviations from the arithmetic mean and represent these deviations by x.

(iii) Square the deviations to obtain values of x .
(iv) Multiply the frequencies of different class-intervals with x2 to find fx2.
Aggregate fx2 column to obtain ∑ fx2.
(v) Apply the formula to obtain the value of standard deviation.
If we want to calculate variance then we can compute σ2 =
Example : We can understand the procedure by taking an example :

σ = where, N = 45, ∑fx2 = 1500
σ =
When the deviations are taken from assumed mean
In some cases, the value of simple mean may be in fractions, them it becomes time consuming to
take deviations and square them. Alternatively, we can take deviations from the assumed mean.
σ =
where N = Number of the items,
dx = deviations from assumed mean (X – A),
f = frequencies of the different groups,
A = assumed mean and
X = Values or mid points.
Step to calculate σ
(i) Take the assumed mean from the given values or mid points.
(ii) Take deviations from the assumed mean and represent them by dx.
(iii) Square the deviations to get dx2 .
(iv) Multiply f with dx of different groups to abtain fdx and add them up to get ∑fdx.
(v) Multiply f with dx2 of different groups to abtain fdx2 and add them up to get ∑fdx2.
(vi) Apply the formula to get the value of standard deviation.

Steps to calculate σ
(i) Take deviations from the assumed mean of the calculated mid-points and divide all deviations by a common factor (i) and represent these values by dx.
(ii) Square these step deviations dx to obtain dx2 for different groups.

(iii) Multiply f with dx of different groups to find fdx and add them to obtain fdx .
(iv) Multiply f with dx2 of different groups to find fdx2 for different groups and add them to obtain ∑fdx2.
(v) Apply the formula to find standard deviation.

Advantages of Standard Deviation

(i) Standard deviation is the best measure of dispersion because it takes into account all the items and is capable of future algebric treatment and statistical analysis.
(ii) It is possible to calculate standard deviation for two or more series.
(iii) This measure is most suitable for making comparisons among two or more series about varibility.

Disadvantages

(i) It is difficult to compute.
(ii) It assigns more weights to extreme items and less weights to items that are nearer to mean. It is because of this fact that the squares of the deviations which are large in size would be proportionately greater than the squares of those deviations which are comparatively small.

Mathematical properties of standard deviation (σ)
(i) If deviations of given items are taken from arithmetic mean and squared then the sum of squared deviation should be minimum, i.e., = Minimum,
(ii) If different values are increased or decreased by a constant, the standard deviation will remain the same. If different values arc multiplied or divided by a constant than the standard deviation will be multiplied or divided by that constant.
(iii) Combined standard deviation can be obtained for two or more series with below given formula:
σ12 =
where:

N1 represents number of items in first series,
N2 represents number of items in second series,
represents variance of first series,
represents variance of second series,
d1 represents the difference between
d2 represents the difference between

represents arithmetic mean of first series,
represents arithmetic mean of second series,
represents combined arithmetic mean of both the series.

Example : Find the combined standard deviation of two series, from the below given information :

	First Series	Second Series
No. of items	10	15
Arithmetic means	15	20
Standard deviation	4	5

Solution : Since we are considering two series, therefore combined standard deviation is computed by the following formula :

σ12 =
where: N1 = 10, N2 = 15, , , σ1 = 4, s2 = 5
=
or =
d1 =
By applying the formula of combined standard deviation, we get :
σ12 =
=
=
(iv) Standard deviation of n natural numbers can he computed as :
σ = where, N represents number of items.

(v) For a symmetrical distribution
+ σ covers 68.27% of items.
+ 2σ covers 95.45% of items.
+ 3σ covers 99.73% of items.

Example : You are heading a rationing department in a State affected by food shortage. Local investigators submit the following report:

Daily calorie value of food available per adult during current period :

Area	Mean	Standard deviation

A	2,500	400
B	2,000	200

The estimated requirement of an adult is taken at 2,800 calories daily and the absolute minimum is 1,350. Comment on the reported figures, and determine which area, in your opinion, need more urgent attention.
Solution : We know that + σ covers 68.27% of items. + 2σ covers 95.45% of items and + 3σ covers 99.73% . In the gjven problem if we take into consideration 99.73%. i.e., almost the whole population, the limits would be + 3σ.

For Area A these limits are :
+ 3σ = 2,500 + (3 × 400) = 3,700
– 3σ = 2,500 – (3 × 400) = 1,300
For Area B these limits are :
+ 3σ = 2,000 + (3 × 200) = 2,600
– 3σ = 2,000 – (3 × 200) = 1,400

It is clear from above limits that in Area A there are some persons who are getting 1300 calories, i.e. below the minimum which is 1,350. But in case of area B there is no one who is getting less than the minimum. Hence area A needs more urgent attention.

(vi) Relationship between quartile deviation, average deviation and standard deviation is given as:
Quartile deviation = 2/3 Standard deviation
Average deviation = 4/5 Standard deviation
(vii) We can also compute corrected standard deviation by using the following formula :
Correct σ =
(a) Compute corrected =
where, corrected ∑f = ∑X + correct items – wrong items
where, ∑X = N.
(b) Compute corrected ∑X2 =∑ X2 + (Each correct item)2 – (Each wrong item}2
where, ∑X2 = Nσ1 +

Example : (a) Find out the coefficient of variation of a series for which the following results are given :

N = 50, ∑X’ = 25, ∑X’2 = 500 where: X’ = deviation from the assumed average 5.
(b) For a frequency distribution of marks in statistics of 100 candidates, (grouped in class inervals of 0 – 10, 10 – 20) the mean and standard deviation were found to be 45 and 20. Later it was discovered that the score 54 was misread as 64 in obtaining frequency distribution. Find out the correct mean and correct standard deviation of the frequency destribution.
(c) Can coefficient of variation be greater than 100%? If so, when?
Solution : (a) We want to calculate, coefficient of variation which is =

Therefore, we are required to calculate mean and standard deviation.
Calculation of simple mean

= where, A = 5, N = 50, ∑X’ = 25

Calculation of standard deviation
σ =
Calculation of Coefficient of variation
C.V. =

(b) Given = 45, σ = 20, N = 100, wrong value = 64, correct value = 54
Since this is a case of continuous series, therefore, we will apply the formula for mean and standard deviation that are applicable in a continuous series.

Calculation of correct Mean
= or N = ∑fX
By substituting the values, we get 100 × 45 = 4500
Correct ∑fX = 4500 – 64 + 54 = 4490
Correct =

Calculation of correct σ
σ = or σ2 =
where, σ = 20, N = 100, = 45
(20)2 =
or 400 =
or 400 + 2025 =
or 2425 × 100 = ∑fX2 = 242500
\ Correct ∑fX2 = 242500 – (64)2 + (54)2 = 242500 – 4096 + 2916 = 242500 –

1180 = 241320
Correct σ =

(c) The formulae for the computation of coefficient of variation is =
Hence, coefficient of variation can be greater than 100% only when the value of standard deviation is greater than the value of mean.
This will happen when data contains a large number of small items and few items are quite large. In such a case the value of simple mean will be pulled down and the value of standard deviation will go up. Similarly, if there arc negative items in a series, the value of mean will come down and the value of standard deviation shall not be affecied because of squaring the deviations.
Example : In a distribution of 10 observations, the value of mean and standard deviation are given as 20 and 8. By mistake, two values are taken as 2 and 6 instead of 4 and 8. Find out the value of correct mean and variance.
Solution : We are given: N – 10, = 20, σ = 3
Wrong values = 2 and 6 and Correct values = 4 and 8
Calculation of correct Mean
=
∑X = 10 × 20 = 200
But ∑X is incorrect. Therefore we shall find correct ∑X.
Correct ∑X = 200 – 2 – 6 + 4 + 8 = 204
Correct Mean =
Calculation of correct variance
σ =
or σ2 =
or (8)2 =
or 64 + 400 =
or ∑X2 = 4640
But this is wrong and hence we shall compute correct ∑X2
Correct ∑X2 = 4640 – 22 – 62 + 42 + 82
= 4640 – 4 – 36 + 16 + 64 = 4680
Correct σ2 =

Revisionary Problems
Example : Compute (a) Inter-quartile range. (b) Semi-quartile range, and (c) Coefficient of quartile deviation from the following data :

Farm Size (acres)	No. of firms	Farm Size (acres)	No. of firms
0 – 40	394	161 – 200	169
41 – 80	461	201 – 240	113
81 – 120	391	24 1 and over	148
121 – 160	334		

Solution :
In this case, the real limits of the class intervals are obtained by subtracting 0.5 from the lower limits of each class and adding 0.5 to the upper limits of each class. This adjustment is necessary to calculate median and quartiles of the series.

Farm Size (acres)	No. of firms	Cumulative frequency (c.f.)
– 0.5 – 10.5	394	394
40.5 – 80.5	461	855
80.5 – 120.5	391	1246
120.5 – 160.5	334	1580
160.5 – 200.5	169	1749
200.5 – 240.5	113	1862
240.5 and over	148	2010

N = 2010

Q1 =
=
Q1 lies in the cumulative frequency of the group 40.5 – 80.5. and l1 = 40.5, f =

461, i = 40, cf0 = 394, = 502.5
Q1 =
Similarly, Q3 =
=
Q3 lies in the cumulative frequency of the group 121 – 160, where the real limits of the class interval are 120.5 – 160.5 and l1 = 120.5, i = 40, f = 334, = 1507.5, c.f. = 1246
Q3 =
Inter-quartile range = Q3 – Q1 = 151.8 – 49.9 = 101.9 acres
Semi-quartile range =
Coefficient of quartile deviation =

Example : Calculate mean and coefficient of mean deviation about mean from the following data :

Marks less than	No. of students
10	4
20	10
30	20
40	40
50	50
60	56
70	60

Solution :
In this question, we are given less than type series alongwith the cumulative frequencies. Therefore, we are required first of all to find out class intervals and frequencies for calculating mean and coefficient of mean deviation about mean.

M.D. about mean =
Coefficient of M.D. about mean =
Example : Calculate standard deviation from the following data :

Class Interval	frequency
– 30 to – 20	5
– 20 to – 10	10
– 10 to – 0	15
0 to 10	10
10 to 20	5
	N = 45

Example : For two firms A and B belonging to same industry, the following details are available :

	Firm A	Firm B
Number of Employees :	100	200
Average wage per month :	Rs. 240	Rs. 170
Standard deviation of the wage per month :	Rs. 6	Rs. 8

Find (i) Which firm pays out larger amount as monthly wages?
(ii) Which firm shows greater variability in the distribution of wages?
(iii) Find average monthly wages and the standard deviation of wages of all employees for both the firms.

Solution : (i) For finding out which firm pays larger amount, we have to find out $\sum X$.

$X =$ or $\sum X = NX$

Firm A : $N = 100, X = 240$, $\sum X = 100 \times 240 = 24000$
Firm B : $N = 200, X = 170$, $\sum X = 200 \times 170 = 34000$
Hence firm B pays larger amount as monthly wages.

(ii) For finding out which firm shows greater variability in the distribution of wages, we have to calculate coefficient of variation.
Firm A : C.V. =

Firm B : C.V. =
Since coefficient of variation is greater for firm B. hence it shows greater variability in the distribution of wages.
(iii) Combined wages : =
where, N1 = 100, = 240, N2 = 200, = 170
Hence =

Combined Standard Deviation :
σ12 =
where N1 = 100, N2 = 200, σ1 = 6, σ2 = 8, = 240 – 193.3 = 46.7
and d1 = = 170 – 193.3 = – 23.3
σ12 =
=

Example : From the following frequency distribution of heights of 360 boys in the age-group 10 – 20 years calculate the :
(i) arithmetic mean;
(ii) coefficient of variation; and
(iii) quartile deviation

Height (cms)	No. of boys	Height (cms)	No. of boys
126 – 130	31	146 – 150	60
131 – 135	44	151 – 155	55
136 – 140	48	156 – 160	43
141 – 145	51	161 – 165	28

Solution : Calculation of , Q.D., and C.V.

Heights	m.p. X	f	(X – 143)/5 dx	fdx	fdx2	c.f.
126 – 130	128	31	– 3	– 93	279	31
131 – 135	133	44	– 2	– 88	176	75
136 – 140	138	48	– 1	– 48	48	123
141 – 145	143	51	0	0	0	174
146 – 150	148	60	+ 1	+ 60	60	234
151 – 155	153	55	+ 2	+ 10	220	289
156 – 160	158	43	+ 3	+ 129	387	332
161 – 165	163	28	+ 4	+ 112	448	360
	N = 45			∑fdx = 182	∑fdx2 = 1618	

(i) = where, N = 360, A = 143, i = 5, ∑fdx = 182

=

(ii) C.V. =

σ =

=

C.V. =

(iii) Q.D. =

Q1 = Size of th observation = observation

Q1 lies in the class 136 – 140. But the real limits of this class is 135.5 – 140.5

Q1 =

Q3 = Size of observition = observation

Q3 lies in the class 151 – 155. But the real limit of this class is 150 – 155.5

Q3 =

Q.D. =

SIMPLE CORRELATION

In the earlier chapters we have discussed univariate distributions to highlight the important characteristics by different statistical techniques. Univariate distribution means the study related to one variable only we may however come across certain series where each item of the series may assume the values of two or more variables. The distributions in which each unit of series assumes two values is called bivariate distribution. In a bivariate distribution, we are interested to find out whether there is any relationship between two variables. The correlation is a statistical technique which studies the relationship between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between the two variables. When two variables are related in such a way that a change in the value of one is accompanied either by a direct change or by an inverse change in the values of the other, the two variables are said to be correlated. In the correlated variables an increase in one variable is accompanied by an increase or decrease in the other variable. For instance, relationship exists between the price and demand of a commodity because keeping other things equal, an increase in the price of a commodity shall cause a decrease in the demand for that commodity. Relationship might exist between the heights and weights of the students and between amount of rainfall in a city and the sales of raincoats in that city.
These are some of the important definitions about correlation. Croxton and Cowden says, "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation".
A.M. Tuttle says, "Correlation is an analysis of the co variation between two or more variables." W.A. Neiswanger says, "Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective." L.R. Conner says, "If two or more quantities vary in sympathy so that the movements in one tends to be accompanied by corresponding movements in others than they are said be correlated."

Utility of Correlation

The study of correlation is very useful in practical life as revealed by these points.
1. With the help of correlation analysis, we can measure in one figure, the degree of relationship existing between variables like price, demand, supply, income, expenditure etc. Once we know that two variables are correlated then we can easily

estimate the value of one variable, given the value of other.
2. Correlation analysis is of great use to economists and businessmen, it reveals to the economists the disturbing factors and suggest to him the stabilizing forces. In business, it enables the executive to estimate costs, sales etc. and plan accordingly.
3. Correlation analysis is helpful to scientists. Nature has been found to be a multiplicity of interrelatedforces.

Difference between Correlation and Causation

The term correlation should not be misunderstood as causation. If correlation exists between two variables, it must not be assumed that a change in one variable is the cause of a change in other variable. In simple words, a change in one variable may be associated with a change in another variable but this change need not necessarily be the cause of a change in the other variable. When there is no cause and effect relationship between two variables but a correlation is found between the two variables such correlation is known as "spurious correlation" or "nonsense correlation". Correlation may exist due to the following:
1. Pure change correlation : This happens in a small sample. Correlation may exist between incomes and weights of four persons although there may be no cause and effect relationship between incomes and weights of people. This type of correlation may arise due to pure random sampling variation or because of the bias of investigator in selecting the sample.

2. When the correlated variables are influenced by one or more variables. A high degree of correlation between the variables may exist, where the same cause is affecting each variable or different cause affecting each with the same effect. For instance, a degree of correlation may be found between yield per acre of rice and tea due to the fact that both are related to the amount of rainfall but none of the two variables is the cause of other.

3. When the variable mutually influence each other so that neither can be called the cause of other. All times it may be difficult to say that which of the two variables is the cause and which is the effect because both may be reacting on each other.

Types of Correlation

Correlation can be categorised as one of the following :

(i) Positive and Negative,
(ii) Simple and Multiple.
(iii) Partial and Total.

(iv) Linear and Non-Linear (Curvilinear)

(i) Positive and Negative Correlation : Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.

(ii) Simple and Multiple Correlation : Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlation, the relationship is studied among three or more variables. For example, the relationship of yield of wheat may be studied with both chemical fertilizers and the pesticides.

(ii) Partial and Total Correlation : There are two categories of multiple correlation analysis. Under partial correlation, the relationship of two or more variables is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant. For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. Total correlation is based upon all the variables.
(iv) Linear and Non-Linear Correlation : When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

Methods of Studying Correlation

There are different methods which helps us to find out whether the variables are related or not.

1. Scatter Diagram Method.

2. Graphic Method.

3. Karl Pearson’s Coefficient of correlation.

4. Rank Method.

We shall discuss these methods one by one.

(1) Scatter Diagram : Scatter diagram is drawn to visualise the relationship between two variables. The values of more important variable is plotted on the X-axis while the values of the variable are plotted on the Y-axis. On the graph, dots are plotted to represent different pairs of data. When dots are plotted to represent all the pairs, we get a scatter diagram. The way the dots scatter gives an indication of the kind of relationship which exists between the two variables. While drawing scatter diagram, it is not necessary to take at the point of sign the zero values of X and Y variables, but the minimum values of the variables considered may be taken.
When there is a positive correlation between the variables, the dots on the scatter diagram run from left hand bottom to the right hand upper corner. In case of perfect positive correlation all the dots will lie on a straight line.
When a negative correlation exists between the variables, dots on the scatter diagram run from the upper left hand corner to the bottom right hand corner. In case of perfect negative correlation, all the dots lie on a straight line.
If a scatter diagram is drawn and no path is formed, there is no correlation.
Students are advised to prepare two scatter diagrams on the basis of the following data :

(i) Data for the first Scatter Diagram :

Demand Schedule

Price (Rs.)	Commodity Demand (units)
6	180
7	150
8	130
9	120
10	125

(iii) Data for the second Scatter Diagram :
Supply Schedule

Price(Rs.)	**CommoditySupply**
50	2,000
51	2,100
52	2,200
53	2,500
54	3,000

55 3,800

56 4,700

Students will find that the first diagram indicate a negative correlation where the second diagram shall reveal a positive correlation.

(2) Graphic Method. In this method the individual values of the two variables are plotted on the graph paper. Therefore two curves are obtained-one for X variable and another for Y variable.

Interpreting Graph

The graph is interpreted as follows:

(i) If both the curves run parallel or nearly parallel or more in the same direction,

there is positive correlation,

(ii) On the other hand, if both the curves move in the opposite direction, there is a negative correlation.

Illustration : Show correlation from the following data by graphic method;

Year	1995	96	97	98	99	2000	2001	2002	2003	2004
Average Income (Rs.)	100	110	125	140	150	180	200	220	250	360
Average Expenditure (Rs.)	90	95	100	120	120	140	150	170	200	260

Solution :

The graph prepared shows that income and expenditure have a close positive correlation. As income increases, the expenditure also increases.

(3) Karl Pearson's Co-efficient of Correlation. Karl Pearson's method, popularly known as Pearson co-efficient of correlation, is most widely applied in practice to measure correlation. The Pearson co-efficient of correlation is represented by the symbol r.

According to Karl Pearson's method, co-efficient of correlation between the variables is obtained by dividing the sum of the products of the corresponding deviations of the various items of two series from their respective means by the product of their standard deviations and the number of pairs of observations. Symbolically, r = where r stands for coefficient of correlation ...(i) where x1, x2, x3, x4 ..... xn are the deviations of various items of the first variable from the mean,

y1, y2, y3,..... yn are the deviations of all items of the second variable from mean, Sxy is the sum of products of these corresponding deviations. N stands for the number of pairs, sx stands for the standard deviation of X variable and sy

stands for the standard deviation of Y variable. sx = and sy =
If we substitute the value of sx and sy in the above written formula of computing r, we get r = or r =
Degree of correlation varies between + 1 and –1; the result will be + 1 in case of perfect positive correlation and – 1 in case of perfect negative correlation. Computation of correlation coefficient can be simplified by dividing the given data by a common factor. In such a case, the final result is not multiplied by the common factor because coefficient of correlation is independent of change of scale and origin.

Illustration : Calculate Co-efficient of Correlation from the following data:

X	50	100	150	200	250	300	350
Y	10	20	30	40	50	60	70

Solution :

The value of r indicates that there exists a high degree positive correlation between lengths and weights.

Illustration : From the following data, compute the co-efficient of correlation between X and Y :

	X Series	**Y Series**
Number of items	15	15
Arithmetic Mean	25	18
Square of deviation from Mean	136	138

Summation of product deviations of X and Y from their Arithmetic Means = 122
Solution : Denoting deviations of X and Y from their arithmetic means by x and y respectively, the given data are : Sx2 = 136, Sxy = 122, and Sy2 = 138
r = =
Short-cut Method: To avoid difficult calculations due to mean being in fraction, deviations are taken from assumed means while calculating coefficient of correlation. The formula is also modified for standard deviations because deviations are taken from assumed means. Karl Perason's formula for short-cut method is given below :
r = or
r =

Illustration : Compute the coefficient of correlation from the following data :

Marks in Statistics	20	30	28	17	19	23	35	13	16	38

Marks in Mathematics 18 35 20 18 25 28 33 18 20 40

Solution :

Direct Method of Computing Correlation Coefficient

Correlation Coefficient can also be computed from given X and Y values by using the below given formula:

r =

The above given formula gives us the same answer as we are getting by taking durations from actual mean or arbitrary mean.

Illustration : Compute the coefficient of correlations from the following data :

Marks in Statistics	20	30	28	17	19	23	35	13	16	38
Marks in Mathematics	18	35	20	18	25	28	33	18	20	40

Solution :

Marks in Statistics X	**Marks in Mathematics Y**	**X_2**	**Y^2**	**XY**
20	18	400	324	360
30	35	900	1225	1050
28	20	784	400	560
17	18	289	324	306
19	25	361	625	475
23	28	529	784	644
35	33	1225	1089	1155
13	18	169	324	234
16	20	256	400	320
38	40	1444	1600	1520

SX = 239 SY = 255 SX^2 = 6357 SY^2 = 7095 SXY = 6624

Substitute the computed values in the below given formula, r =

=

=

Coefficient of Correlation in a Continuous Series

In the case of a continuous series, we assume that every item which falls within a given class interval falls exactly at the middle of that class. The formula, because of the presence of frequencies is modified as follows:
r =

Various values shall be calculated as follows :
(i) Take the step deviations of variable X and denote it as dx.
(ii) Take the step deviations of variable Y and denote it as dy.
(iii) Multiply dx dy and the respective frequency of each cell and write the figure obtained in the right-hand upper comer of each cell.
(iv) Add all the cornered values calculated in step (iii) to get Sfdxdy.
(v) Multiply the frequencies of the variable X by the deviations of X to get Sfdx.
(vi) Take the squares of the deviations of the variable X and multiply them by the respective frequencies to get Sfdx2.
(vii) Multiply the frequencies of the variable Y by the deviations of Y to get Sfdy.
(viii) Take the squares of the deviations of the variable Y and multiply them by the respective frequencies to get Sfdy2.
(ix) Now substitute the values of Sfdxdy, Sfdx, Sfdx2, Sfdy, Sfdy2 in the formula to get the value of r.

Properties of Coefficient of Correlation

Following are some of the important proportion of r :
(1) The coefficient of correlation lies between – 1 and + 1 (– 1 £ r £ + 1)
(2) The coefficient of correlation is independent of change of scale and origin of the variable X and Y.

(3) The coefficient of correlation is the geometric mean of two regression coefficients.
r =

Merits of Pearson's coefficient of correlation : The correlation of coefficient summarizes in one figure the degree and direction of correlation. Value varies between +1 and –1.

Demerits of Pearson's coefficient of correlation : It always assumes linear relationship between the variables; in fact the assumption may be wrong. Secondly, it is not easy to interpret the significance of correlation coefficient. The method is time consuming and affected by the extreme items.

Probable Error of the coefficient of correlation : It is calculated to find out how far the Pearson's coefficient of correlation is reliable in a particular case.

P.E of coefficient of correlation = where r = coefficient of correlation and N = number of pairs of items. If the probable error calculated is added to and subtracted from the coefficient of correlation, it would give us such limits within which we can expect the value of the coefficient of correlation to vary. If r is less than probable error, then there is no real evidence of correlation. If r is more than 6 times the probable error, the coefficient of correlation is considered highly significant. If r is more than 3 times the probable error but less than 6 times, correlation is considered significant but not highly significant. If the probable error is not much and the given r is more than the probable error but less then 3 times of it, nothing definite can be concluded.

(4) Rank Correlation : There are many problems of business and industry when it is not possible to measure the variable under consideration quantitatively or the statistical series is composed of items which can not be exactly measured. For instance, it may be possible for the two judges to rank six different brands of cigarettes in terms of taste, whereas it may be difficult to give them a numerical grade in terms of taste. In such problems. Spearman's coefficient of rank correlation is used. The formula for rank correlation is : r = or where r stands for rank coefficient of correlation. D refers to the difference of ranks between paired items.

N refers to the number of paired observations.

The value of rank correlation coefficient varies between +1 and –1. When the value of r = +1, there is complete agreement in the order of ranks and the ranks will be in the same order. When r = – l, the ranks will be in opposite direction showing complete disagreement in the order of ranks. Let' us understand with the help of an illustration.

Illustration : Ranks of 10 individuals at the start and at the finish of a course of training are given :

Individual :	A	B	C	D	E	F	Q	H	I	J
Rank before :	1	6	3	9	5	2	7	10	8	4
Rank after :	6	8	3	7	2	1	5	9	4	10

Calculate coefficient of correlation.

Solution :

Individual Rank before Rank after (R1 – R2)

	R1	R2	D	D2
A	1	6	5	25
B	6	8	2	4
C	3	3	0	0
D	9	7	2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	6	36

N = 10 SD2 = 100

By applying the formula,

r =

When we are given the actual data and not the ranks, it becomes necessary for us to assign the ranks. Ranks can be assigned by taking either the highest value as one or the lowest value as one. But if we start by taking the highest value or the lowest value we must follow the same order for both the variables to assign ranks.

Illustration : Calculate rank correlation from the following data :

X	:	17	13	15	16	6	11	14	9	7	12
Y	:	36	46	35	24	12	18	27	22	2	8

In some case it becomes necessary to rank two or more items an identical rank. In such cases, it is customary to give each item an average rank. Therefore, if two items are equal for 4th and 5th rank, each item shall be ranked 4.5 i.e., . It means, where two or more items are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of ranks which these items would have got had they differed slightly from each other. When equal ranks are assigned to some items, the rank correlation formula is also adjusted. The

adjustment consists of adding (m2 – m) to the value of SD2 where m stands for number of items whose ranks are identical.

r =

Let us take an example to understand this.

Illustration : Compute the rank correlation coefficient from the following data:

Section A	:	115	109	112	87	98	98	120	100	98	118
Section B	:	75	73	85	70	76	65	82	73	68	80

Item 98 is repeated three times in series A. Hence m = 3. In series B the item 73 is repeated two times and so m = 2.

r =

r =

REGRESSION ANALYSIS

The statistical technique correlation establishes the degree and direction of relationship between two or more variables. But we may be interested in estimating the value of an unknown variable on the basis of a known variable. If we know the index of money supply and price-level, we can find out the degree and direction of relationship between these indices with the help of correlation technique. But the regression technique helps us in determining what the general price-level would be assuming a fixed supply of money.

Similarly if we know that the price and demand of a commodity are correlated we can find out the demand for that commodity for a fixed price. Hence, the statistical tool with the help of which we can estimate orpredict the unknown variable from known variable is called regression. The meaning of the term “Regression”is the act of returning or going back. This term was first used by Sir Francis Galton in 1877 when he studied the relationship between the height of fathers and sons. His study revealed a very interesting relationship.

All tall fathers tend to have tall sons and all short fathers short sons but the average height of the sons of a group of tall fathers was less than that of the fathers and the average height of the sons of a group of short fathers was greater than that of the fathers. The line describing this tendency of going back is called“Regression Line”. Modern writers have started to use the term estimating line instead of

regression linebecause the expression estimating line is more clear in character. According to Morris Myers Blair, regressionis the measure of the average relationship between two or more variables in terms of the original units of the data.

Regression analysis is a branch of statistical theory which is widely used in all the scientific disciplines. It is a basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. The uses of regression analysis are not confined to economic and business activities. Its applications are extended to almost all the natural, physical and social sciences. The regression technique can be extended to three or more variables but we shall limit ourselves to problems having two variables in this lesson. Regression analysis is of great practical use even more than the correlation analysis. Some of the uses of the regression analysis are given below :

(i) Regression Analysis helps in establishing a functional relationship between two or more variables. Once this is established it can be used for various analytic purposes.

(ii) With the use of electronic machines and computers, the medium of calculation of regression equation particularly expressing multiple and non-linear relations has been reduced considerably.

(iii) The regression analysis is very useful for prediction purposes. Once a functional relationship is established the value of the dependent variable can be estimated from the given value of the independent variables.

Difference between Correlation and Regression

Both the techniques are directed towards a common purpose of establishing the degree and direction of relationship between two or more variables but the methods of doing so are different. The choice of one or the other will depend on the purpose. If the purpose is to know the degree and direction of relationship, correlation is an appropriate tool but if the purpose is to estimate a dependent variable with the substitution of one or more independent variables, the regression analysis shall be more helpful. The point of difference are discussed below :

(i) Degree and Nature of Relationship : The correlation coefficient is a measure of degree of covariability between two variables whereas regression analysis is used to study the nature of relationship between the variables so that we can predict the value of one on the basis of another. The reliance on the estimates or predictions depend upon the closeness of relationship between the variables.

(ii) Cause and Effect Relationship : The cause and effect relationship is

explained by regression analysis. Correlation is only a tool to ascertain the degree of relationship between two variables and we can not say that one variable is the cause and other the effect. A high degree of correlation between price and demand for a commodity or at a particular point of time may not suggest which is the cause and which is the effect. However, in regression analysis cause and effect relationship is clearly expressed— one variable is taken as dependent and the other an independent.

The variable which is the basis of prediction is called independent variable and the variable that is to be predicted is called dependent variable. The independent variable is represented by X and the dependent variable by Y.

Principle of Least Squares

Regression refers to an average of relationship between a dependent variable with one or more independent variables. Such relationship is generally expressed by a line of regression drawn by the method of the "Least Squares". This line of regression can be drawn graphically or derived algebraically with the help of regression equations. According to Tom Cars, before the equation of the least line can be determined some criterion must be established as to what conditions the best line should satisfy. The condition usually stipulated in regression analysis is that the sum of the squares of the deviations of the observed Y values from the fitted line shall be minimum. This is known as the least squares or minimum squared error criterion. A line fitted by the method of least squares is the line of best fit. The line satisfies the following conditions :

(i) The algebraic sum of deviations above the line and below the line are equal to zero.

$S(X - Xc) = 0$ and $S(Y - Yc) = 0$

Where .XC and YC are the values derived with the help of regression technique.

(ii) The sum of the squares of all these deviations is less than the sum of the squares of deviations from any other line, we can say

$S (X - Xc)2$ is smaller than $S (X - A)2$ and

$S (Y - Yc)2$ is smaller than $S (Y - A)2$

Where A is some other value or any other straight line.

(iii) The line of regression (best fit) intersect at the mean value of the variables i.e., and

(iv) When the data represent a sample from a larger population, the least square line is the best estimate of the population line.

Methods of Regression Analysis

We can study regression by the following methods :
1. Graphic method (regression lines)

2. Algebraic method (regression equations)

We shall discuss these methods in detail.
1. Graphic Method : When we apply this method different points are plotted on a graph paper representing different pairs of variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn between these points either by free hand or by a scale in such a way that the squares of the vertical or horizontal distances between the points and the line of regression is minimum. It should be drawn in such a manner that the line leaves equal number of points on both sides. However, to ensure this is rather difficult and the method only renders a rough estimate which can not be completely free from subjectivity of person drawing it. Such a line can be a straight line or a curved line depending upon the scatter of points and relationship to be established. A non-linear free hand curve will have more element of subjectivity and a straight line is generally drawn. Let us understand it with the help of an example:

Example :

Height of father (Inches)	**Height of sons**
65	68
63	66
67	68
64	65
68	69
62	66
70	68
66	65
68	71
67	67
69	68
71	70

Solution : The diagram given below shows the height of fathers on x-axis and the height of sons on y-axis. The line of regression called the regression of y on x is drawn between the scatter dots.

Fig. 1

Another line of regression called the regression line of x on y is drawn amongst the same set of scatter dots in such a way that the squares of the horizontal distances between dots are minimised.

Fig. 2 Fig. 3

It is clear that the position of the regression line of x on y is not exactly like that of the regression lime of y on x. In the following figure both the regression of y on x and x on y are exhibited.

Fig. 4

When there is either perfect positive or perfect negative correlation between the two variables, the two regression lines will coincide and we will have only one line. The farther the two regression lines from each other, the lesser is the degree of correlation and vice-versa. If the variables are independent, correlation is zero and the lines of regression will be at right angles. It should be noted that the regression lines cut each other at the point of average of x and y, i.e., if from the point where both the regression lines cut each other a perpendicular is drawn on the x-axis, we will get the mean value of x series and if from that point a horizontal line is drawn on the y-axis we will get the mean of y series.

2. Algebraic Method : The algebraic method for simple linear regression can be understood by two methods:

(i) Regression Equations

(ii) Regression Coefficients

Regression Equations : These equations are known as estimating equations. Regression equations are algebraic expressions of the regression lines. As there are two regression lines, there are two regression equations :
(i) x on y is used to describe the variations in the values of x for given changes in y.
(ii) y on x is used to describe the variations in the values of y for given changes in x.
The regression equations of y on x is expressed as

$Yc = a + bX$

The regression equations of x on y is expressed as

Xc = a + bY

In these equations a and b are constants which deretmine the position of the line completely. These constants are called the parameters of the line. If the value of any of these parameters is changed, another line is determind. Parameter a refers to the intercept of the line and b to the slope of the line. The symbol Yc and Xc refers to the values of Y computed and the value of X computed on the basis of independent variable in both the cases. If the values of both the parameters are obtained, the line is completely determined. The values of these two parameters a and b can be obtained by the method of least squares. With a little algebra and differential calculus it can be shown that the following two equations, are solved simultaneously, will give values of the parameters a and b such that the least squares requirement is fulfilled;

For regression equation Yc = a + bX

Sy = Na + bSx

Sxy = aSx + bSx2

For regression equation Xc = a + bY

Sx = Na + bSY

Sxy = aSy + bSy2

These equations are usually called the normal equations. In the equations Sx, Sy, Sxy, Sx2, Sy2 indicate totals which are computed from the observed pairs of values of two variables x and y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values. Let us understand by an example.

Example : From the following data obtain the two regression equations :

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

Solution :

Computation of Regression Equations

Regression coefficients of x on y is

bxy =

bxy =

bxy = where x = and y =

Regression Coefficient of Y on X is

byx =

byx =

byx = where x = and y =

Example : Calculate the regression coefficients from data given below :

	Series x	**Series y**	
Average	25	22	
Standard deviation	4	5	r = 0.8

Solution : The coefficient of regression of x on y is bxy =

The coefficient of regression of y on x is byx =

Properties of Regression Coefficients

(i) The coefficient of correlation is the geometric mean of the two regression coefficients, r =
(ii) Both the regression coefficients are either positive or negative. It means that they always have identical sign i.e., either both have positive sign or negative sign.
(iii) The coefficient of correlation and the regression coefficients will also have same sign.

(iv) If one of the regression coefficient is more than unity, the other must be less than unity because the value of coefficient of correlation can not exceed one (r = ± 1)
(v) Regression coefficients are independent of the change in the origin but not of the scale.
(vi) The average of regression coefficients is always greater than correlation coefficient.
We can compute the regression equations with the help of regression coefficients by the following equations:

1. **Regression equation X on Y**

Where is the mean of X series is the mean of X series is the regression coefficient of x on y

2. Regression equation Y on X =

We can explain this by taking an example :

Example : Calculate the following from the below given data :

(a) the two regression equations,

(b) the coefficient of correlation and

(c) the most likely marks in Statistics when the marks in Economics are 30

Marks in Economics : 25 28 35 32 31 36 29 38 34 32

Marks in Statistics : 43 46 49 41 36 32 31 30 33 39

Solution : Calculation of Regression Equations and Correlation Coefficient

Standard Error of Estimate

Standard error of an estimate is the measure of the spread of observed values from estimated ones, expressed by regression line or equation. The concept of standard error of estimate is analogous to the standard deviation which measures the variation or scatter of individual items about the arithmetic mean. Therefore, like the standard deviation which is the average of square of deviations about the arithmetic mean, the standard error of an estimate is the average of the square of deviations between the actual or the observed values and the estimated values based on the regression equation. It can also be expressed as the root of the measure of unexplained variations divided by N – 2 :

Syx =

and Sxy =

where Syx refers to standard error of estimate of Y values on X values. Sxy refers to standard error of estimate of X values on Y values.
Yc and Xc are the estimated values of Y and X variables by means of their regression equations respectively.
N – 2 is used for getting an unbiased estimate of standard error. The usual explanation given for this division by N – 2 is that the two constants a and b were calculated on the basis of original data and we lose two degrees of freedom.

Degrees of freedom mean the number of classes to which values can be assigned at will without violating any restrictions.

However a simpler method of computing Syx and Sxy is to use the following formulae :

Syx =

and Sxy =

The standard error of estimate measures the accuracy of the estimated figures. The smaller the values of standard error of estimate, the closer will be the dots to the regression line and the better the estimates based on the equation for this line. If standard error of estimate is zero, then there is no variation about the line and the correlation will be perfect. Thus with the help of standard error of estimate it is possible for us to ascertain how good and representative the regression line is as a description of the average relationship

between two series.

Example : Given the following data :

X	:	6	2	10	4	8
Y	:	9	11	5	8	7

And two regression equations Y = 11.09 – 0.65 X and X = 16.4 – 1.3 Y. Calculate the standard error of estimate i.e. Syx and Sxy.

Solution : We can calculate Xc and Yc values from these regression equations.

MATRIX

Definition of a Matrix

$$\mathbf{A}_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

- Rectangular array of real numbers
- *m* rows by *n* columns
- Named using capital letters
- First subscript is row, second subscript is column

Terminology

- A matrix with *m* rows and *n* columns is called a matrix of order *m* x *n*.
- A square matrix is a matrix with an equal number of rows and columns. Since the number of rows and columns are the same, it is said to have order *n*.
- The main diagonal of a square matrix are the elements from the upper left to the lower right of the matrix.
- A row matrix is a matrix that has only one row.
- A column matrix is a matrix that has only one column.
- A matrix with only one row or one column is called a vector.

Converting Systems of Linear Equations to Matrices

Each equation in the system becomes a row. Each variable in the system becomes a column. The variables are dropped and the coefficients are placed into a matrix. If the right hand side is included, it's called an augmented matrix. If the right hand side isn't included, it's called a coefficient matrix.

The system of linear equations ...

x + y - z = 1
3x - 2y + z = 3
4x + y - 2z = 9
becomes the augmented matrix ...

X	**Y**	**Z**	**rhs**
1	1	-1	1
3	-2	1	3

$$\left[\begin{array}{cc|c} 4 & 1 & -2 \\ & & 9 \end{array}\right]$$

Elementary Row Operations

Elementary Row Operations are operations that can be performed on a matrix that will produce a row-equivalent matrix. If the matrix is an augmented matrix, constructed from a system of linear equations, then the row-equivalent matrix will have the same solution set as the original matrix.

When working with systems of linear equations, there were three operations you could perform which would not change the solution set.

1. Interchange two equations.
2. Multiply an equation by a non-zero constant.
3. Multiply an equation by a non-zero constant and add it to another equation, replacing that equation.

When a system of linear equations is converted to an augmented matrix, each equation becomes a row. So, there are now three elementary row operations which will produce a row-equivalent matrix.

1. Interchange two rows
2. Multiply a row by a non-zero constant
3. Multiply a row by a non-zero constant and add it to another row, replacing that row.

Row-Echelon and Reduced Row-Echelon Forms

These are Row-equivalent forms of a matrix. One can easily solve a system of linear equations when matrices are in one of these forms.

Row-Echelon Form

A matrix is in row-echelon form when the following conditions are met.

1. If there is a row of all zeros, then it is at the bottom of the matrix.
2. The first non-zero element of any row is a one. That element is called the leading one.
3. The leading one of any row is to the right of the leading one of the previous row.

Notes

- The leading one of a row does not have to be to the *immediate* right of the leading one of the previous row.
- A matrix in row-echelon form will have zeros below the leading ones.
- Gaussian Elimination places a matrix into row-echelon form, and then back substitution is required to finish finding the solutions to the system.
- The row-echelon form of a matrix is not necessarily unique.

Reduced Row-Echelon Form

A matrix is in reduced row-echelon form when all of the conditions of row-echelon form are met and all elements above, as well as below, the leading ones are zero.

1. If there is a row of all zeros, then it is at the bottom of the matrix.
2. The first non-zero element of any row is a one. That element is called the leading one.
3. The leading one of any row is to the right of the leading one of the previous row.
4. All elements above and below a leading one are zero.

Notes

- The leading one of a row does not have to be to the *immediate* right of the leading one of the previous row.
- A matrix in row-echelon form will have zeros both above and below the leading ones.
- Gauss-Jordan Elimination places a matrix into reduced row-echelon form.
- No back substitution is required to finish finding the solutions to the system.
- The reduced row-echelon form of a matrix is unique.

Operations with Matrices

Equality

Two matrices are equal if and only if

- The order of the matrices are the same
- The corresponding elements of the matrices are the same

Addition

- Order of the matrices must be the same
- Add corresponding elements together
- Matrix addition is commutative
- Matrix addition is associative

Subtraction

- The order of the matrices must be the same
- Subtract corresponding elements
- Matrix subtraction is not commutative (neither is subtraction of real numbers)
- Matrix subtraction is not associative (neither is subtraction of real numbers)

Scalar Multiplication

A scalar is a number, not a matrix.

- The matrix can be any order
- Multiply all elements in the matrix by the scalar
- Scalar multiplication is commutative
- Scalar multiplication is associative

Zero Matrix

- Matrix of any order
- Consists of all zeros
- Denoted by capital O
- Additive Identity for matrices
- Any matrix plus the zero matrix is the original matrix

Matrix Multiplication

$A_{m\times n} \times B_{n\times p} = C_{m\times p}$

- The number of columns in the first matrix must be equal to the number of rows in the second matrix. That is, the inner dimensions must be the same.
- The order of the product is the number of rows in the first matrix by the number of columns in the second matrix. That is, the dimensions of the product are the outer dimensions.
- Since the number of columns in the first matrix is equal to the number of rows in the second matrix, you can pair up entries.
- Each element in row *i* from the first matrix is paired up with an element in column *j* from the second matrix.
- The element in row *i*, column *j*, of the product is formed by multiplying these paired elements and summing them.
- Each element in the product is the sum of the products of the elements from row *i* of the first matrix and column *j* of the second matrix.
- There will be *n* products which are summed for each element in the product.

See a complete example of matrix multiplication.

Matrix multiplication is not commutative

- Multiplication of real numbers is.
- The inner dimensions may not agree if the order of the matrices is changed.

Do not simply multiply corresponding elements together

- Since the order (dimensions) of the matrices don't have to be the same, there may not be corresponding elements to multiply together.
- Multiply the rows of the first by the columns of the second and add.

There is no matrix division

- There is no defined process for dividing a matrix by another matrix.
- A matrix may be divided by a scalar.

Identity Matrix

- Square matrix
- Ones on the main diagonal
- Zeros everywhere else
- Denoted by I. If a subscript is included, it is the order of the identity matrix.
- I is the multiplicative identity for matrices
- Any matrix times the identity matrix is the original matrix.
- Multiplication by the identity matrix is commutative, although the order of the identity may change

Identity matrix of size 2

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Identity matrix of size 3

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Properties of Matrices

Property	Example
Commutativity of Addition	A + B = B + A
Associativity of Addition	A + (B + C) = (A + B) + C
Associativity of Scalar Multiplication	(cd) A = c (dA)

Scalar Identity	1A = A(1) = A
Distributive	c (A + B) = cA + cB
Distributive	(c + d) A = cA + dA
Additive Identity	A + O = O + A = A
Associativity of Multiplication	A (BC) = (AB) C
Left Distributive	A (B + C) = AB + AC
Right Distributive	(A + B) C = AC + BC
Scalar Associativity / Commutativity	c (AB) = (cA) B = A (cB) = (AB) c
Multiplicative Identity	IA = AI = A

Properties of Real Numbers that aren't Properties of Matrices

Commutativity of Multiplication

- You can not change the order of a multiplication problem and expect to get the same product. AB≠BA
- You must be careful when factoring common factors to make sure they are on the same side. AX+BX = (A+B)X and XA + XB = X(A+B) but AX + XB doesn't factor.

Zero Product Property

- Just because a product of two matrices is the zero matrix does not mean that one of them was the zero matrix.

Multiplicative Property of Equality

- If A=B, then AC = BC. This property is still true, but the converse is not necessarily true. Just because AC = BC does not mean that A = B.

- Because matrix multiplication is not commutative, you must be careful to either pre-multiply or post-multiply on both sides of the equation. That is, if A=B, then AC = BC or CA = CB, but AC≠CB.

There is no matrix division

- You must multiply by the inverse of the matrix

Evaluating a Function using a Matrix

Consider the function $f(x) = x^2 - 4x + 3$ and the matrix A

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

The initial attempt to evaluate the f(A) would be to replace every x with an A to get $f(A) = A^2 - 4A + 3$. There is one slight problem, however. The constant 3 is not a matrix, and you can't add matrices and scalars together. So, we multiply the constant by the Identity matrix.

$f(A) = A^2 - 4A + 3I$.

Evaluate each term in the function and then add them together.

$$A^2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} * \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix}$$

$$-4\,A = -4 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} -4 & -8 \\ -12 & -16 \end{bmatrix}$$

$$3I = 3 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$f(A) = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix} + \begin{bmatrix} -4 & -8 \\ -12 & -16 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 3 & 9 \end{bmatrix}$$

Factoring Expressions

Some examples of factoring are shown. Simplify and solve like normal, but remember that matrix multiplication is not commutative and there is no matrix division.

$2X + 3X = 5X$
$AX + BX = (A+B)X$
$XA + XB = X(A+B)$
$AX + 5X = (A+5I)X$
$AX+XB$ does not factor

Solving Equations

A system of linear equations can be written as $AX=B$ where A is the coefficient matrix, X is a column vector containing the variables, and B is the right hand side. We'll learn how to solve this equation in the next section.

If there are more than one system of linear equations with the same coefficient matrix, then you can expand the B matrix to have more than one column. Put each right hand side into its own column.

Matrix Multiplication

Matrix multiplication involves summing a product. It is appropriate where you need to multiply things together and then add. As an example, multiplying the number of units by the per unit cost will give the total cost.

The units on the product are found by performing unit analysis on the matrices. The labels for the product are the labels of the rows of the first matrix and the labels of the columns of the second matrix.

The Inverse of a Matrix

So, what is the inverse of a matrix?

Well, in real numbers, the inverse of any real number a was the number a^{-1}, such that a times a^{-1} equaled 1. We knew that for a real number, the inverse of the number was the reciprocal of the number, as long as the number wasn't zero.

The inverse of a square matrix A, denoted by A^{-1}, is the matrix so that the product of A and A^{-1} is the Identity matrix. The identity matrix that results will be the same size as the matrix A. Wow, there's a lot of similarities there between real numbers and matrices. That's good, right - you don't want it to be something completely different.

$A(A^{-1}) = I$ or $A^{-1}(A) = I$

There are a couple of exceptions, though. First of all, A^{-1} does not mean 1/A. Remember, "There is no Matrix Division!" Secondly, A^{-1} does not mean take the reciprocal of every element in the matrix A.

Requirements to have an Inverse

1. The matrix must be square (same number of rows and columns).
2. The determinant of the matrix must not be zero (determinants are covered in section 6.4). This is instead of the real number not being zero to have an inverse, the determinant must not be zero to have an inverse.

A square matrix that has an inverse is called **invertible** or **non-singular**. A matrix that does not have an inverse is called **singular**.

A matrix does not have to have an inverse, but if it does, the inverse is unique.

Finding the Inverse the Hard Way

The inverse of a matrix A will satisfy the equation $A(A^{-1}) = I$.

1. Adjoin the identity matrix onto the right of the original matrix, so that you have A on the left side and the identity matrix on the right side. It will look like this [A | I].
2. Row-reduce (I suggest using pivoting) the matrix until the left side is the Identity matrix. When the left side is the Identity matrix, the right side will be the Inverse [I | A^{-1}]. If you are unable to obtain the identity matrix on the left side, then the matrix is singular and has no inverse.
3. Take the augmented matrix from the right side and call that the inverse.

Shortcut to the Finding the Inverse of a 2×2 Matrix

The inverse of a 2×2 matrix can be found by ...

1. Switch the elements on the main diagonal
2. Take the opposite of the other two elements
3. Divide all the values by the determinant of the matrix (since we haven't talked about the determinant, for a 2×2 system, it is the product of the elements on the main diagonal minus the product of the other two elements).

Example for the shortcut

Let's go with an original matrix of

$$\begin{bmatrix} 7 & -2 \\ 3 & 5 \end{bmatrix}$$

Step 1, switch the elements on the main diagonal would involve switching the 5 and 7.

$$\begin{bmatrix} \mathbf{5} & -2 \\ & \end{bmatrix}$$

$$\begin{bmatrix} 3 & \mathbf{7} \end{bmatrix}$$

Step 2, take the opposite of the other two elements, but leave them where they are.

$$\begin{bmatrix} 5 & \mathbf{2} \\ \mathbf{-3} & 7 \end{bmatrix}$$

Step 3, find the determinant and divide every element by that. The determinant is the product of the elements on the main diagonal minus the product of the elements off the main diagonal. That means the determinant of this matrix is 7(5) - (-3)(2) = 35 + 6 = 41. We divide every element by 41.

The inverse of the original matrix is ...

$$\begin{bmatrix} 5/41 & 2/41 \\ -3/41 & 7/41 \end{bmatrix}$$

Now, you're saying, wait a minute - you said there was no matrix division. There is no division by a matrix. You may multiply or divide a matrix by a scalar (real number) and the determinant is a scalar.

Using the Calculator

Now that you know how to find the identity matrix by hand, let's talk practicality. The calculator will do it for you.

Entering a Matrix

1. Press the Matrix key (right below the X key). On the TI-83+, you will need to hit 2nd Matrix.
2. Arrow to the Edit submenu.
3. Choose a Matrix to work with. You have five to choose from with the TI-82 and ten to choose from with the TI-83. Typically, you will use [A]. Try to avoid using [E] for unspecified reasons that will be specified if you take Finite Mathematics.
4. Enter the number of rows, press enter, and then enter the number of columns, followed by enter.
5. You now enter each element in the matrix, reading from left to right and top to bottom. Press enter after each number. You may use the arrow keys to move around if you make a mistake.
6. Quit (2nd Mode) when you are done entering all the numbers.

Using Matrices

Whenever you need to access a matrix that you have created, just hit the Matrix key and choose the appropriate matrix. I would suggest that you start using Matrix 1, Matrix 2, etc, instead of Matrix, arrow down, enter. It will go faster, and you will be doing a lot with these matrices.

Finding the Inverse of a Matrix on a Calculator

Enter the expression $[A]^{-1}$ by going Matrix 1, and then hitting the x^{-1} key. It will not work if you try to raise the matrix to the -1 power as in [A]^(-1).

You may have to use the right or left arrow keys to scroll through the entire matrix to write it down. Please give exact answers whenever possible.

One way of giving exact answers is to have the calculator convert the decimals to fractions for you. After all, fractions really are your friends (and I seriously mean that here). You can have the calculator do a decimal to fraction conversion by hitting Math, Enter, Enter.

Also, if you get an answer like 1.2E-12, chances are really good that number is zero and it is because of inaccuracies in the calculator that you are getting that response. Convert the number to zero.

Why was it we needed an inverse?

I am so glad you asked that.

One of the major uses of inverses is to solve a system of linear equations. You can write a system in matrix form as AX = B.

Now, pre-multiply both sides by the inverse of A. Make sure you meet these two conditions.

1. You must place the inverse of the matrix adjacent to the matrix. That is because Inverses need to be next to each other (very loose mathematically, but think back to functions) in order to undo each other.
2. If you multiply by putting something in front of the left side (pre-multiply), it has to go in front of the right side. If you put something behind (post-multiply) the left side, it has to go behind the right side.

Matrix Multiplication is NOT Commutative!

$A^{-1}(AX) = A^{-1}(B)$ ... pre-multiply both sides by A^{-1}
$(A^{-1}A)\ X = A^{-1}\ B$ ... use the associative property to regroup factors
$I\ X = A^{-1}\ B$ ... when you multiply inverses together, they become the identity matrix
$X = A^{-1}\ B$ ... the identity matrix is like multiplying by 1.

If AX = B, then $X = A^{-1}\ B$

So what you're asking in your normal cynical way is "You've just solved another equation, what does that have to do with anything?"

Solving Systems of Linear Equations

Consider the system of linear equations

$$3x + 2y - 5z = 12$$
$$x - 3y + 2z = -13$$
$$5x - y + 4z = 10$$

Write the coefficients in an A matrix.

$$\begin{matrix} \mathbf{X} & \mathbf{y} & \mathbf{z} \end{matrix}$$

$$\begin{bmatrix} 3 & 2 & -5 \\ 1 & -3 & 2 \\ 5 & -1 & 4 \end{bmatrix}$$

Write the variables in an X matrix.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

Write the constants in a B matrix.

$$\begin{bmatrix} 12 \\ -13 \\ 10 \end{bmatrix}$$

Verify that AX = B

This step isn't really needed, but I wanted to show you that this thing really does work.

AX will be a (3×3) × (3×1) = 3×1 matrix. The B matrix is also a 3×1 matrix, so at least the dimensions work out right.

Here's A times X.

$$\begin{bmatrix} 3 & 2 & -5 \\ 1 & -3 & 2 \\ 5 & -1 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3x + 2y - 5z \\ 1x - 3y + 2z \\ 5x - 1y + 4z \end{bmatrix}$$

Notice that turns out to be the left side of the system of equations. The B is the right hand side, so we have achieved equality. Woohoo! You can write a system of linear equations as AX = B.

So, if you can write a system of linear equations as AX=B where A is the coefficient matrix, X is the variable matrix, and B is the right hand side, you can find the solution to the system by $X = A^{-1} B$.

Place the coefficient matrix into [A] on the calculator and the right hand side into [B].

If you asked the calculator to find the inverse of the coefficient matrix, it would give you this for A^{-1}

$$\begin{bmatrix} 5/44 & 3/88 & 1/8 \\ -3/44 & -37/88 & 1/8 \\ -7/44 & -13/88 & 1/8 \end{bmatrix}$$

You could do that, and then multiply that by B, but it would be easier just to put the whole expression into the calculator and get the answer directly. Even what is shown below is more work than is necessary.

$X = A^{-1} B = ...$

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 5/44 & 3/88 & 1/8 \\ -3/44 & -37/88 & 1/8 \\ -7/44 & -13/88 & 1/8 \end{bmatrix} \begin{bmatrix} 12 \\ -13 \\ 10 \end{bmatrix} = \begin{bmatrix} 191/88 \\ 519/88 \\ 111/88 \end{bmatrix}$$

There you go, x = 191/88, y = 519/88, and z = 111/88. That would be a real pain to solve by hand.

This is easy, why don't we always do this?

The main reason is because it doesn't always work.

1. Inverses only exist for square matrices. That means if you don't the same number of equations as variables, then you can't use this method.

2. Not every square matrix has an inverse. If the coefficient matrix A is singular (has no inverse), then there may be no solution or there may be many solutions, but we can't tell what it is.
3. Inverses are a pain to find by hand. If you have a calculator, it's not so bad, but remember that calculators don't always give you the answer you're looking for.

The Determinant of a Square Matrix

A determinant is a real number associated with every square matrix. I have yet to find a good English definition for what a determinant is. Everything I can find either defines it in terms of a mathematical formula or suggests some of the uses of it. There's even a definition of determinant that defines it in terms of itself.

The determinant of a square matrix A is denoted by "det A" or | A |. Now, that last one looks like the absolute value of A, but you will have to apply context. If the vertical lines are around a matrix, it means determinant.

The line below shows the two ways to write a determinant.

$$\begin{vmatrix} 3 & 1 \\ 5 & 2 \end{vmatrix} = \det \begin{bmatrix} 3 & 1 \\ 5 & 2 \end{bmatrix}$$

Determinant of a 2×2 Matrix

The determinant of a 2×2 matrix is found much like a pivot operation. It is the product of the elements on the main diagonal minus the product of the elements off the main diagonal.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Properties of Determinants

- The determinant is a real number, it is not a matrix.
- The determinant can be a negative number.
- It is not associated with absolute value at all except that they both use vertical lines.
- The determinant only exists for square matrices (2×2, 3×3, ... n×n). The determinant of a 1×1 matrix is that single value in the determinant.
- The inverse of a matrix will exist only if the determinant is not zero.

Expansion using Minors and Cofactors

The definition of determinant that we have so far is only for a 2×2 matrix. There is a shortcut for a 3×3 matrix, but I firmly believe you should learn the way that will work for all sizes, not just a special case for a 3×3 matrix.

The method is called expansion using minors and cofactors. Before we can use them, we need to define them.

Minors

A minor for any element is the determinant that results when the row and column that element are in are deleted.

The notation M_{ij} is used to stand for the minor of the element in row i and column j. So M_{21} would mean the minor for the element in row 2, column 1.

Consider the 3×3 determinant shown below. I've included headers so that you can keep the rows and columns straight, but you wouldn't normally include those. We're going to find some of the minors.

	C_1	C_2	C_3
R_1	1	3	2
R_2	4	1	3
R_3	2	5	2

Finding the Minor for R_2C_1

The minor is the determinant that remains when you delete the row and column of the element you're trying to find the minor for. That means we should delete row 2 and column 1 and then find the determinant.

	C_2	C_3
R_1	3	2
R_3	5	2

$= 3(2) - 5(2) = 6 - 10 = -4$

As you can see, the minor for row 2 and column 1 is $M_{21} = -4$.

Let's try another one.

Finding the Minor for R_3C_2

This time, we would delete row 3 and column 2.

	C_1	C_3

R1	1	2	= 1(3) - 4(2) = 3 - 8 = -5
R2	4	3	

So the minor for row 3, column 2 is $M_{32} = -5$.

Matrix of Minors

When you're just trying to find the determinant of a matrix, this is overkill. But there is one extremely useful application for it and it will give us practice finding minors.

The matrix of minors is the square matrix where each element is the minor for the number in that position.

Here is a generic matrix of minors for a 3×3 determinant.

	C1	**C2**	**C3**
R1	M_{11}	M_{12}	M_{13}
R2	M_{21}	M_{22}	M_{23}
R3	M_{31}	M_{32}	M_{33}

Let's find the matrix of minors for our original determinant. Here is the determinant.

	C1	**C2**	**C3**
R1	1	3	2
R2	4	1	3
R3	2	5	2

Here is the work to find each minor in the matrix of minors.

	C1	**C2**	**C3**
R1	$\begin{vmatrix} 1 & 3 \\ 5 & 2 \end{vmatrix}$	$\begin{vmatrix} 4 & 3 \\ 2 & 2 \end{vmatrix}$	$\begin{vmatrix} 4 & 1 \\ 2 & 5 \end{vmatrix}$

	= 2 - 15 = -13	= 8 - 6 = 2	= 20 - 2 = 18
R_2	$\begin{vmatrix} 3 & 2 \\ 5 & 2 \end{vmatrix}$ = 6 - 10 = -4	$\begin{vmatrix} 1 & 2 \\ 2 & 2 \end{vmatrix}$ = 2 - 4 = -2	$\begin{vmatrix} 1 & 3 \\ 2 & 5 \end{vmatrix}$ = 5 - 6 = -1
R_3	$\begin{vmatrix} 3 & 2 \\ 1 & 3 \end{vmatrix}$ = 9 - 2 = 7	$\begin{vmatrix} 1 & 2 \\ 4 & 3 \end{vmatrix}$ = 3 - 8 = -5	$\begin{vmatrix} 1 & 3 \\ 4 & 1 \end{vmatrix}$ = 1 - 12 = -11

Finally, here is the matrix of minors. Again, you don't need to put the labels for the row and columns on there, but it may help you.

	C_1	**C_2**	**C_3**
R_1	-13	2	18
R_2	-4	-2	-1
R_3	7	-5	-11

Cofactors

A cofactor for any element is either the minor or the opposite of the minor, depending on where the element is in the original determinant. If the row and column of the element add up to be an even number, then the cofactor is the same as the minor. If the row and column of the element add up to be an odd number, then the cofactor is the opposite of the minor.

Ooh - did you get that? Odd changes signs, even is the same sign. Deja Vu. We've been talking about that ever since section 3.2 on polynomials.

Sign Chart

Rather than adding up the row and column of the element to see whether it is odd or even, many people prefer to use a sign chart. A sign chart is either a + or - for each element in the matrix. The first element (row 1, column 1) is always a + and it alternates from there.

Note: The + does not mean positive and the - negative. The + means the same sign as the minor and the - means the opposite of the minor. Think of it addition and subtraction rather than positive or negative.

Here is the sign chart for a 2×2 determinant.

	C_1	C_2
R_1	+	-
R_2	-	+

Here is the sign chart for a 3×3 determinant.

	C_1	C_2	C_3
R_1	+	-	+
R_2	-	+	-
R_3	+	-	+

Matrix of Cofactors

Again, if all you're trying to do is find the determinant, you do not need to go through this much work.

The matrix of cofactors is the matrix found by replacing each element of a matrix by its cofactor. This is the matrix of minors with the signs changed on the elements in the - positions.

	C_1	C_2	C_3
R_1	-13	-2	18
R_2	4	-2	1
R_3	7	5	-11

Expanding to Find the Determinant

Here are the steps to go through to find the determinant.

1. Pick any row or column in the matrix. It does not matter which row or which column you use, the answer will be the same for any row. There are some rows or columns that are easier than others, but we'll get to that later.
2. Multiply every element in that row or column by its cofactor and add. The result is the determinant.

Let's expand our matrix along the first row.

$$\begin{vmatrix} 1 & 3 & 2 \\ 4 & 1 & 3 \\ 2 & 5 & 2 \end{vmatrix}$$

From the sign chart, we see that 1 is in a positive position, 3 is in a negative position and 2 is in a positive position. By putting the + or - in front of the element, it takes care of the sign adjustment when going from the minor to the cofactor.

$$+1 \begin{vmatrix} 1 & 3 \\ 5 & 2 \end{vmatrix} - 3 \begin{vmatrix} 4 & 3 \\ 2 & 2 \end{vmatrix} + 2 \begin{vmatrix} 4 & 1 \\ 2 & 5 \end{vmatrix}$$

$= 1 (2 - 15) - 3 (8 - 6) + 2 (20 - 2)$
$= 1 (-13) - 3 (2) + 2 (18)$
$= -13 - 6 + 36$
$= 17$

The determinant of this matrix is 17.

As I said earlier, it doesn't really matter which row or column you use.

Let's try it again, but this time expand on the second columns. As an effort to save time, the minors for that column (from the matrix of minors) were 2, -2, and -5. The original elements were 3, 1, and 5. The 3 and 5 are in negative positions.

determinant = $- 3 (2) + 1 (-2) - 5 (-5) = -6 -2 + 25 = 17$

Expand on any row or any column, you'll get 17.

However, you can't do diagonals. If we try the main diagonal, you get

$+ 1 (-13) + 1 (-2) + 2 (-11) = -13 -2 - 22 = -37$

Some rows or columns are better than others

1. Pick the row or column with the most zeros in it.
   Since each minor or cofactor is multiplied by the element in the matrix, picking a row or column with lots of zeros in it means that you will be multiplying by a lot of zeros.

Multiplying by zero doesn't take very long at all. In fact, if the element is zero, you don't need to even find the minor or cofactor.

2. Pick the row or column with the largest numbers (or variables) in it.
   The elements in the row or column that you expand along are not used to find the minors. The only place that they are multiplied is once, in the expansion. If you pick the row or column with the smallest numbers, then every minor will be the product of larger numbers.
   If you pick a row or column that has variables in it, then you will only have to multiply by the variables once, during the expansion.

Inverse of a Matrix (revisited)

Let's consider our original determinant as a matrix this time.

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 1 & 3 \\ 2 & 5 & 2 \end{bmatrix}$$

Find the **matrix of minors** as explained above.

$$\begin{bmatrix} -13 & 2 & 18 \\ -4 & -2 & -1 \\ 7 & -5 & -11 \end{bmatrix}$$

Turn it into a **matrix of cofactors** by changing the signs on the appropriate elements based on the sign chart.

$$\begin{bmatrix} -13 & -2 & 18 \\ 4 & -2 & 1 \\ 7 & 5 & -11 \end{bmatrix}$$

Find the **adjoint** by transposing the matrix of cofactors.

To transpose a matrix, you switch the rows and columns. That is, the rows become columns and the columns become rows. The Transpose of a matrix can be found using the TI-82 or TI-83 calculator by entering the name of the matrix and then choosing Matrix, Math, and then option 2, a superscripted T, like $[A]^T$.

$$\begin{bmatrix} -13 & 4 & 7 \end{bmatrix}$$

$$\begin{bmatrix} -2 & -2 & 5 \\ 18 & 1 & -11 \end{bmatrix}$$

Finally divide the adjoint of the matrix by the determinant of the matrix. In this problem, the determinant is 17, so we'll divide every element by 17. The resulting matrix is the **inverse** of the original matrix.

$$\begin{bmatrix} -13/17 & 4/17 & 7/17 \\ -2/17 & -2/17 & 5/17 \\ 18/17 & 1/17 & -11/17 \end{bmatrix}$$

The inverse of a matrix is found by dividing the adjoint of the matrix by the determinant of the matrix. Don't try that on your calculator since the calculator won't let you divide a matrix by a scalar. You will have to multiply by the inverse of the determinant instead.

If you check it with your calculator, you can verify that the inverse actually is the adjoint divided by the determinant.

Since the inverse is the adjoint divided by the determinant, we can see why the inverse doesn't exist if the determinant is zero. That would cause division by zero, which is undefined.

Larger Order Determinants

Let's find the determinant of a 4x4 system.

	C_1	C_2	C_3	C_4
R_1	3	2	0	1
R_2	4	0	1	2
R_3	3	0	2	1
R_4	9	2	3	1

Pick the row or column with the most zeros in it. In this case, that is the second column.

For each element in the original matrix, its minor will be a 3×3 determinant. We'll have to expand each of those by using three 2×2 determinants.

This is why we want to expand along the second column. The minors are multiplied by their elements, so if the element in the original matrix is 0, it doesn't really matter what the minor is and we can save a lot of time by not having to find it. In the second column, you won't need to find two of the minors because their corresponding element in the second column is zero.

$$-2\begin{vmatrix} 4 & 1 & 2 \\ 3 & 2 & 1 \\ 9 & 3 & 1 \end{vmatrix} + 0\begin{vmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{vmatrix} - 0\begin{vmatrix} ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{vmatrix} + 2\begin{vmatrix} 3 & 0 & 1 \\ 4 & 1 & 2 \\ 3 & 2 & 1 \end{vmatrix}$$

We could actually fill in those middle two minors, but since they're multiplied by 0, it doesn't really matter what they are. In fact, you could just as easily skip them.

Now, there are two 3x3 determinants left to find.

In the first 3x3 determinant, there are no zeros, so pick the row or column with the largest numbers. That would be column 1, so expand along the first column.

Notice the 4 is in a positive position. The sign charts begin over with each new determinant. The position of the number in the original matrix does not matter, only its position in the current matrix.

$$\begin{vmatrix} 4 & 1 & 2 \\ 3 & 2 & 1 \\ 9 & 3 & 1 \end{vmatrix} = +4\begin{vmatrix} 2 & 1 \\ 3 & 1 \end{vmatrix} - 3\begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix} + 9\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix}$$

$= 4 (2 - 3) - 3 (1 - 6) + 9 (1 - 4) = 4 (-1) - 3 (-5) + 9 (-3) = -4 + 15 - 27 = -16$

Consider the other 3×3 matrix. In this one, there is a 0 in the row 1 and column 2. Either one of those would be a good pick for expansion, but since row 1 has slightly larger numbers, we'll expand along the first row.

$$\begin{vmatrix} 3 & 0 & 1 \\ 4 & 1 & 2 \\ 3 & 2 & 1 \end{vmatrix} = +3\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} - 0\begin{vmatrix} ? & ? \\ ? & ? \end{vmatrix} + 1\begin{vmatrix} 4 & 1 \\ 3 & 2 \end{vmatrix}$$

$= 3 (1 - 4) - 0 (\text{doesn't matter}) + 1 (8 - 3) = 3 (-3) + 1 (5) = -9 + 5 = -4$

When you go to find the determinant, remember that there were elements from the original 4×4 matrix that were times each of those 3×3 determinants. The first one was -2 and the second one was +2.

Determinant = $-2 (-16) + 2 (-4) = 32 - 8 = 24$

Worst case scenario

To find a 3x3 determinant with no zeros, you have to find three 2x2 determinants.

To find a 4x4 determinant with no zeros, you have to find four 3x3 determinants, each of which then becomes three 2x2 determinants for a total of twelve 2x2 determinants.

To find a 5x5 determinant with no zeros, you have to find five 4x4 determinants, each of which then becomes four 3x3 determinants, each of those becoming three 2x2 determinants for a total of sixty 2x2 determinants.

Using the Calculator

After that last problem, you've got to be asking yourself if there isn't an easier way. Well, yes, there is, as long as the determinant doesn't have any variables in it. You can use the calculator.

The notation that the TI-82 or TI-83 calculator uses is the Det A notation. So, after entering the matrix into one of the available matrices on the calculator, enter DET by going Matrix, Math, and choosing option 1. Then put in the name of the matrix that you're using.

You don't need to use parentheses (unless you have a TI-83), but you can if you want to find the determinant of a product "det ([A]*[B])" or the determinant of a transpose "det $([A]^T)$" as opposed to the transpose of the determinant "$(\det [A])^T$". By the way, the calculator won't find the transpose of a determinant because the determinant is a scalar (real number) and the calculator only knows how to find the transpose of a matrix. The transpose of a scalar is that scalar.

Triangular Matrices

You're really going to like finding determinants of these matrices.

Upper Triangular Matrix

> A matrix in which all the non-zero elements are either on or above the main diagonal. That is, all the non-zero values are in the upper triangle. Everything below the diagonal is a zero.

Lower Triangular Matrix

> A matrix in which all the non-zero elements are either on or below the main diagonal.
>
> That is, all the non-zero values are in the lower triangle. Everything above the diagonal is zero.

Diagonal Matrix

> A matrix in which all the non-zero elements are on the main diagonal. Everything off the main diagonal is a zero.

The determinant of a triangular matrix or a diagonal matrix is the product of the elements on the main diagonal.

Elementary Row Operations

There were three elementary row operations that could be performed that would return an equivalent system. With determinants, since the determinant of a transpose is the same as the determinant of the matrix, the elementary row operations can also be applied to columns.

By performing row-reduction (using pivoting on a 1 if you like), you can place a matrix into triangular form. Once it's in triangular form, then all you have to do is multiply by the elements on the main diagonal and you have the determinant.

Let's look at each of the three elementary row operations.

1. If you interchange two rows or two columns in a determinant, the resulting determinant will differ only in sign. That is, if you swap rows or columns, the resulting determinant is the opposite of the original determinant.
2. If you multiply a row or column by a non-zero constant, the determinant is multiplied by that same non-zero constant.
3. If you multiply a row or column by a non-zero constant and add it to another row or column, replacing that row or column, there is no change in the determinant.

That last operation is equivalent to pivoting on a one!

Warning, if your pivot is a number other than one, then you are multiplying each row that you change by the pivot element. So, if you pivot on a 3 and you change two rows, then the resulting determinant will be 3*3 = 9 times as great as the original determinant.

As long as you pivot on a one, you'll be okay.

You do not have to place the matrix into reduced row-echelon form or even row-echelon form. You are free to stop the reduction at any point and expand using minors and cofactors. What I suggest is pivot where there is a one, and then expand.

Determinants that are Zero

The determinant of a matrix will be zero if

1. An entire row is zero.
2. Two rows or columns are equal.
3. A row or column is a constant multiple of another row or column.

Remember, that a matrix is invertible, non-singular, if and only if the determinant is not zero. So, if the determinant is zero, the matrix is singular and does not have an inverse

SET THEORY

Basic definitions

A set is any collection of objects, for example, set of numbers. The objects of a set are called the elements of the set. A set may be specified by listing its elements. For example, {1,3,6} denotes the set with elements 1, 3 and 6. This is called the list form for the set. Note the curly brackets.∗ * Typographical terms: { opening curly bracket } closing curly bracketWe usually use capital ∗ letters A, B, C, etc., to denote * capital letter = upper case letter sets. The notation x ∈ A means "x is an element of A".∗ But * Alternatively we may say "x belongs to A" or "A contains x". x 6∈ A means "x is not an element of A". Example 1.1.1 1 ∈{1,3,6}, 3 ∈{1,3,6}, 6 ∈{1,3,6} but 2 6∈{1,3,6}.

Predicate and list form of definition of a set A set can also be specified in predicate form∗, that is by giving * or descriptive form a distinguished property of the elements of the set (or an explicit∗ description of the elements in the set). For example, * explicit = specific, definite

18 we can define set B by B = {x : x is a possitive integer less than 5}. The way to read this notation is "B is the set of all x such that x is a positive integer less than 5". The curly brackets indicate a set and the colon∗ * Typographical terms: : colon 0 : 0 is used to denote "such that", and, not surprisingly, is read "such that".1 The same set B can be given by listing its elements, or in list form: B = {1,2,3,4}.

Equality of sets Two sets are equal∗ if they have exactly the same elements. * We also say: two sets coincide. Thus {1,2,3,4} = {x : x is a possitive integer less than 5}. 1I have received this delightful email from David Rudling:

I have been working through your lecture notes at home now that I am retired and trying to catch up on not going to university when younger. I have noticed that when introducing : as the symbol for "such that" in set theory you have not added an asterisk commentary note mentioning the American usage of the vertical bar | as an alternative which your students will undoubtedly encounter. Might I have the temerity to suggest that an asterisk comment on this would be helpful

Indeed, in some books you can find this notation for sets: B = {x | x is a possitive integer less than 5}.

In list form the same set is denoted whatever order the elements are listed and however many times each element is listed. Thus {2,3,5} = {5,2,3} = {5,2,3,2,2,3}. Note that {5,2,3,2,2,3} is a set with only 3 elements: 2, 3 and 5.

Example 1.3.1 {x : x is a letter in the word GOOD } = {D,G,O}. The set {2}is regarded as being different from the number 2. A set of numbers is not a number. {2}is a set with only one element which happens to be the number 2. But a set is not the same as the object it contains: {2}6= 2. The statement 2 ∈{2} is correct. The statement {2}∈{2} is wrong.

Example The sets of letters in the words GOOD and DOG are equal.

The set

{x : x is an integer such that x2 = −1} has no elements. This is called an empty set∗. It was said * Some books call it null set. earlier that two sets are equal if they have the same elements. Thus if A and B are empty sets we have A = B. Mathematicians have found that this is the correct viewpoint, and this makes our first theorem.∗ * The word theorem means a statement that has been proved and therefore became part of mathematics. We shall also use words proposition and lemma: they are like theorem, but a proposition is usually a theorem of less importance, while lemma has no value on its own and is used as a step in a proof of a theorem. Theorem. If A and B are empty sets then A = B. Proof.∗ The sets A and B are equal because they cannot be * The word proof indicates that an argument establishing a theorem or other statement will follow. non-equal. Indeed, for A and B not to be equal we need an element in one of them, say a ∈ A, that does not belong to B. But A contains no elements! Similarly, we cannot find an element b ∈ B that does not belong to A – because B contains no elements at all. ——————————— Corollary.∗ There is only one empty set, THE∗ empty set. * Corollary is something that easily follows from a theorem or a proposition. * Notice the use of definite article THE.The empty set is usually denoted by ∅. Thus {x : x is an integer such that x2 = −1} = ∅.

A set cannot be an element of itself

We have complete freedom of forming sets, but one rule is of absolute importance: you cannot form a set containing itself as an element: A 6∈ A for all sets A!∗ * We shall revisit this principle later in the lectures, when we shall consider the so-called self-referential statements and various paradoxes associated with them. This means that when we are forming a set, we assume that its elements are somehow already given to us; but the set itself is not made yet, it is still in the process of construction. In particular, there is no set of all sets – because this set would contain itself as an element.

Questions from students ∗ * This section contains no compulsory material but still may be useful. 1. My question is: Are all empty sets equal? No matter the conditions. For example is {x : x is positive integer less than zero} equal to {x : x is an integer between 9 and 10}

Answer. Yes, all empty sets are equal. To see that in your example, let us denote A = {x : x is positive integer less than zero} and

B = {x : x is an integer between 9 and 10} So, I claim that A = B. If you do not agree with me, you have to show that A is different from B. To do so, you have to show me an element in one set that does not belong to another set. Can you do that? Can you point to an offending element if both sets have no elements whatsoever? Indeed, can you point to a "positive integer less than zero" which is not an "integer between 9 and 10"? Of course, you cannot, because there are no positive integers less than zero. Can you point to an "integer between 9 and 10" which is not a "positive integer less than zero"? Of course, you cannot, because there are no integers between 9 and 10. Hence you cannot prove that A is not equal to B. Therefore you have to agree with me that A = B.

Subsets; Finite and Infinite Sets

Subsets Consider the sets A and B where A = {2,4} and B = {1,2,3,4,5}. Every element of the set A is an element of the set B. We say that A is a subset of B and write A ⊆ B, or B ⊇ A. We can also say that B contains A.∗ * Also: A is contained in B, A is included in B. The expression B ⊇ A is read "B is a superset of A", or B contains A. Notice that the word "contains" is used in set theory in two meanings, it can be applied to elements and to subsets: the set{a,b,c}contains an element a and a subset {a}. Symbols used are different: a ∈{a,b,c}, {a}⊆{a,b,c}, and a 6= {a}.

Venn diagram

Figure 1: A diagram of A ⊆ B (which is the same as B ⊇ A). Figure 1 is a simple example of a Venn diagram for showing relationships between sets. Figure 2 is an example of a Venn diagram for three sets G, L, C of uppercase letters of the Greek, Latin and Cyrillic alphabets, respectively. Some basic facts:

• A ⊆ A for every set A. Every set is a subset of itself. [Indeed every element of A is an element of A. Hence, by definition of a subset, A is a subset of A.] • The empty set is a subset of every set: ∅ ⊆ A for any set A. [Indeed every element of ∅ is an element of A because there is no any elements in ∅.]

Figure 2: Venn diagram showing which uppercase letters are shared by the Greek, Latin and Cyrillic alphabets (sets G, L, C, respectively). • If A ⊆ B and B ⊆ C then A ⊆ C.∗ * We say that ⊆ is a transitive relation between sets. Notice that the relation ∈ "being an element of" is not transitive. relation = connection, bond• If A ⊆ B and B ⊆ A then A = B.

The set of subsets of a set Example 2.2.1 Let A = {1,2}. Denote by B the set of subsets of A. Then B = {∅,{1},{2},{1,2}}. Notice that 1 ∈{1} and 1 ∈ A, but it is not true that 1 ∈ B. On the other hand, {1}∈ B, but it is not true that {1}∈A .

Example 2.2.2 The subsets of {1,2,3} are ∅, {1}, {2}, {3}, {1,2}, {1,3}, {2,3}, {1,2,3}. Note: don't forget the empty set∅and the whole set{1,2,3}.Thus {1,2,3} has 8 subsets.

Theorem. If A is a set with n elements then A has 2n subsets. Here, 2n = 2×2×⋯×2 with n factors. Proof. Let A = {a1,a2,...,an}. How many are there ways to choose a subset in A? When choosing a subset, we have to decide, for each element, whether we include this elements into our subset or not. We have two choices for the first element: 'include' and 'do not include', two choices for the second element, etc., and finally two choices for the nth element: 2×2×⋯×2 choices overall. ——————————

Proper subsets If A ⊆ B and A 6= B we call A a proper subset of B and write A ⊂ B to denote this.∗ * If A ⊂ B, we also write B ⊃ A. Similarly, A ⊆ B is the same as B ⊇ A Example 2.3.1 Let A = {1,3}, B = {3,1}, C = {1,3,4}. Then A = B true A ⊂ B false C ⊆ A false A ⊆ B true A ⊆ C true C ⊂ C false B ⊆ A true A ⊂ C true Compare with inequalities for numbers: 2 6 2 true, 1 6 2 true, 2 < 2 false, 1 < 2 true.

A set with n elements contains 2n −1 proper subsets.

Finite and infinite sets

A finite set is a set containing only finite number of elements. For example,{1,2,3}is finite. If A is a finite set, we denote by |A| the number of elements in A. For example, |{1,2,3}| = 3 and |∅| = 0. A set with infinitely many elements is called an infinite set. The set of all positive integers (also called natural numbers) N = {1,2,3,..., } is infinite; the dots indicate that the sequence 1,2,3 is to be continued indefinitely.∗ * indefinitely = for ever, without end The set of all non-negative integers∗ is also infinite: * There is no universal agreement about whether to include zero in the set of natural numbers: some define the natural numbers to be the positive integers {1,2,3,...}, while for others the term designates the non-negative integers {0,1,2,3,...}. In this lecture course, we shall stick to the first one (and more traditional) convention: 0 is not a natural number. N0 = {0,1,2,3,..., }.

More examples of infinite sets: Z = {...,−2,−1,0,1,2,...} (the set of integers) {...,−4,−2,0,2,4,...} (the set of all even integers) {...,−3,−1,1,3,...} (the set of all odd integers) Q denotes the set of all rational numbers (that is, the numbers of the form n/m where n and m are integers and m 6= 0), R the set of all real numbers (in particular, √2 ∈ R and π ∈R), C the set of all complex numbers (that is, numbers of the form x + yi, where x and y are real and i is a square root of −1, i2 = −1).∗ * The lettersABCDEFGHIJKLMOPRSTUVWXYZ are called blackboard bold and were invented by mathematicians for writing on a blackboard instead of bold letters ABC... which are difficult to write with chalk. They are all infinite sets. We have the following inclusions: N⊂N0 ⊂Z⊂Q⊂R⊂C.

Operations on Sets

A∩B A∪B

Sets A and B and their intersection A∩B and union A∪B.

Intersection Suppose A and B are sets. Then A∩B denotes the set of all elements which belong to both A and B: A∩B = {x : x ∈ A and x ∈ B}. A∩B is called the intersection of A and B.∗ * The typographic symbol ∩ is sometimes called “cap”. Notice that the name of a typographical symbol for an operation is not necessary the same as the name of operation. For example, symbol plus is used to denote addition of numbers, like 2 + 3. Example 3.1.1 Let A = {1,3,5,6,7} and B = {3,4,5,8}, then A∩B = {3,5}.

3.2 Union A∪B denotes the set of all elements which belong to A or to B: A∪B = {x : x ∈ A or x ∈ B}. A∪B is called the union of A and B.∗ * The typographic symbol ∪ is sometimes called “cup”. Notice that, in mathematics, or is usually understood in the inclusive sense: elements from A∪B belong to A or to

B or to both A and B; or, in brief, to A and/or B. In some human languages, the connective∗ ‘or’ is understood in the * “Connective” is a word like ‘or’, ‘and’, ‘but’, ‘if’, ...exclusive sense: to A or to B, but not both A and B. We will always understand ‘or’ as inclusive ‘and/or’. In particular, this means that A∩B ⊆ A∪B. Example 3.2.1 Let A = {1,3,5,6,7}, B = {3,4,5,8}, then A∪B = {1,3,4,5,6,7,8}. If A and B are sets such that A∩B = ∅, that is, A and B have no elements in common, we say that A is disjoint from B, or that A and B are disjoint∗ (from each other). * Or that A and B do not intersect.

Example 3.2.2 A = {1,3,5}, B = {2,4,6}. Here A and B are disjoint.

Universal set and complement

In any application of set theory all the sets under consideration will be subsets of a background set, called the universal set. For example, when working with real numbers the universal set is the set R of real numbers. We usually denote the universal set by U. U is conveniently shown as a “frame” when drawing a Venn diagram. All the sets under consideration are subsets of U and so can be drawn inside the frame.

Let A be a set and U be the universal set. Then A0 (called the complement∗ of A and pronounced “A prime”) denotes * Notice that the complement A0 is sometimes denoted ¬A and pronounced “not A”, or A (pronounced “A bar”), or Ac (“A compliment”) the set of all elements in U which do not belong to A: A0 = {x : x ∈ U and x 6∈ A}.

Let U = {a,b,c,d,e,f}, A = {a,c}, B = {b,c,f}, C = {b,d,e,f}. Then B∪C = {b,c,d,e,f}, A∩(B∪C) = {c}, A0 = {b,d,e,f} = C, A0∩(B∪C) = C ∩(B∪C) = {b,d,e,f} = C.

It will be convenient for us to modify predicate notation: instead of writing {x : x ∈ U and x satisfies ...} we shall write {x ∈ U : x satisfies ...}

3.4 Relative complement

If A and B are two sets, we define the relative complement of B in A as ArB = {a ∈ A : a / ∈ B}.

Example 3.4.1 If

A = {1,2,3,4}

and

B = {2,4,6,8}

then

ArB = {1,3}.

This operation can be easily expressed in terms of intersection and taking the complement: ArB = A∩B0.

3.5 Symmetric difference

The symmetric difference of sets A and B is defined as A4B = (ArB)∪(B rA).

Sets A, B, and A4B.

It can be seen (check!) that A4B = {x : x ∈ A or x ∈ B, but x 6∈ A∩B} and also that A4B = (A∪B)r(A∩B). Example 3.5.1 If A = {1,2,3,4,5} and B = {4,5,6,7} then A4B = {1,2,3,6,7}.

Please notice that the symmetric difference of sets A and B does not depend on the universal set U to which they belong; the same applies to conjunction A∧B, disjunction A∨B, and relative complement ArB; it is the complement A0 where we have to take care of the universal set.

3.6 Boolean Algebra When dealing with sets, we have operations ∩, ∪ and 0. The manipulation of expressions involving these symbols is called Boolean algebra (after George Boole, 1815–1864). The identities of Boolean algebra∗ are as follows. (A, B and C denote * Or “laws” of Boolean algebra. arbitrary sets all of which are subsets of U.)

A∩B = B∩A A∪B = B∪Acommutative laws (1) A∩A = A A∪A = Aidempotent laws (2)

A∩(B∩C) = (A∩B)∩C A∪(B∪C) = (A∪B)∪C associative laws (3)

A∩(B∪C) = (A∩B)∪(A∩C) A∪(B∩C) = (A∪B)∩(A∪C)distributive laws (4)

A∩(A∪B) = A A∪(A∩B) = Aabsorbtion laws (5)

identity laws:

A∩U = A A∪U = U A∪∅ = A A∩∅ = ∅

complement laws:

(A0)0 = A A∩A0 = ∅ U0 = ∅ A∪A0 = U ∅0 = U (7)

(A∩B)0 = A0∪B0 (A∪B)0 = A0∩B0De Morgan's laws (8)

We shall prove these laws in the next lecture. Meanwhile, notice similarities and differences with laws of usual arithmetic. For example, multiplication is distributive with respect to addition: a×(b + c) = (a×b) + (a×c), but addition is not distributive with respect to multiplication: it is not true that a + (b×c) = (a + b)×(a + c).

Notice also that the idempotent laws are not so alien to arithmetic as one may think: they hold for zero, 0 + 0 = 0, 0×0 = 0.