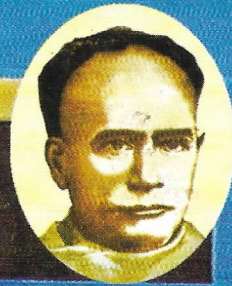


DISTANCE LEARNING MATERIAL



B.V. BADAL DINESH BHAVAN
DIRECTORATE OF DISTANCE EDUCATION



VIDYASAGAR UNIVERSITY
DIRECTORATE OF DISTANCE EDUCATION
MIDNAPORE 721 102

M. SC. IN BOTANY
PART - I

Paper : II (1st & 2nd Half)

Module No. : 13, 14, 15, 17, 18, 18(A), 19, 19(A) & 19 (B)

DIRECTORATE OF DISTANCE EDUCATION

VIDYASAGAR UNIVERSITY

MIDNAPORE -721 102



M.Sc. in Botany

Part-I :: Paper -II (First Half)

Module No. 13, 14, 15, 17, 18, 18(A), 19, 19(A) & 19(B)

M.Sc. in Botany

Part-I :: Paper -II (First Half)

Module No. 13

M. Sc. In Botany

Paper – II (First Half) ● Module No. - 13

1. Introduction Definition of terms: Systematics, Taxonomy, Classification, Nomenclature, Identification

SYSTEMATICS AND TAXONOMY

The terms systematics and taxonomy are synonymous although to a purist these two terms have certain distinctions. While taxonomy is considered to be basically concerned with the study of classification including the principles, rules and procedures, systematics is recognized as a much wider field of study covering diversity of organisms, their relationship, classification and evolution. Simpson (1961) defined systematics as a scientific study of kinds and diversity of organisms, and of any and all relationships between them.

Taxonomy, although the oldest discipline of biological science, was introduced into the field of botanical studies for the first time by A.P de Candolle in 1813 as a combination of taxis (arrangement) and nomes (rules or laws). Taxonomy, once considered to be theories of classification, has widened its aspects and prospects considerably through ages. Definition of taxonomy coincides with systematics and is concerned with identification, nomenclature, classification, interrelationship and evolution of organisms. Stace (1989) defined taxonomy as the study and description of variations in living organisms, the investigation of the causes and consequence of this variation and the manipulation of the data obtained to produce a system of classification.

CLASSIFICATION

Classification is an arrangement of plants into groups on the basis of certain principles. It is a process in one hand and product on the other. As a process it involves certain methods and criteria and as a product it serves as an information storage and retrieval system about the concerned groups or taxa (species, genus, family, order, class, division etc.) Taxonomic entities are classified in different fashions.

1. Artificial system of classification: This is utilitarian, based on arbitrary, easily observable characters, usually one or a few at a time, so that the resulting groups consist of unrelated members. The sexual system of classification given to us by Linnaeus fits into this category.
2. Natural system of classification are those which are based on evaluation of as many characters as possible so that the resulting groups consist of related members. Thus, these systems use overall similarity in grouping taxa. This concept was initiated by M. Adanson (1763) and culminated by Bentham and Hooker (1862-1883).
3. Phylogenetic systems are based on the evolutionary descent of the organisms and their various groups or categories or taxa. However, these systems are natural in construction and evolutionary in expression. The phylogenetic interpretation incorporates information about origin, interrelationship, development and fate of different taxa. System of Engler and Prantl (1887-1915), John Hutchinson (1926, 1930, 1959 and 1973), Armen Takhtajan (1980, 1987, 1997), Arthur Cronquist (1981, 1988), Robert Thorne (1983, 1992, 2000), Rolf Dahlgren (1983, updated in 1989 by his wife Gertur Dahlgren) are phylogenetic with great reputation.

NOMENCLATURE

Nomenclature is the procedure of determination of the correct name for a taxon on the basis of rules and recommendations of the International Code of Botanical Nomenclature (ICBN). Updated every six years or so, the code maintains the system of picking up a single correct name out of numerous scientific names available for a taxon, with particular circumscription, position and rank. To avoid inconvenient name changes for certain taxa, a list of conserved names is provided in the code.

The code (ICBN) aims at provision of a stable method of naming taxonomic groups, avoiding and rejecting the use of ambiguous and erroneous names. The code consists of 3 components:

- I Principles
- II Rules and recommendations
- III Provision for modification of the code.

In addition, the code includes the following appendices:

- I. Names of hybrids
- II. Nomina familiarum conservanda et rejicienda
- III. A. Nomina generica conservanda et rejicienda
- III B. Nomina specifica conservanda et rejicienda
- IV. Nomina utique rejicienda
- V. Opera utique appressa

The last three useful appendices were included for the first time in the Tokyo Code. The first (IIB) includes the specific names conserved and rejected the second i.e. IV the name and all combinations based on these names, which are ruled as rejected under Art. 56, and none is to be used; and the last (V) one deals with the list of publications (and the category of taxa therein) which are not validly published in accordance with the code.

IDENTIFICATION

Determination of the name of a species and its systematic position is basic to all taxonomic methods and pre-requisite to all scientific researches concerning plants and their economic use. Identification of all categories in taxonomic hierarchy needs appropriate characterization and standard authentic keys and for a species particularly comparison with authentic specimens preserved in herbaria (a direct method) is necessary in addition to dichotomous keys, polyclaves or computer-aided devices (an indirect method). In most cases, the characters derived from the gross morphology are adequate for provisional identification of a taxon. Confirmation of such identification may be obtained from other data sources. If the specimen does not agree either with the predetermined specimen or with the literature, then it is to be treated as so far unknown to science. A new name may be given to it following all rules laid by ICBN.

HISTORY OF DEVELOPMENT OF TAXONOMY

The dependence of man on plants for every aspect of life since the dawn of civilization has always induced in him the urge to know more about the plants. Establishment of identity of plants their description and classification led to the development of taxonomy. Medicinal plants were recognized as early as Vedic age. The reputed Atharva Veda, written around 2000 B.C. contains a wealth of information about medicinal plants and their uses. Texts on medicinal plants in ancient times were plentiful being prepared by Chinese, Egyptians, Assyrians, Aztecs and others. Vrikshayurveda (science of plant life) is one of the earliest books

written by Parasara (250 B.C. to 120 B.C.). This treatise was the basis of botanical studies preparatory to medical science in those days.

The ancient Greek philosophers like Plato, Aristotle, Theophrastus and others, who compiled the principles of plant classification in Greek and Latin (around 400 to 285 B.C.) are considered to be founders of taxonomy. Theophrastus is credited with more than 200 works, but most of these survive only as fragments or quotations in the work of other writers. The most famous ones among his writings include Enquiry into Plants and The Causes of Plants. The plants were classified by him into trees, shrubs, undershrubs and herbs and about 480 plants were listed in his Historia Plantarum. Distinctions between centripetal (racemose) and centrifugal (cymose) types of inflorescence, superior and inferior ovaries and the polypetalous and gamopetalous corolla were also made.

One of the most authentic books to have served Europeans for 1500 years from I.A.D. was de Materia Medica authored by Pedanios Dioscorides. This book made a substantial contribution primarily to the then practice of medicine and secondarily to the development of taxonomy by including description and use of about 600 species of plants. It was de Vegetabilis of Albert Magnus (1193-1280) which was very widely used for the next two hundred years. The credit goes to him to distinguish the monocotyledons from the dicotyledons for the first time.

During the fifteenth and sixteenth century many famous books on medicinal plants (Herbals) were published which also contributed to the development of taxonomy through plant descriptions and illustrations. These two hundred years are also branded as the Age of Herbals during which invention of movable type printing and improvement in the science of navigation augmented the growth of knowledge about plants. Otto Brunfels (1489-1534), Jerome Bock (1498-1554), Leonhart Fuchs (1526-1609), Methius de l'obel (1515-1568), John Gerard (1545-1612) and few others are among the herbalists of this age who would be ever remembered for their work on medicinal plants-their use and characterization.

It is from the sixteenth and seventeenth century onwards that the science of botany started developing as an independent discipline. Since taxonomy is basic to all other branches of botany, in the initial stages botany started developing through contributions in the area of taxonomy. There were attempts to study more and more plants and more number of characters in order to arrive at a satisfactory classification.

It was A. Caesalpino (1519-1603), an Italian botanist, to give a classification of plants based on definite morphological features in his famous work de Plantis (1583). He had realized that the flowers and fruits are more reliable in classification than the habit. Further advancement in taxonomy emanated from the endeavours of Jean Bauhin, Gaspar (Caspar) Bauhin, Jhon Ray, J.P. de Tournefort and others. Jhon Ray's Methodus Plantarum (1682) and Historia Plantarum (1686-1704) deserve special mention since he adopted the principle that all the parts of the plants should be considered in classification, a principle now recognized as the cornerstone of a natural system. Tournefort is credited for organization of 698 genera in his Elements de Botanique (1694) many of which were validated by Linnaeus (eg. *Betula*, *Castanea*, *Fagus*, *Quercus*, *Alnus*, *Abutilon*).

SEXUAL SYSTEM-A NEW APPROACH

In 1694 Rudolf Jacob Camerarius experimentally proved sexuality in plants and set a turning point in the trend of plant classification. It inspired Linnaeus to formulate a comprehensive artificial system based on sexual characters, viz. androecial and gynoecial characters. Finding the inadequacy of Tournefort's system (1694) he applied his own system in the second edition of Hortus Uplandicus (1732). This also served the basis of his later publications such as Systema Naturae (1735), Critica Botanica (1737a), Flora Lapponica (1737b), Hortus Cliffortianus (1737c) and Genera Plantarum (1737d), Species Plantarum (1753). His sexual system of classification was very simple and convenient since only one or two characters had served

the basis for the same. Although it failed to show natural affinities of plants, the effort provided the idea for preparation of artificial keys in modern Floras and Monographs. The most significant contribution of Linnaeus to Botany was introduction of binomial system of plant nomenclature. The date 1st May 1753 has been fixed up as the starting point of plant nomenclature.

APPROACHES BASED ON NATURAL RELATIONSHIP

It was Michel Adans on (1727-1806), a French explorer to Africa, who rejected all artificial systems including that of Linnaeus in favour of a natural system. He developed his own system in his book *Families des Plantes* (1763). His system was a natural system of classification wherein he had given two postulates: (i) in classification all characters should be of equal weight and (ii) taxa are to be based on correlation between these attributes. Through these postulates he had sown the seeds of numerical taxonomy, which took about 200 years to get translated into reality due to the efforts of Sokal and Sneath (1963).

The subsequent systems of classification were based on form-relationship. *Genera Plantarum* (1789) of A.L. de Jussieu marked the beginning of a well-planned natural system and its date of publication is officially recognized as the starting point for nomenclature of plant families. Jussieu's system was further developed by Augustin P. de Candolle (1778-1841) in his *Theorie Elementaire de la Botanique* (1813). It was this book through which he introduced the term taxonomy as the theory of plant classification. His monumental work entitled "*Prodrum Systematis Naturalis Regni Vegetabilis*" was initiated in 1824 and he could live to see its seven volumes getting published. Ten more volumes were published by others after his death. The lucky choice of Ranaian group as the starting point in the linear sequence of the families had done much to popularize the system (Davis and Haywood, 1963).

In the immediate post-Linnaean period classification of cryptogams remained almost unattended. In 1813 de Candolle for the first time assembled all pteridophytes in a separate group prior to which *Lycopodium* was classed as moss, *Salvinia* as a liverwort, cycads were kept with Ferns and *Equisetum* among conifers. The first distinction between phanerogams and cryptogams was made by Brogniart (1825) and between angiosperms and gymnosperms by Brown (1827) in his work entitled '*Prodrum Florae Novae Hollandiae*'. Endlicher (1836-1850) divided the plant kingdom into thallophytes (algae, fungi and lichens) and cromophytes (mosses, ferns and seed plants). The bryophytes were recognized by Braun (1859) and the most scientific demarcation among different plant groups was schemed by Eichler (1883).

By the middle years of the 19th century, the philosophy of a natural system of plant classification was thoroughly entrenched in the mind of botanists. The system of de candolle provided the basis to Bentham and Hooker to prepare their *Genera Plantarum* (1862-1883) which is a very useful compendium of generic descriptions arranged in a natural scheme. The generic descriptions therein were models of completeness and precision. Their system achieved a lot of appreciation since the work was mostly based on the plant specimens in the British and continental Herbaria and least on literature. Although the publication of *Genera Plantarum* was later to that of Darwin's *Origin of Species* it did not implement the doctrine of evolution and remained pre-evolutionary in concept. Like all other pre-Darwinian systems reliance was laid on the dogma that the species are special creations and therefore constant and immutable.

PHYLOGENETIC APPROACHES

Post-Darwinian systems were evolutionary in approach and can be arranged into two main groups according to the concept of the primitive angiospermous flower. The first view was proposed by Alexander Braun in 1859 in his *Flora der Provinz Brandenburg*. This influenced his successor A.W.Eicher (1883) in Berlin and was modified by Adolf Engler in his Guide to Breslau Botanic Garden (1886) and fully elaborated in his *Syllabus der Pflanzfamilien*, the first edition of which appeared in 1892 and the latest i.e. the 12th edition

in 1964 after having been modified by Melchior. Engler in collaboration with Karl Prantl, expanded the system in *Die Natürlichen Pflanzenfamilien* (1887-1915). This system covers the whole plant kingdom and surpasses all other systems in its quantum of information. It has utilized information emanating from embryology, morphology, anatomy, geographical distribution and presented adequate illustration, bibliography of pertinent literature and keys to identification.

The Englerian concept considers unisexual and naked flowers borne in separate aments or catkins as the most primitive state which have progressively been elaborated into bisexual diplochlamydeous flowers. Bisexual flowers were derived from a single cluster of male and female flowers held in the same inflorescence that simulated a flower (pseudanthium), while the subtending perianth evolved later.

The most primitive flowers were wind-pollinated like the cones of gymnosperms and were derived from a gymnospermous ancestor with unisexual strobilus bearing either micro- or mega-sporophylls. Melchior's revision of Engler's *Syllabus* (1964) involved profound changes and rearrangements. Dicots were treated as more primitive than monocots, various orders were splitted and families there in were rearranged. Nevertheless, the essence of the Englerian concept of primitive flowers is maintained in original form.

Engler's concept received support from Wettstein (1901), Hayata (1921), Pulle (1938, 1950), Rendle (1904, 1925, 1930, 1938), Skottsberg (1940) and others who had also framed their own systems keeping the axial theme more or less similar.

The non-Englerian concept is familiar as the Ranaian concept. The pioneer architect of this concept was Charles E. Bessey (1883) of the University of Nebraska whose final system of classification (1915) received admiration from a number of taxonomists including Cronquist for strobilus bearing in its axis megasporophylls above and microsporophylls below. The lower sporophylls were transformed into sepals and petals through sterilization. The primitive angiospermous flowers have many free perianth members; numerous, free stamens and carpels arranged spirally on the long conspicuous thalamus. The nearest approach to this type of floral organization is made by the extant *Magnolia*. Arber and Parkin (1907) provided the theoretical basis for this view. The angiosperms owe their origin to a so far unknown gymnosperm related to the rare fossil *Cycadeoidea dactenensis*, a Mesozoic plant of Bennettitales that had borne 'anthostrobili' in the axis of cycad like leaves.

A somewhat similar type phylogenetic system was received from Hans Hallier (1905, 1908, 1912) whose independent endeavour concerned an exhaustive survey of literature, herbarium specimens and palaeobotanical evidences. This system of classification, according to Takhtajan, was more synthetic and had a deeper insight into the morphological evolution and phylogeny of angiosperms, arose from an unknown and extinct tribe of cycads that was related to Bennettitales and had affinities with Marattiales. Monocots were thought to be evolved from an extinct dicot ancestral to Lardizabalaceae.

Jhon Hutchinson (1926, 1934, 1959, 1964-67, 1973), a British botanist, proposed a system of classification resembling Bessey's but differing in certain fundamental aspects. Hutchinson derived the flowering plants from a hypothetical proangiosperm and divided them into three lines: the Monocotyledons, Herbaceae, Dicotyledones and Lignosae. Dicotyledones. Monocots were believed to arise from herbaceous Raniales. The woody line (lignosae) was derived from the woody Magnoliales and Herbaceae from Raniales.

Although Hutchinson's treatment of individual families and genera and especially that of the monocotyledons is excellent, creation of woody and herbaceous lines has resulted into alignment of unrelated taxa and done much to detract from overall value of his work.

MODERN APPROACHES

A number of taxonomists henceforth got involved in improving schemes of classification based on new information from various sources. Data from palaeobotany, phytochemistry, ultra structure, numerical

taxonomy, serology etc. have helped in enriching taxonomy, solving many problems related to circumscription, phylogeny and evaluation of utilitarian aspects of plants. Some of the noteworthy taxonomists and their contributions are discussed in the following.

The philosophical tradition of Bessey and Halliers phylogenetic systems of classification had an impact on Armen Takhtajan (1954, 1958, 1966, 1969, 1980, 1987 and 1997) and Arthur Cronquist (1957, 1968, 1981 and 1988), R. Dahlgren (1975, 1980, 1983) and R.L. Thorne (1976, 1981, 1983, 1992, 2000). The system of Takhtajan has similarity in fundamental aspects with that of Cronquist and so is the system of Dahlgren with that of Thorne. An updated revision of classification of Dahlgren was published by his wife Gertrud Dahlgren (1989). Takhtajan's recent system of classification (1997) recognizes 11 subclasses of dicots and 6 of monocots. These are further subdivided into superorders, orders and families; 56 super orders, 175 orders and 459 families have been put under dicots and 16 superorders, 58 orders and 133 families under monocots. Arthur Cronquist on the other hand had recognized 6 and 5 subclasses of dicots and monocots respectively. Unlike Takhtajan, Cronquist did not recognize superorders. The Dicots were directly divided into 64 orders and 318 families, and monocots were divided into 19 orders and 65 families. Whereas Takhtajan attaches more importance to cladistics Cronquist gives more importance to phenetic relationship in the placement of various groups.

The realignments in Dahlgren's system (1987) are based on a large number of characteristics mainly phytochemistry, ultra structure and embryology. It includes 25 super orders in dicots and 8 in monocots. The diagram, which is a cross section of the phylogenetic tree passed across the top, is very useful for mapping character distribution in various groups which may go a long way in developing an ideal system of classification.

R. F. Thorne's system of classification (1992) was revised by him in 2000 after having realized that much new information has become available about the classification of Angiospermae, especially in the currently popular fields of cladistics; micro-morphological and molecular taxonomy. His recent efforts were aimed towards bringing into light the up-to-date knowledge about monocots (2000a) and dicots (2000b), keeping his philosophy essentially the same. Thorne retained Angiospermae as a class and had modified considerably the subclasses of Cronquist and Takhtajan because of their polyphyletic nature. Thorne divided angiosperms into ten, hopefully monophyletic subclasses, viz Magnoliidae, Ranunculidae, Caryophyllidae, Dilleniidae, Rosidae, Asteridae, Lamiidae under dicots and Liliidae, Alismatidae, and Commelinidae under monocots.

PHENETIC APPROACHES

New approaches started developing over the last few decades in search of greater objectivity for better understanding of taxonomy. One of such efforts is numerical taxonomy or phenetics, which has evolved from availability and development of electronic computers in the late 1950s and in the 1960s (Sneath, 1957 a and b; Michener and Sokal, 1957; Sokal and Sneath, 1963). It involves numerical evaluation of the affinity between taxonomic units and the ordering of these units into taxa on the basis of their affinity. A number of standard publications (Sokal and Sneath, 1963; Rohlf and Sokal, 1965; Sneath and Sokal, 1973; Neff and Marcus, 1980; Duncan and Baum, 1981; Gordon, 1981; Dunn and Everitt, 1982; Felsenstein, 1981, 1983; Legendre and Legendre, 1983; McNeil, 1983) came out in succession till the early eighties, which have provided useful taximetric techniques and shown their applications. The numerical methods in general are not specifically sensitive to convergent evolution, sibling species or to isolating mechanisms.

CLADISTIC APPROACHES

Another objective approach emanated in form of phylogenetic systematics due to the endeavors of W Hennig (1966) which was termed cladistics by Mayr (1969) and developed by Nelson and Patnick (1981) and Bremer and Wanntorp (1981). Since mid 1970s many botanists particularly those of USA started working in this direction. An American botanist, W.H.Wagner, working independently on Hawaiian ferns developed a cladistic method familiar as ground-plan divergence method. In principle cladistics attempts to analyse phylogenetic data objectively very much like numerical taxonomy which analyses phenetic data. Cladistic methods are largely based on the principle of parsimony according to which the most likely evolutionary route is the shortest hypothetical pathway of changes that explains the pattern under observation. Journals like Cladistics (since 1985), Advances in Cladistics (since 1981) are being published and society called 'Willi Hennig Society' has been established. Though there is a claim of superiority of principle and practices of cladistics, it needs much improvisation in the future.

BIOSYSTEMATIC APPROACHES

A new discipline of biology concerning development of taxonomy in relation to population sampling and experimental procedures emerged in form of Biosystematics, which was introduced, by Camp and Gilly in 1943. Biosystematics has been providing a rich source of understanding of taxa and reassessment of their circumscription, interrelationship and classification on the basis of analysis of variation pattern and genetic relationship.

HISTORY OF BOTANICAL NOMENCLATURE

Although Casper Bauhin (1737) introduced the concept of binomical nomenclature the credit for establishing this system of naming a species goes to Carolus Linnaeus. This system has been totally followed in his book entitled "Species Plantarum" for which its date of publication i.e. 1st May 1753 has been officially recognized as the starting point of plant nomenclature. The early rules of this system were set forth by Linnaeus in his Critica Botanica (1737). It was A.P. de Candolle (1813) who had given explicit instructions on nomenclatural procedures through his book *Theorie Elementaire de la Botanique*. The pioneer taxonomist to have prepared an Index of latin names of the then known plants was Steudel. His creation "*Nomenclator Botanique*" (1821) is still remembered today with gratitude and honour for having taught us the importance of an Index for standardization of nomenclature and also for having given the raw materials for present day Indices.

The first organized effort for standardization of plant nomenclature was made by Alphonse de Candolle who had given the leadership for holding the first International Botanical Congress at Paris in 1867. The rules adopted therein is familiar as **Paris Code**. For not having been satisfied with the Paris code, American Botanists adopted a separate **Rochester Code** (1892) which introduced the concept of types. The **Paris code** was replaced by **Vienna code** (1905) which declared "Species Plantarum" of Linnaeus (1753) as the **starting point**, rejected tautonyms and made Latin diagnosis essential for new species. Moreover, a list of conserved generic names (**Nomina Generica Conservanda**) was approved. Being dissatisfied with the Vienna Code also, adherents to the Rochester code adopted the American code (1907) which rejected the list of conserved names as well as the need for Latin diagnosis.

It was not until the 5th International Botanical Congress held at Cambridge (1930) that differences in opinions were removed and a truly International Code evolved, accepting the type concept, rejecting the tautonyms, making Latin diagnosis compulsory for concerned taxon in nomenclatural proposals and approving conservation of generic names. The code is being reviewed periodically and amended in the Botanical Congress. The 15th International Botanical Congress was held at Tokyo in 1993 and the 16th in St

Louis in 1999 Revision of the code is based on the realization that the system of nomenclature should be simple, comprehensive and precise so that it proves useful to all in all theoretical as well as applied aspects

PHASES OF TAXONOMIC STUDIES

Taxonomic studies concerned with characterization and classification of living entities is accomplished through the following phases.

- I The pioneer or exploratory phase: This is the most important pre-requisite for taxonomic studies since it is a natural science and is based on field survey or exploration. During field work the species are identified at least provisionally.
- II The consolidation phase: The specimens procured from the field are brought to the Herbarium not only for preservation but also for preparing their descriptions and confirming identification of each species in consultation with authentic herbarium specimens and literature.
The first two phase are very much associated with preparation of Flora, Distributional Records, Monographs, Reviews etc.
- III The Biosystematic phase: This phase is concerned with indepth study of taxa based not only on the above two phases but also on geographical variations and correlative ecological features, genetic experiments, cytological observations, physiological features of samples representing populations of a species.
- IV The Encyclopaedic phase: This is compilation and co-ordination of the above-mentioned phases. The first two phases are mainly descriptive and based on gross morphological features as a result of which these correspond to the "alpha" taxonomy and the last two phases dealing with greater number of data sources correspond to "omega" taxonomy of Turrill (1938). Thus, the distinction lies in the depth of knowledge (Davis and Heywood, 1963)

AIMS OF TAXONOMY

Taxonomy, which is mainly concerned with identification, nomenclature and classification, can be said to have the following aims:

- 1 to provide a standard procedure of identification,
- 2 to provide the basis for scientific communications about various aspects of plants and animals,
- 3 to classify on the basis of nature affinities of organisms as far as possible,
- 4 to assemble the knowledge of plant recourses so as to implicate the same to our well-being,
- 5 to prepare an inventory of living organisms for use in flora, fauna or stock-taking of biodiversity and
- 6 to have provisions for imparting training to biology students and scholars so that they can develop complementarity with the synthetic approaches of modern taxonomy.

PROCEDURES OF TAXONOMY

These are three aspects of taxonomic procedure, viz. identification, nomenclature and classification. This is basic to all disciplines of biological science and is concerned with study and description of variations in living organisms for which it necessitates through field observations and perfect morphological description. The thorough morphological description and the voucher specimen collected from field are used in identification of the species for reference to the keys available in literature and matching with the authentic specimens preserved in Herbaria respectively. These conventional procedures are associated with classical taxonomy.

Since taxonomy has undergone modernization through its synthetic approach, procedures to reveal the embryological, palynological features etc. of the species get concerned with characterization and

classification of plants and interpretation of their evolution.

BIOSYSTEMATICS

Definition: The systematics of living organisms concerned with development of taxonomy through establishment of variation pattern on the basis of population sampling is known as biosystematics. Biosystematics has also been termed in various other ways, viz. experimental taxonomy, modern taxonomy, neosystematics, genenomy, genecology and so on.

Variation - phenotypic plasticity

Biosystematics is concerned with the study and description of variations in the populations of living organisms. The variation is considered as the outcome of interaction of genes to a specific ecological conditions and a population is certain to show certain characters that can be correlated with its habitat characteristics.

In biosystematics it is very essential to determine whether the variation is permanent or temporary. In the former case the variation results from genetic changes induced by the environment.

On the other hand if the variation in populations growing in different ecological conditions disappear when grown together in the same habitat and environment the case is treated as phenotypic plasticity. In cases of phenotypic plasticity the ecological isolation is so meager that these hardly lead to genetic drift and speciation. Such populations are termed ecads or ecophenes and may be considered forerunner of genetic variations.

Principles of biosystematics

1. Biosystematics is largely concerned with cytogenetic and ecological aspects of taxonomy and involves studies in the field as well as experimental gardens and laboratories. In broad terms it appears that biosystematics covers in one hand the neo-systematics taking into account genetic and evolutionary nature of groups and the micro evolutionary phenomena on the other hand.
2. It is based on the concept of reproductive isolation and genetic structure of populations or biotic units.
3. It relies upon the descriptive phase of taxonomy and for delimitation of different categories it incorporates evidences from morphology and various other sources as genetics, ecology, physiology, biochemistry, embryology, palynology etc.

Procedures

Biosystematic studies are performed in these steps:

1. Thorough sampling of the taxon and its populations: It involves field studies to identify populations showing variation in correlation with their environment. Voucher specimens are collected for preservation and their propagules (seeds, rhizomes, tubers etc.) for cultivation in experimental garden for further cytogenetic anatomical. Palynological, chemical studies etc. The field observations are very carefully recorded.
2. Herbarium study: A preliminary knowledge about the range variations within a species can be obtained from the study of the specimens of each species preserved in herbaria. This would also enable detection of discontinuities in population samples. The information drawn from herbarium study can be correlated with other kinds of studies. The discontinuities are used to identify genetic isolation and eventually delimitation of taxa. The observations have to be confirmed by experimental studies.
3. Hybridization: The ability of the different population systems to hybridize is studied by growing (cultivating) their representatives in experimental garden under uniform conditions. The next step

involves the study of vigour and fertility of the hybrids. This enables identification of genetic barrier (reproductive isolation) developed between population systems.

4. Cytological study: The homology of chromosomes in the hybrids is determined on the basis of observation on their behaviour during meiosis.
5. Compilation and categorization: The information or data obtained from the above steps is compiled with data from comparative morphology and geographical distribution. After compilation, the data is used in assigning the population systems the respective biosystematic category.
6. Modern methods: Since genetic experiments in the experimental garden are time consuming and inconvenient, method for study of genetic relationship have started concerning semantides, i.e. information containing molecules such as DNA-DNA hybridization, DNA-RNA hybridization, nucleotide sequencing, comparative protein sequencing, allozyme comparison, serology etc.

RELATION WITH TAXONOMY

Biosystematics, although a very young discipline, is essentially an expansion of classical taxonomy (Davis and Heywood, 1963). It still employs comparative morphology as a primary means of describing and defining taxa. For delimitation and discrimination of taxa it uses genetic experiments and pays much emphasis on responses of genetic structure of populations to different ecological conditions. Thus, biosystematics has the potential to improvise, revise and enrich classical taxonomy and transform it into a synthetic science. Biosystematics in general, has been of tremendous help in (i) delimiting the natural biotic units and (ii) application of nomenclature to these units in such a way so that it conveys precise information regarding their defined limits, relationships, variability and dynamic structure (Camp and Gilly, 1943). However one should very carefully welcome biosystematics in the existing system of nomenclature. Biosystematics can certainly prove helpful in the understanding of interrelationship and interpreting microevolution.

BIOSYSTEMATIC CATEGORIES

Biosystematics have developed certain categories for experimentally investigated biotic units. The most widely accepted categories, in order of ascending phyletic value are ecotype, ecospecies, cenospecies and comparium.

Ecotype: It is the unit in biosystematics adapted to a particular environment but capable of producing fully fertile hybrids with other related ecotypes. Such ecotypes are inseparable by genetic barriers, although they develop certain variations in correlation with their environments.

Ecospecies: A group formed by association of one or more ecotypes, which are able to exchange their genes without extending detriment to the offsprings. Related ecospecies, on hybridization usually produce sterile F_1 hybrids. Thus such ecospecies are separated by ecological barrier in one hand and partial genetic barrier precluding free exchange of genes on the other.

Cenospecies: A group formed by uniting one or more ecospecies of common evolutionary origin. Related cenospecies rarely produce sterile hybrids. Thus genetic barrier between related cenospecies is complete.

Comparium: A group formed by similar cenospecies that do not cross or hybridize.

BOTANIC GARDENS AND HERBARIA

Gardens which are specifically designed and sustained to sponsor academic as well as beneficial aspects of plants are categorized as botanical gardens. The first garden for the purpose of science and education was maintained by Theophrastus in Lyceum, Athens, probably bequeathed to him by his teacher, Aristotle.

Credit for establishing a full-fledged botanical garden goes to Luca Ghini (1490-1556) who developed it at Pisa in Italy in 1544. This was followed by establishment of such gardens at Padua and Florence of Italy in 1545. At present there are thousands of botanic gardens all over the world and are looked after by different institutions. Of these about 800 gardens are documented in the International Directory of Botanical Gardens. A botanical garden today is a green area maintained by an organization for growing various types of plants so as to fulfill aesthetic, conservational, educational, recreational, utilitarian and various scientific purposes. Although botanical gardens house plant species which the climate of the area can support, there are many gardens which have controlled systems to sustain species collected from different phytogeographical and eco-climatic regions. Glasshouses in temperate regions provide adequate warmth to sub-tropical and tropical species through what is known as manipulated green house effects. Similarly in tropical gardens, specially cooled indoor growing houses are maintained for sustaining cold loving plants.

Botanical gardens have always been the major source of inspiration as well as information to many taxonomists to develop systems of classification almost since the time of Theophrastus. Linnaeus while working at Uppsala and Bernard de Jussieu at Versailles created the then the most rational artificial and natural systems of classification respectively.

Botanical gardens play the following important roles:

1. **Aesthetic value-** The plants of a botanical garden always induce aesthetic pleasure, e.g. the Great Banyan Tree in the Indian Botanical Garden, Sibpur.
2. **Supply of material/specimens for botanical research.**
3. **Imparts training and teaching to general people, students and scholars about taxonomy, economic botany and ecology.**
4. **Supports research projects**
5. **Conservation of threatened species:** Botanical gardens are important in conserving genetic diversity in general and endangered species in particular (*ex situ conservation*). The proceedings of the Symposium on Threatened and Endangered species, sponsored by New York Botanical Garden in 1976 ("Extinction is forever") and the conference on the practical role of botanical gardens in conservation of rare and threatened species sponsored by the Royal Botanical Gardens, Kew ("Survival and Extinction") have brought to our knowledge the role of botanical gardens in conservation.
6. **Seed Bank:** Most of the botanical gardens act as Seed Banks and more than 500 botanical gardens of the world operate seed exchange programmes, offer annual lists of available species and freely exchangeable seeds.
7. **Academic Institution and Documentation centres:** Most of the botanical gardens have a linkage with a herbarium and library and collectively play very important academic role.
8. **Public Services:** Botanical gardens provide information to the general public on identification of native and exotic species, methods of propagation and also planting material through sale or exchange.
9. **Ecological role:** Botanical gardens optimize the environmental state even of the nearby cities or townships and act very much as the human lungs to refresh the air.
10. **Introduction of Germ-plasm:** Germplasm of many economic plants are introduced from other countries and the species are initially acclimatized prior to their spread within the country for cultivation.

Some of the world famous botanical gardens are enumerated in the following

1. Botanic Garden of Padua, Padua, Italy (1545)
2. Botanic Garden of Heidelberg, University of Heidelberg, Germany (1593)
3. National Museum of Natural History, Paris, France (1635)
4. Botanical Garden and Museum, Berlin, Germany (1646)
5. Royal Botanic Garden, Edinburgh, Scotland (1670)
6. Botanic Garden of the Academy of science, Leningrad, Russia (1712)
7. University Botanic Gardens, Cambridge, England (1762)
8. Indian Botanic Garden, Howrah, India (1787)
9. Botanic Garden of Harvard University, Cambridge, Mass., USA (1807)
10. Botanic Gardens, Munich, Germany. (1809)
11. Royal Botanic Garden, Kew, England (1841)
12. Missouri Botanical Garden, St. Louis, Mo., USA (1859)
13. Arnold Arboretum of Harvard University, Jamaica Plains, Mass., USA (1872)
14. The New York Botanical Garden, New York, USA (1890)
15. Montreal Botanic Garden, Montreal, Canada (1936)

HERBARIUM

The herbarium is a repository of plant specimens arranged in the sequence of a standard natural system of classification. According to Fosberg (1946) herbarium is a great filing system for information about plants, both primarily in form of actual specimens of the plants and secondarily in form of published information, pictures and recorded notes.

Herbaria are integrated with academic institutions, Botanical gardens, Research institutions, scientific societies etc.

The Index Herbariorum Part I includes the names of all Herbaria of distinction and Part II of the same publication includes the name of reputed plant collectors and present location of their specimens. Some of the world famous herbaria are listed below with the respective acronyms within brackets

Royal Botanic Garden, Kew (K)

V. L. Komarov Botanical Institute, Leningrad (LE)

Museum National d'Histoire Naturelle, Paris (P)

British Museum (Natural History), London (BM)

Central National Herbarium, Sibpur, Howrah, India (CAL)

Harvard University, Cambridge, USA (A+FH+GH)

Some of the famous herbaria in India besides CAL are mentioned in the following

1. Herbarium of the Industrial Section at the Indian Museum, Calcutta
2. National Botanical Garden, Lucknow
3. Blatter Herbarium, Bombay
4. Eastern Circle Herbarium, BSI, Shillong
5. Western Circle Herbarium, BSI, Pune
6. Southern Circle Herbarium, BSI, Coimbatore
7. Herbarium of the Forest Research Institute Dehra Dun
8. Northern Circle Herbarium, BSI, Dehra Dun
9. Central Circle Herbarium, BSI, Allahabad

Importance of Herbarium in taxonomic studies

- 1 Herbaria present pictures of local, regional and world flora through preserved plant specimens.
- 2 It serves as the chief basis for taxonomic researches aiming towards preparation of Reviews, Monographs, Flora, Manual etc.
- 3 Since herbaria are repository of type specimens critical nomenclatural reassignment can be made.
- 4 It is concerned with time-based documentation of local flora
- 5 The students section of Herbarium as well as the entire Herbarium is useful in imparting training to students of plant sciences, forest officers, conservators and others interested in plant sciences.
- 6 Herbaria not only serves taxonomists but also the economic botanists, ethnobotanists, morphologists, geneticists etc.
- 7 Herbaria are concerned with the identification of specimens submitted to them in one hand and by giving authentic specimens on loan.
- 8 Herbaria are always associated with Libraries enriched with different types of taxonomic literature including pictorial atlas, albums, field notes and records, icons, encyclopaedia, documentary microfilms etc. As such Herbaria are research institutions which are well organized to augment different types of botanical research and academic accomplishment.

Model Questions:

Answer the following (of one mark each)

1. What is Flora?
2. What is ICBN?
3. Who was the author of the book "de Materia Medica"?
4. Who proved sexuality in plants and when?
5. Who introduced the term taxonomy to plant sciences?
6. Define phenetics.
7. Define cladistics.
8. What was the name of the book written by Bentham and Hooker?
9. Name any book of Englar & Prantl.
10. What is a Herbal?
11. Define ecotype.
12. Mention the name of any internationally famous botanical garden of India with its locations.
13. Which date is officially declared as the starting point of plant nomenclature?
14. Name one famous botanical garden outside India and mention its location.
15. Name one internationally famous Herbarium and mention its location.
16. What is alpha taxonomy?
17. What is omega taxonomy?
18. What is Paris Code?
19. What is meant by 'ex-situ' conservation?
20. What is meant by 'age of herbals'?

Questions of 2 marks each

1. Distinguish between systematics and taxonomy
2. What were the postulates given by Michel Adanson?
3. Mention the characters of the most primitive angiosperm flower according to Engler.
4. What is Biosystematics? Why is it often termed genecology?
5. Which International Botanical Congress was for the first time truly international? Where was it held?
6. Distinguish between the terms phenetics and cladistics.
7. Why is numerical taxonomy called neo-Adansonian classification?
8. What are the contents of ICBN?
9. How many subclasses of dicots and monocots were recognized by Takhtajan (1997) and Thorne (2000)?
10. Name two journals on cladistics.

Questions of 5 marks each

1. Write notes on:
 - a) Nomenclature
 - b) Classification
 - c) Principles of biosystematics
 - d) Biosystematic categories
 - e) Taxonomic procedure
2. What is taxonomy? State its aims.
3. Explain phenotypic plasticity.
4. Elucidate the relationship of biosystematics with taxonomy.
5. Explain the Ranaian concept of primitive flowers.

Questions of 10 mark each.

1. What is biosystematics? Discuss different steps to such studies.
2. What is Herbarium? Discuss its importance in taxonomic studies.
3. What is Botanical Garden? Discuss its importance / functions.
4. Discuss in brief history of botanical nomenclature.
5. Prepare a concise account on phylogenetic, phenetic, cladistic and biosystematic approaches.

M.Sc. in Botany

Part-I :: Paper -II (First Half)

Module No. 14

M.Sc. Botany
Part - I Paper - II (1st half)
Module No. - 14

NUMERICAL TAXONOMY

Organisms are classified on the evidence obtained from their characters, therefore, it becomes necessary to employ all the characters for the ideal or a natural classification. But since each individual may possess thousands of characters, it becomes impracticable to use all characters and as Mayr has remarked, the number is limited by the patience of the investigator. This naturally leads to the problem of selection of suitable characters.

Taxonomists such as Bentham and Hooker, Engler and Prantl, Tipso, Hutchinson and others have employed morphological and anatomical characters their major source of evidence while more recent workers including Takhtajan, Cronquist, Thorne, Dahlgren and others have based their classifications on relatively large number of attributes. It is for this reason that most of these later classifications have gained a wider acceptance. They are more natural.

The use of as many characters as possible, or ideally, all the characters for classification was proposed by Adanson (1763) and such classifications are called Adansonian classifications. The two important postulates are: (i) in constructing the classification each attribute selected is of equal weight and (ii) taxa are based on correlations between these attributes.

The Adansonian principles have received great support since the 1960s and have developed new methods in taxonomy included under a general term **numerical taxonomy**. It involves the numerical evaluation of the affinity or similarity between taxonomic units and the ordering of these units into taxa on the basis of affinities. This is essentially an extension of the Adansonian classification using mathematical procedure with the primary aims of repeatability and objectivity (Davis and Heywood, 1963). There are two aspects of such procedures – the construction of taxonomic groups and discrimination.

Construction of Taxonomic Groups

In the past, taxonomic groups were constructed using characters by some form of weighting – physiological importance of character in the pre-Adansonian taxonomy and phylogenetic importance in the post – Darwinian period or intuitive correlation weighting. In angiosperms, due to the absence of reliable fossil evidence, phylogenetic weighting becomes largely negative. Phenetic classification, on the other hand, is based on overall affinity as judged by using as many characters or as much evidence as is available. In practice it employs correlation weighting by mechanical methods of comparison. Various techniques are involved and newer ones are being incorporated (Sokal and Sneath, 1963; Estabrook and Rogers, 1966; Rohlf, 1965; Rohlf and Sokal 1965; Bonner, 1964; Crawford and Wishart, 1967 and many others). The successive steps recognised by Sneath (1962) are :-

1. Operational Taxonomic Units (OTUs)

The fundamental unit that can logically be classified by numerical methods is the individual organism. But generally it is not possible to use numerous individuals of the same species of each of several taxonomic groups to compute a classification. Further, such a study would reveal resemblance's at the intraspecific level and usually would not offer much scope for higher levels. It is, therefore, customary to employ species as a unit for this purpose. Since the taxonomic units employed in numerical methods are not always comparable to formal taxonomic units, they are termed as operational taxonomic units (OTUs).

If a numerical taxonomic study of higher categories is to be undertaken, a higher taxon which represents a cluster of various polymorphic tax should be employed as the OTU. Another solution is to use only a single representative of the polymorphic group.

2. Unit Characters

A unit character has been defined as a taxonomic character of two or more states, which within the study at hand cannot be subdivided logically, except for subdivision brought about by changes in the method of coding (Sokal and Sneath, 1963). Further, only the phenotypic characters are used for this basic information. Thus, the presence or absence of an awn in a grass spikelet may be a unit character. The organisational level for unit characters may differ from character to character. But as a rule, each character state should contribute one new item of information.

The proper selection of characters is a critical point in the application of numerical taxonomy, as it is in other disciplines of taxonomy. Certain characters are clearly disqualified for numerical taxonomy and these are listed by Sokal and Sneath (1963) as inadmissible characters. According to these authors, it is

undesirable to use : (i) attributes which are not a reflection of the genotypes of the organisms themselves; (ii) any property which is a logical consequence of another, either partly or wholly; and (iii) characters which do not vary within the entire sample of organisms.

A large number of characters at least more than 50, must be selected. Several hundreds will be more significant but too few characters are not reliable. These characters have then to be coded or given some symbol or mark.

- (a) **Two -state coding :** This is the simplest form of coding where characters are divided into + and - or as 1 and 0. The positive characters are recorded as + or as 1 and negative characters as - or as 0. In case the organ possessing a given characters is missing in an organism. the character must be scored NC. In other words, the symbol NC means "no comparison".
- (b) **Multi-state coding:** The multi-state characters may be either quantitative or qualitative. Quantitative characters can each be expressed by a single numerical value, e.g. amount of pubescence on a leaf. Such characters can be coded into number of states (1,2,3 ...) corresponding to their range of variation. Quantitative multi-state characters cannot be arranged in some order, and no reliable sequence can be established. In such cases, qualitative characters are conveniently converted into some new characters. Many a times, it is convenient to convert multi-state characters into two-state characters. In additive coding, for example, multiple characters with four states could be coded as

Two state characters			
	1	2	3
Multiple states	0-	-	-
	1+	-	-
	2+	+	+
	3+	+	+

In this way, a multistate character of n states is turned into n-1 two-state characters. The obtained by scoring the characters in OTUs are then arranged in a table in a matrix form as above and compared.

3. Measurement of Resemblance

There have been three methods devised so far for estimating phenetic resemblance between the taxonomic groups, namely (i) co-efficients of association; (ii) co-efficients of correlation; and (iii) measures of taxonomic distance using the convention of multidimensional space with the dimension for each characters.

4. Cluster Analysis

a taxonomic system can be constructed from the resemblances among the OTUs. To form taxa different OTUs are grouped together on the basis of affinities found by measurement of resemblances. These group of OTUs are termed clusters.

Clustering is achieved in two ways : (i) by employing the attributes one at a time- monothetic systems, or (ii) according to all their attributes considered simultaneously. The monothetic method obviously leads to artificial clustering while the second method gives a natural grouping. One cluster or a group is separated from the other by the dividing line which indicates a distinct gap between the two. It is often easy to separate categories of higher rank, i.e. above the genus level, clearly. In case of intraspecific clusters, there are often few discontinuities.

There are several techniques to describe structure in matrices of similarity co-efficient. One of the common techniques is the **differential shading of the similarity matrix**. In this method, similarity co-efficients are grouped into five to ten evenly spaced classes. Each of these classes are represented by different degrees of shading in the squares of half matrix. The highest value is generally shown darkest and the lowest value lightest, as in the following Figure. Then the half matrix can be seen as a pattern to different shades, limited by a diagonal of squares with the darkest shade. Thus, on rearrangement of the sequence of OTUs, clusters can be more sharply defined, as in the Figure below.

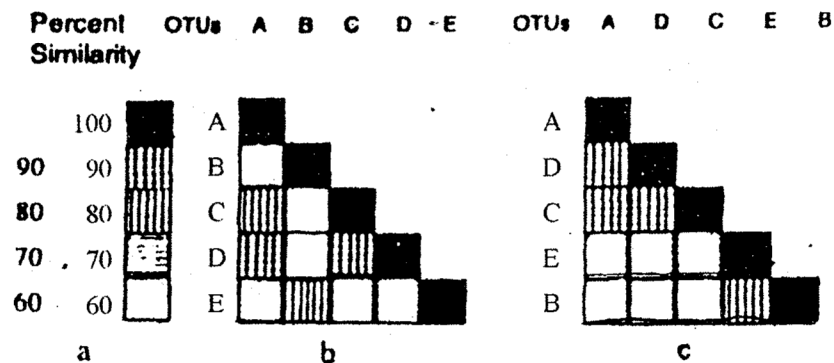


FIG. :- Shaded similarity matrices : (a) percentage similarity. (b) with the OTU's arranged haphazardly. (c) after rearrangement of OTU's.

The groups of related OTUs based on high similarity coefficients can be analysed by a large number of numerical techniques such as **elementary cluster analysis**, clustering by single, complete or average linkage, central or nodal clustering and so on.

The groups of similar organisms recognised in this manner are termed **phenons**. The clusters of phenons are then rearranged in a dendrogram which summarises the main features of the cluster analysis.

5. Phenons and Rank

The groups or clusters established by numerical methods may be equivalent with those of classical taxonomic methods, i.e. usual rank categories such as genus, tribe or family. But these terms have evolutionary and nomenclatural background. If this is to be avoided, Sokal and Sneath (1962) have provided a new expression – phenons. Their level of affinity is indicated by prefacing them with a number. A group affiliated at 80 and above in the similarity scale may thus be termed 80-phenon. In figures, clusters with different shade indicate 60 to 90 phenons. These terms are intended to cover the groups produced by any form of cluster analysis or from any form of similarity co-efficients. Although phenons may be equivalent to various taxonomic groups, the term “phenon” is not synonymous with “taxon”.

The delimitation of phenons is done by drawing a horizontal line across the dendrogram at a similarity value. A line at 75%, for example, creates five 75-phenons 11, 7, 3, 5, 6, 4, 9, 10 and 2, 8; while that at 80% creates six 80-phenons. Such a dendrogram will have a reference to a given taxon and cannot be transferred to any other study. In the following dendrogram, if OTUs 1 to 10 had been species, an 80-phenon line could indicate 6 subgenera and 65-phenon line two genera. It should however be remembered that phenons are arbitrary and relative groups.

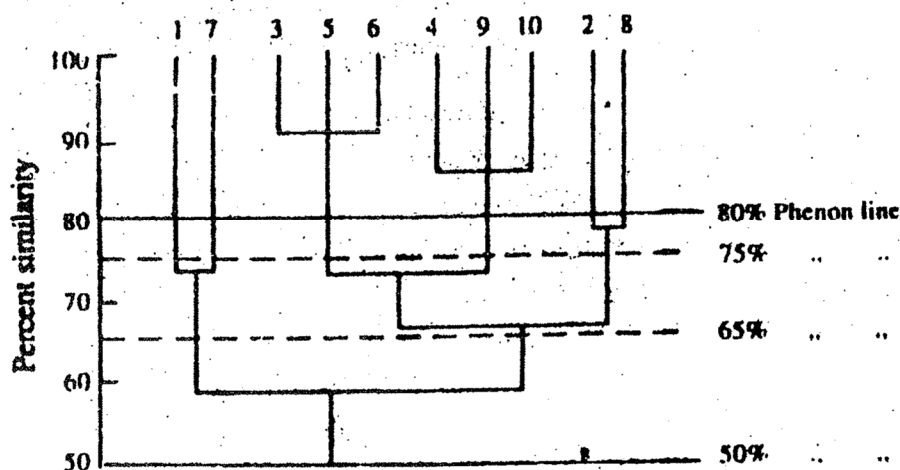


FIG. Dendrogram to show formation of phenons.

Discrimination

If taxonomic groups chosen for the study show overlapping of characters, discrimination should be used to select them. Various techniques, such as discriminant analysis, have been devised for such purposes. The best methods for delimiting tax are based on the utilisation of maximum number of characters with similar weightage given to them.

Nomenclature and Numerical Taxonomy

Modern nomenclature does not concern itself with the problems of delimitation of taxa. It serves

only as a reference point to the taxonomic names. The limits are debatable, subjective and forever changeable. Numerical taxonomy, on the other hand, is very useful in delimitation of taxa by exact estimation of affinities (although phenetic). Thus, there is no scope for "personal opinion" or "decision of taxonomists". The limits may be objective, utilitarian, permanent and fixed by common consent.

Applications of Numerical Taxonomy

The taxonomists who are interested in the study of similarities and differences are now using numerical methods on an increasing scale. There has been considerable work on bacteria, other micro-organisms and several animal groups, using numerical methods. However, its application in plant taxonomy has not been comparable to that in other groups. From angiosperms, genera such as *Oryza* (Morishima and Oka, 1961), *Solanum* (Soria and Heiser, 1961), *Sarcostemma* (Johnson and Holm, 1968) and other groups including Farinosae of Engler (Hamann, 1961) and a few others have been tried by numerical taxonomy for their delimitation. The results, in general, have been in conformity with the earlier works based on classical methods. It is only in certain taxa (Farinosae-Hamann, 1961) that the earlier assemblages have been shown to be unnatural. Numerical taxonomic studies on the genus *Ononis* (Papilionaceae) by Cook (1969) are in confirmation with the system proposed earlier by Sirjaev (1932).

In general, the results achieved so far by numerical taxonomy, have not been appreciably different from those achieved by other methods. It is probably because the methods employed need much improvement. A colloquium held at Andrew's University, Scotland in 1968 brought out several modifications in methods, and since then many have attempted to reclassify a number of plant taxa without any appreciable deviation from the formal taxonomic classification.

Merits and Demerits of Application of Numerical Methods

According to Davis and Heywood (1963), it would be better to welcome these procedures with caution, since these methods are only an extension of the orthodox procedures. They have raised some doubts. First, the methods will clearly be useful in phenetic—classifications, not phylogenetic. Similarly, the proponents of "biological" species concept, may not accept the specific limits bound by these methods. Even the practicing taxonomist might use his brain more efficiently than the machine which is fed with non-relevant selection of characters. Character selection is the weak link in this approach. The statistical methods are likely to give less satisfactory solution of characters chosen for comparison are inadequate.

Stearn (1968) indicated that different taxonomic procedures may yield different results. A major difficulty for the beginner is to choose a procedure for his purpose. Another difficulty concerns the number

of characters (from 40-100) needed in order to obtain satisfactory results by these mechanical aids. Taxonomists usually manage with far less characters. Further, so far it has not been seen that the results achieved by mechanical means are in any way more acceptable than those visualised by practicing taxonomists. It is desirable to ascertain whether a large number of characters would really give satisfactory results than those using a smaller number. Stearn (1964), after applying taxonomic procedures to the Jamaican species of *Columnnea* and *Alloplestus* (Gesneriaceae), came to realise that it seemed a pity not to make further use of these. This survey, as Stearn (1968) has concluded, demonstrated the capacity of computer-aided taxonomic methods to build from an assemblage of characters a grouping of species comparable in validity to one made by conscious taxonomic effort. It also indicated that the number of characters used is less important than their range.

Johnson and Holm (1968), after analysing their data on the genus *Saccostemma* (Asclepiadaceae) by various taxonomic methods, have concluded that the numerical classification based on correlation coefficient bears closer resemblance to the classical taxonomic classification. However, they expressed their view that thorough analysis of character sets will lead to a better understanding a process of evolution and the role of environment in determining patterns of variation.

Dale (1968), while presenting the basic procedures which underline numerical taxonomic methods, concluded that any taxonomist proposing to use such methods must be careful in his choice and be wary of what may seem to be unimportant details.

Cullen (1968), while reviewing the botanical problems of numerical taxonomy, extended his welcome to the advance of numerical techniques which, according to him, may well provide means of checking and improving classifications by orthodox taxonomists. He has also realised that the numerical classifications are not likely to supplant orthodox ones – they may either conform to them or, if very different, exist side by side with them.

Clifford (1970) seems to have used numerical methods for a better classification of the grasses and concluded that there is a greater probability of their evolution from the palms. This conclusion was also drawn by Meeuse (1966) on the basis of the ovarian structure and origin. In both palms and grasses, the ovary has an ecarpellate origin, with greater reduction in the grasses than the palms. Corner (1966) also has indicated that the embryos of *Archontophoenix* possess a coleorrhiza as in grasses, a feature otherwise unknown in palms.

QUESTIONS

1. Define Numerical Taxonomy. Described briefly, the successive steps as recognised by Sneath (1962) for undertaking such studies. Comment on the merits and demerits of Numerical Taxonomy.

REFERENCES FOR MODULE NO. 14

1. Taxonomy of Angiosperms – V.N. Naik (1991), Tata McGraw-Hill Publishing Company Limited, New Delhi.
2. Introduction to the Principles of Taxonomy – V.V. Sivarajan (1991), Oxford & IBH Publishing Co. PVT. LTD., Calcutta.
3. Systematic Botany – S.C. Datta (1991), Wiley Eastern Ltd., Calcutta.
4. Plant Taxonomy and Biosystematics – C.A. Stace (1994), Edward Arnold, London.
5. Plant systematics – S.B. Jones and A.E. Juchsing (1979), McGraw-Hill, New York.
6. Chromosomes – A. Sharma (1991), Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi.

M.Sc. in Botany

Part-I : : Paper -II (First Half)

Module No. 15

M.Sc. Botany

Part - I Paper - II (1st half)

Module No. - 15

STUDY OF CLASSIFICATION

Taxonomy is the study and description of the variation of organisms, the investigation of the causes and the manipulation of the data obtained to produce a system classification. Thus classification is the most important integer of Taxodemy and its history dates back to the dawn of plant science.

Classification is a process in one hand and object or end product on the other. As a process it is the production of a logical system of categories each containing a certain number of organisms which allows easier reference to the components i.e. kinds of organisms. As an object classification is that system itself of which there are many sorts (Stace, 1989). In simpler terms classification can be defined as the ordering of organisms into groups and arranging these groups in a system of hierarchy so that it can serve as a information storage as well as a retrieval system.

There has been an inherent tendency in man to identify and study plants almost since the dawn of civilisation since plants have been meeting various necessities of his life. The rich botanical literature authored by Chinese, Assyrians, Aztecs, Egyptians and Indians long before the Christian era have been discovered. One of the earliest of these is attributed to the legendary emperor of China-Shen nong who is reputed to have lived about 3600 years before the Christian era. Another botanical literature entitled Cheng lei pen tsao presented by Tang Shen Weig earned a lot of popularity. The well-known *Atharva Veda*, written in the Vedic age i.e, around 2000 B.C. is a treasure house of information on medicinal plants and their uses. 'Vrikshayurveda' is another book of ancient India in which the author Parashara propounded a system of classification based on comparative morphology of plants besides giving an account of external morphology, nature and properties of soil, 14 forest types in India, internal structure of leaves etc. A large number of families (Ganas) were so clearly defined that they are even recognisable today (Majumdar, 1927, 1946). These gems of botanical literature are poorly known in the West and the work in Greek and Latin are considered to have laid the foundation of botany.

Classifications Based on Habit

Theophrastus (ca. 370-285 B.C.) was the greatest botanical writer of classical antiquity and he is regarded as the intellectual grandfather of modern botany (Greene, 1909). The writings of Theophrastus on which a large quantum of his reputation rests include 'Enquiry into plants' and 'Historia Plantarum'. He classified plants on the basis of their habit into trees, shrubs, undershrubs, herbs; distinguished between centripetal (racemose) and centrifugal (cymose) inflorescence, superior and inferior ovary and polypetalous and gamopetalous corolla.

A somewhat similar classification was given by Pliny, the Elder (Caius Plinius Secundus) in 'Historia Naturalis' (77 A.D.). Pedanios Dioscorides, a contemporary of Pliny gave information about 600 species of plants together with their medicinal use in a book entitled 'Materia Medica' in 1 A.D. that served the need for the next 1500 years.

The thousand years or so from the decline of Rome or the Renaissance was a period when little academic accomplishment was there in Europe. In such an era of general ignorance Albert Magnus was the best naturalist who became popular as the 'Aristotle of the Middle Ages'. His classification (1193-1280)

recognised monocot and dicot plants as Corticatae and Tunicatae respectively His 'de Vegetabilis' was widely used for the next 200 years.

Age of herbals

With the advent of Renaissance the cultural and intellectual fabric of the West altered. There was a great quest for knowledge as a consequence of which fifteenth and sixteenth century botany experienced a general awakening in the field of medicine and plant characterisation in form of illustration and description. Many world famous 'Herbals' appeared during these two centuries which earned fame as the 'Age of Herbals'. Invention of movable type printing in 1440 in Europe afforded compilation of books on medicinal plants (Herbals). The then development in navigation technology facilitated exploration and concomitant expansion of botanical knowledge. Noteworthy contributions, though indirect, were received from herbalists like Jerome Bock, Otto Brunfels, Leonhart Fuchs, P. Mathioli, Carolus Clusius, Matthias de L'obel, John Gerard and many others whose intentions were to prove the novelty and superiority of their respective Herbals over others. However, the classifications designed by them had very little systematic basis.

Purely Botanical Approaches

From the sixteenth and seventeenth century onwards the botanists switched over from the utilitarian approach to that of a scientific one. It was Andreae Caesalpino (1512-1603), an Italian botanist, to give a classification of plants based on definite morphological features in his famous work *de Plantis* (1583). He had realised that the flowers and fruits are more reliable in classification than the habit. Further advancement in taxonomy emanated from the endeavours of Jean Bauhin, Gaspard (Caspar) Bauhin, John Ray, J.P. de Tournefort and others. John Ray's *Methodus Plantarum* (1682) and *Historia Plantarum* (1686-1704) deserve special mention since he adopted a principle that all parts of the plants should be considered in classification – a principle now recognised as the corner-stone of a natural system. Tournefort is credited for organisation of 698 genera in his *Eléments de Botanique* (1694) many of which were validated by Linnaeus (e.g. *Betula*, *Castanea*, *Fagus Quercus*, *Ulmus*, *Abutilon*).

Use of Sexual Characters in Classification

The systems of classification framed since Theophrastus till the end of the seventeenth century were artificial based mainly on habit of plants. In 1694 Rudolf Jacob Camerarius experimentally proved sexuality in plants and set a turning point in the trend of plant classification. It inspired Linnaeus to formulate a comprehensive artificial system based on sexual characters viz., androecial and gynoecial characters.

Finding the inadequacy of Tournefort's system (1694) he applied his own system in the second edition of *Hortus Upplandicus* (1732). This also served the basis of his later publications such as *Systema Naturae* (1735), *Critica Botanica* (1737a), *Flora Lapponica* (1737b), *Bortus Cliffortianus* (1737c) and *Genera Plantarum* (1737d), *Species Plantarum* (1753). His sexual system of classification was very simple and convenient since only one or two characters had served the basis for the same. Although it failed to show natural affinities of plants, the effort provided the idea for preparation of artificial keys in modern Floras and Monographs. The most significant contribution of Linnaeus to Botany was introduction of binomial system of plant nomenclature. The date 1st May, 1753 has been fixed up as the starting point of plant nomenclature.

Natural System of Classification

The inadequacy of Linnaean system was realised as soon as botanists started accumulating plant wealth from different parts of the world. Michel Adanson (1727-1806), a French explorer to Africa, rejected all artificial systems including that of Linnaeus in favour of a natural system. He developed his own system in *Familles des Plantes* (1763) which is remembered for the use of as many characters as possible and two postulates: (i) in constructing classification each attribute selected is of equal weight and (ii) taxa are based on correlation between these attributes. Through his work he had sown the seeds of phenetic system which after spending dormancy period of 200 years germinated into numerical taxonomy mainly due to the efforts of Sokal and Sneath (1963). Historians of Science note that Adanson foreshadowed what is now called phenetic taxonomy.

The subsequent systems of classification were based on form-relationship and attempted to express natural relationship among taxa. *Genera Plantarum* (1789) of A. L. de Jussieu marked the beginning of a well planned natural system and is still honoured for perfection in organisation of many orders (=modern families) and its year of publication i.e. 1789 is marked out as the starting point for nomenclature of plant families. Jussieu's system was further developed by Augustin Pyramide de Candolle (1778-1841) in his *Theorie Elementaire de la Botanique* (1813). He is remembered for introducing the term 'Taxonomy' as the theory of plant classification. He was the pioneer in organising pteridophytes as a separate group, although erroneously as a co-ordinate group with monocots. His monumental work entitled '*Prodrum Systematis Naturalis Regni Vegetabilis*' was initiated in 1824 and he could live to see only its seven volumes getting published. Ten more volumes were published by others after his death. The lucky choice of Ranalian group as the starting point in the linear sequence of the families had done much to popularise the system (Davis and Heywood, 1963).

In the immediate post-Linnaean period classification of Cryptogams remained almost unattended. In 1813 de Candolle for the first time assembled all pteridophytes in a separate group prior to which

Lycopodium was classed as a moss, *Salvinia* as a liverwort, cycads were kept with Ferns and *Equisetum* among conifers. The first distinction between phanerogams and cryptogams was made by Brongniart (1825) and between angiosperms and gymnosperms by Brown (1827) in his work entitled 'Prodrumus Florae Novae Hollandiae'. Endlicher (1836-1850) divided the plant kingdom into Thallophytes (algae, fungi and lichens) and Cormophytes (mosses, ferns and seed plants). The bryophytes were recognised by Braun (1859) and the most scientific demarcation among different plant groups was schemed by Eichler (1883).

By the middle years of the 19th century, philosophy of natural system of plant classification was thoroughly entrenched in the mind of botanists. The system of de Candolle provided the basis to Bentham and Hooker to prepare their *Genera Plantarum* (1862-1883) which is a very useful compendium of generic descriptions arranged in a natural scheme. The generic description therein were models of completeness and precision. Their system achieved a lot of appreciation since the work was mostly based on the plant specimens in the British and Continental Herbaria and least on literature. Although the publication of *Genera Plantarum* was later to that of Darwin's *Origin of Species* it did not implement the doctrine of evolution and remained pre-evolutionary in concept. Like all other pre-Darwinian systems reliance was laid on the dogma that the species are special creations and therefore constant and immutable.

Phylogenetic Systems of Classification

Post Darwinian systems were evolutionary in approach and can be arranged into two main groups according to the concept of the primitive angiospermous flower. The first familiar as Englerian concept, was proposed by Alexander Braun in 1859 in his *Flora der Provinz Brandenburg*. This influenced his successor A.W. Eichler (1813) in Berlin and was modified by Adolf Engler in his *Guide to Breslau Botanic Garden* (1886) and fully elaborated in his *Syllabus der Pflanzenfamilien*, the first edition of which appeared in 1892 and the latest i.e. the 12th edition in 1964 after having been modified by Melchior. Engler in collaboration with Karl Prantle, expanded the system in *Die Natürlichen Pflanzenfamilien* (1887-1915). This system covers the whole plant kingdom and surpasses all other systems in its quantum of information. It has utilised information emanating from embryology, morphology, anatomy, geographical distribution and presented adequate illustrations, bibliography of pertinent literature and keys to identification.

The Englerian concept considers unisexual and naked flowers borne in separate aments or catkins as the most primitive state which have progressively been elaborated into bisexual, diplochlamydous flowers. Bisexual flowers were derived from a single cluster of male and female flowers held in the same inflorescence that simulated a flower (pseudanthium), while the subtending perianth evolved later.

The most primitive flowers were wind pollinated like the cones of gymnosperms and were derived from a gymnospermous ancestor with unisexual strobilus bearing either micro- or mega-sporophylls,

Melchior's revision of Engler's *Syllabus* (1964) involved profound changes and rearrangements. Dicots were treated as more primitive than monocots, various orders were splitted and families there in were rearranged. Nevertheless, the essence of the Englerian concept of primitive flowers is maintained in original form.

Engler's concept received support from Wettstein (1901), Hayata (1921), Pulle (1938, 1950), Rendle (1904, 1945, 1930, 1938), Skottsberg (1940) and others who had also framed their own systems keeping the axial theme more or less similar. Among these systems that of Richard von Wettstein (1901) is appreciated most since it is phylogenetically more rational. He considered dicots to be more primitive than monocots the latter having been derived from Ranalian stocks. Woody plants were regarded by him as more primitive than herbs, many flowered inflorescence to be more ancient than solitary flowers and spiral flowers than cyclic ones.

The non-Englerian concept is familiar as the Ranalian concept. The pioneer architect of this concept was Charles E. Bessey (1883) of the University of Nebraska whose final system of classification (1915) received admiration from a number of taxonomists including Cronquist for having considered Ranales as the most primitive order derived from a gymnosperm with a bisexual strobilus bearing in its axis megasporophylls above and microsporophylls below. The lower sporophylls were transformed into sepals and petals through sterilisation. The primitive angiospermous flowers had many free perianth members; numerous, free stamens and carpels arranged spirally on the long conspicuous thalamus. The nearest approach to this type of floral organisation is made by the extant *Magnolia*. Arber and Parkin (1907) provided the theoretical basis for this view. The angiosperm owe their origin to a so far unknown gymnosperm related to the rare fossil *Cyca Jexileia dacotensis*. A Mesozoic plant of Bennettiales that had borne 'anthostrobil' in the axial of its cycad like leaves.

A somewhat similar type phylogenetic system was received from Hans Hallier (1905, 1908, 1912) whose independent endeavour concerned an exhaustive survey of literature, herbarium specimens and palaeobotanical evidences. This system of classification, according to Takhtajan, was more synthetic and had the deeper insight into the morphological evolution and phylogeny of flowering plants than other contemporary systems including that of Bessey. Hallier believed that angiosperms arose from an unknown and extinct tribe of Cycaos that was related to Bennettiales and had affinities with Marattiales. Monocots were thought to be evolved from an extinct dicot ancestral to Lardizabalaceae.

John Hutchinson (1926, 1934, 1959, 1964-67, 1973), a British botanist, proposed a system of classification resembling Bessey's but differing in certain fundamental aspects. Hutchinson derived the flowering plants from a hypothetical proangiosperm and divided them into three lines: the Monocotyledons, Herbaceae Dicotyledones and Lignosae Dicotyledones. Monocots were believed to arise from herbaceous Ranales. The woody line (Lignosae) derived from the woody Magnoliales and the Herbaceae from Raniales.

Although Hutchinson's treatment of individual families and genera and especially that of the monocotyledons is excellent, creation of woody and herbaceous lines has resulted into alignment of unrelated taxa and done much to detract from overall value of his work.

The philosophical tradition of Bessey and Halliers' phylogenetic systems of classification had an impact on Armen Takhtajan (1954, 1958, 1966, 1969, 1980, 1987). A Cronquist (1957, 1968, 1981), R. Dahlgren (1975, 1980, 1983), R.L. Thorne (1968, 1976, 1983), G.L. Stebbins (1974) and others who have used data from anatomy, palaeobotany, biochemical systematics, serology, embryology, cytogenetics, Scanning and Transmission Electron Microscopy etc. in conjunction with information from the traditional source i.e. morphology so as to give us a phylogenetically fertile and substantial systems of classification. It was Oswald Tippo (1942) and Karl Mez (1926, 1936) to have developed systems of classification based on anatomy and serology respectively. K.R. Sporne (1976, 1980) has sought for an objective methodology for phylogenetic interpretation. He prepared a list of characters significantly correlated among dicot families and performed simple statistical tests to determine Advancement Index for each family. On the basis of 30 pairs of such characters Sporne (1980) ascribed lowest and highest Advancement Indices to Magnoliaceae (25) and Dipsacaceae (87) respectively. Although the catalogue of increasing advancement index is in no way linear sequence of evolution it can improve existing classification and settle arguments in respect of taxonomic and phylogenetic relationship. Sporne presented a bird's eye view of evolution in form of a phylogram drawn in a circle, the radius representing Advancement Index and position on circumference the likely degree of divergence. The Danish botanist Rolf Dahlgren (1975) originally prepared a phytoqram with living orders placed at the topmost plane of the diagram on the ancestral branches (absent in Sporne's diagram) many of which are extinct. In his later schemes (1977, 1983) he dropped the branching portion of the phylogram and used only the topmost plane. Dahlgren follows the view that none of the extant groups of flowering plants is ancestral to any other present day group. The Magnoliaceae-Ranunculaceae group is not ancestral to others but it has simply retained many primitive characters. Similar representation is also noticed in Thorne's work (1983). Dahlgren *et al.* (1985) divided dicots into 25 super orders and monocots into 10 superorders. Robert F. Thorne (1983) created 19 and 9 superorders for dicots and monocots respectively. Although claimed to have been original the novelty of the systems published during the last decade or so is mootable since they are deeply rooted in the substrate of Takhtajan and Cronquist's systems of classification.

The system published by Takhtajan in 1954 has undergone modifications and elaborations in his later publications but the framework and essence remained the same. He divided Magnoliophyta (Angiosperms) into two classes viz., Magnoliopsida (Dicots) and Liliopsida (Monocots). The former was subdivided into 7 subclasses and 20 superorders and the Liliopsida was subdivided into 3 subclasses and 8 superorders. Monocots have been shown to have undergone parallel evolution with Nymphaeales from a common Magnoliale having monosulcate pollen and wood without vessels. His system is based on clear

evolutionary principles and rational putative relationship. The excellency of Takhtajan's system lies in his attempts to solve satisfactorily many problems related to polyphyly or monophyly primitiveness of Magnoliales, secondary nature of anemophilous families with reduced flowers in aments etc. Keeping Magnoliales at the base the evolution of dicots and monocots along with their putative relationships have been shown very efficiently. Although it can not claim to have made the final arrangement of the orders and families particularly in subclasses Rosidae and Liliidae the system has many positive points to its credit. Takhtajan's system seems to be very complicated mainly due to creation of the superorders between subclasses and orders and recognition of small families and orders in pursuit of sharp definition of groups. He (1980) recognised 419 families out of which 342 were dicots and 77 monocots representing 71 and 21 orders respectively. Moreover derivation of monocots from stocks ancestral to Nymphaeales receives objection from many corners Stebbins, (1974) was of the opinion that the similarities between Nymphaeales and Monocots is probably a consequence of convergent evolution and the ancestry of them are completely obscured in the geological history. There is an objection to Takhtajan's idea of South-east Asia as the cradle of angiosperms'. It is now believed that the places of North Gondwana particularly between South Africa and South America where gradually appeared the Atlantic Ocean were the centre of origin of angiosperms.

The system of classification of flowering plants as designed by Arthur Cronquist (1981) is in similar philosophical dogma and is a refinement over Takhtajan's system. In full justification of the title "An Integrated System of Classification of Flowering Plants" he (1981) integrated all available data from different sources avoiding undue optimism about the present importance of data that reflect a very sparse sampling. He has used classical data with continuing expansion of data-base for chemical and micromorphological features. The system is compatible with the fossil record and is most comparable with that of Arman Takhtajan. Although both Takhtajan and Cronquist agree on main outlines of angiosperm evolution, there are certain differences in close decisions at micro-level. Takhtajan placed more reliance on serology and less on other chemical characters than Cronquist. Unlike Takhtajan Cronquist (1981) was reluctant to assign the status of family or orders to small groups which could be moderately and comfortably accommodated in larger groups. As a consequence the number of families were reduced by Cronquist to 318 and 65 in case of classes Magnoliopsida and Liliopsida respectively. The former class was subdivided into 6 subclasses and 64 orders and the latter class into 5 subclasses and 19 orders. The basic difference with Takhtajan's system lies in its simplicity and natural assemblage, the submergence of Ranunculidae into Magnoliidae, rejection of superorders, cleavage of Liliidae *sensu* Takhtajan into Liliidae, Commelinidae and Zingiberidae. However, at some levels too much reliance on certain single character like centrifugal stamens, free central placentation etc. hampers the otherwise natural classification. The arrangement of families in Liliales, particularly with reference to submergence of Amaryllidaceae into Liliaceae, although tentative, has been a subject of criticism (Treib, 1974, 1975).

Biosystematic approaches :

A new discipline of biology concerning development of taxonomy in relation to population sampling and experimental procedures emerged in form of Biosystematics which was introduced by Camp and Gilly in 1943 as 'Biosystematy'. This type of study provides a rich source of understanding of a taxa and reassessment of their circumscription, interrelationship and classification. Several categories of experimentally investigated biotic units are in vogue, the most popular among which is Turesson's (1922) terms used by Clausen et al. (1940), viz. ecotype, ecospecies, cenospecies, comparium in order of ascending phyletic value. Biosystematics aims to prepare itself to acquire such dimensions so as to prove its superiority over the conventional ones and to set up a system of classification providing all types of data with equal fidelity.

In conclusion it may be said that systems of plant classification have been undergoing evolution due to a progressive increase in the resolving power of taxonomy through ages. The synthetic approach in taxonomy has set up new windows with the help of cytogenetics, embryology, ecology, mathematics, chemistry, palynology, serology etc. to let in more light on our conception about the taxonomic entities and more fresh air so as to activate us in understanding their phylogeny. No system of classification till date can claim to have made the final arrangement of taxa. However, if collaboration among various disciplines persists the endeavour is sure to offer unequivocal way out to the quandaries faced by traditional taxonomy and yield a perfect system of classification.

References :

- Arber, E.A.N. and Parkin, J. 1907. On the origin of angiosperms. *J. Linn. Soc. Bot.* 38; 29-80.
- Bessey, C.E. 1883. Evolution and classification. *Bot. Gaz* 18 : 329-332.
- 1897. Phylogeny and taxonomy of living plants. *Bot. Gaz.* 24 : 145 - 178.
- 1915. The phylogenetic taxonomy of flowering plants. *Ann. Missouri Bot. Gard.* 2 : 109-164.
- Bremer, K. and Wanntrop, H. 1981. The cladistic approach to plant classification in : *Advances in Cladistics* (V.A Funk and D.R. Brooks eds.). The New York Botanic Garden, New York.
- Brongniart, A.T. 1825. *Essai d'une classification des champignons*. Paris.
- Clausen, J.Kech, D.D. and Hiesey, W.M. 1940. Experimental studies in the nature of species. I. Effect of varied environment on Western North American Plants Carnegie Inst. Wash. Publ. 520, Washington, D.C. pp. 452.
- Cronquist, a. 1957. Outline of a new system of families and orders of dicotyledons. *Bull. Jard. Bot. Brux.* 27: 13-40.

Advent of Numerical Taxonomy

New approaches have developed over the last few decades in search of greater objectivity for better understanding of taxonomic entities and their relationship. One of such efforts is numerical taxonomy which (Sneath, 1957 a & b; Michener and Sokal and Sneath, 1963). It involves the numerical evaluation of the affinity between taxonomic units and the ordering of these units into taxa on the basis of their affinity. This is often viewed as an extension of Adansonian classification using mathematical procedure with the primary aims of repeatability and objectivity (Davis and Heywood, 1963). Numerical taxonomy covers all aspects of quantitative studies in systematics, that utilise multivariate techniques. It had its origin in the phenetic approach to classification (Sokal and Sneath, 1963) which from the description by Jardine and Sibson (1971) and the discussion in relation to alternative approaches by Mc Neill (1980) appears to be a natural or general purpose classification in which a large number of general statements can be made about the taxa recognised. Although numerical taxonomy has developed in the context of phenetic philosophy, it is not restricted to the construction of phenetic classification. It has integrated itself with biosystematics. A number of standard publications (Sokal and Sneath, 1963; Rolf and Sokal, 1965; Sneath and Sokal, 1973; Neff and Marcus, 1980; Duncan and Baum, 1981; Gordon 1981; Dunn and Everitt, 1982; Felsenstein, 1982, 1983; Legendre and Legendre, 1983; Mc Neil 1983) are at our disposal that provide useful introduction to taximetric techniques and their applications. The numerical methods, in general, are not specifically sensitive to convergent evolution, sibling species or to isolating mechanisms.

Cladistic Approaches :

The other objective approach emanated in the field of phylogenetic systematics due to the endeavours of W. Hennig (1966) which was termed cladistics by later workers (Nelson and Platnick, 1981; Bremer and Wilmott, 1991). Since mid 1970's many botanists particularly those of USA started walking in this direction. A separate but similar approach came in form of Wagner's (1980) 'Ground-plasm Divergence' method based on his work on Hawaiian ferns. This method attempts to analyse phylogenetic data objectively in a manner parallel to that in taximetrics which seeks to introduce objectivity to natural system. The relationship among plants is assessed on the basis of their evolutionary behaviour and principle of parsimony i.e. the shortest hypothetical pathway of changes that explains the present phenetic pattern. In the mean time journals like *Cladistics* (since 1985), *Advances in Cladistics* (since 1981) are being published and a society called Willi Hennig Society has been established. Though there is a claim of its superiority of principles and practices of classification, no subject can rival cladistics in the degree of argument and controversy.

DIRECTORATE OF DISTANCE EDUCATION

VIDYASAGAR UNIVERSITY

MIDNAPORE 721102



M.Sc. In Botany

Part- I :: Paper II (Second Half)

Module No.- 17, 18, 18 (A)

Bioinformatics and

Computer Applications

M.Sc. In Botany
Part- I :: Paper II (Second Half)
Module No.- 17

Contributor - Prof. P. C. Dhara

Module No. : 17

Module Structure

1.0 INTRODUCTION

2.0 AIM OF THE MODULE

3.0 BASIC STRUCTURE OF COMPUTER

3.1. Hardware

3.2 Software

4.0 COMPUTER LANGUAGE

4.1 Low level language

4.2 High level language

5.0 DESKTOP

6.0 WINDOWS '98

6.1 Main features of Windows '98

6.2 Window Control

7.0 MY COMPUTER

8.0 RECYCLE BIN

9.0 INTERNET

9.1 Modem

9.2 ISP

9.3 World Wide Web (WWW)

9.4 Browsing, Working and Downloading

10.0 BIOINFORMATICS

10.1 Fields of Application of Bioinformatics

10.2 Searching the Internet

10.3 Software frame works for bioinformatics

10.4 Programs and programming languages

10.5 Running programs over the Internet

10.6 Database management in bioinformatics

11.0 CONCLUSION

12.0 SUMMARY

13. BIBLIOGRAPHY

14.0 MODEL QUESTIONS

14.1 Short questions

14.2 Long questions

Computer application in biology and bioinformatics

1.0 INTRODUCTION

People tried to use computing devices for a long time ago. Modern computer has become a successful computing machine as a result of continuous efforts of the scientists. In 1950s modern computers came into the market and they were used only in limited areas. Now, computers become part and parcel of the human life.

A computer may be described as an 'information processor'. There are several advantages of using computers.

- i) Perform complex and repetitive calculation rapidly and accurately.
- ii) Storage of large amounts of information for subsequent manipulation.
- iii) Make decisions.
- iv) Provide information to the user.
- v) Draw and print graphs
- vi) Converse with users through terminals.

Computers are now affecting every sphere of human activity and bringing about many changes in industry, government, education, medicine, scientific research, law, social sciences and even in arts like music and painting. These are presently used, among other applications, to

- * Design buildings, bridges and machines.
- * Assist in railway reservation.
- * Control inventories to minimize material cost.
- * Grade examination and process results.
- * Aid in teaching.
- * Control space vehicles.
- * Play games like chess and video games.

Nowadays computer are being applied in all the branches of sciences including biological sciences.

2.0 AIM OF THE MODULE:

The areas of applications of computer are confined only by limitations of human creativity and imagination. In fact any task that can be carried out systematically, using a precise step-by step method, can be performed by a computer. Therefore, it is essential for every educated person today to know about a computer, its strength, its weakness and its internal structure.

Earlier, computers were used in the fields of engineering and technology. However, for the last few years computers are being used in biology and medicine. There is enough scope to utilize the computer in various fields of biology and medicine. Therefore, the biologists, medical practitioners and medical oriented staffs should be computer literate. Like other branches of biology, the benefits of the computer are utilized in Botany. It can be used to solve various problems in different fields of botany, like, taxonomy, plant physiology, forestry, plant biochemistry, genetics etc. Creation of botanical database is one of the important field applications of computer.

The objective of this module is to delineate important characteristics of computer hardware, software and the way of using computer in biological science. We shall deal with the basic architecture of the computer, different software, operating system, components of desktop, internet, web based information and an idea of bioinformatics.

3.0 BASIC STRUCTURE OF COMPUTER:

A computer can be broadly divided into two parts:

- a) Hardware
- b) Software

3.1. Hardware: Hardware refers to the physical structure of a computer, that is, all of the electronic and electro-mechanical devices associated with a computer. The physical components of the computer are included in the following two divisions:

- i) Central Processing Unit (CPU).
- ii) Peripheral devices.

The architecture of a computer is in Fig. 1.

3.1.1 Central Processing Unit (CPU): It is the central part of the computer. It can be compared to the brain of a human body. The CPU is connected to all other parts of the computer. It controls input, output, memory and other electronic circuits of the computer. The computational works which are done through the computer are executed by the CPU. The Central Processing Unit has the following subunits: (i) control Unit (ii) Arithmetic Logic Unit (iii) Memory.

3.1.1.1 Control Unit (CU): It is the supervisor of the CPU. This unit coordinates the activities of all other units within the system. The important functions of this unit are as follows: (a) It controls the transfer of data and other information among different subunits of CPU. (b) It helps to initiate and supervise the functions of arithmetic-logic unit. A user gives the instruction through computer program which is stored in the memory. The control unit takes those instructions from the memory, decodes them and directs the different units to execute the specific instructions. For example, in a user's program there is an instruction to multiply A with B. The control unit will fetch the values stored in the memory locations A and B and send them to arithmetic-logic unit. The CU will direct the arithmetic-logic unit to perform the multiplication with the values of A and B and to send the result to the memory. It will also send the result to the output unit if there is instruction in the user's program. Thus the control unit performs supervisory functions of the CPU.

3.1.1.2 Arithmetic-Logic Unit (ALU): This subunit is an electronic circuit of CPU. It performs all arithmetic as well as logical functions of the computer program. The arithmetic function includes addition, subtraction, multiplication, division, and exponentiation. Various logical operations, e.g., greater than, lesser than, equals to, not equals to, greater than equals to, lesser than equals to etc are also executed by the logic circuits of ALU.

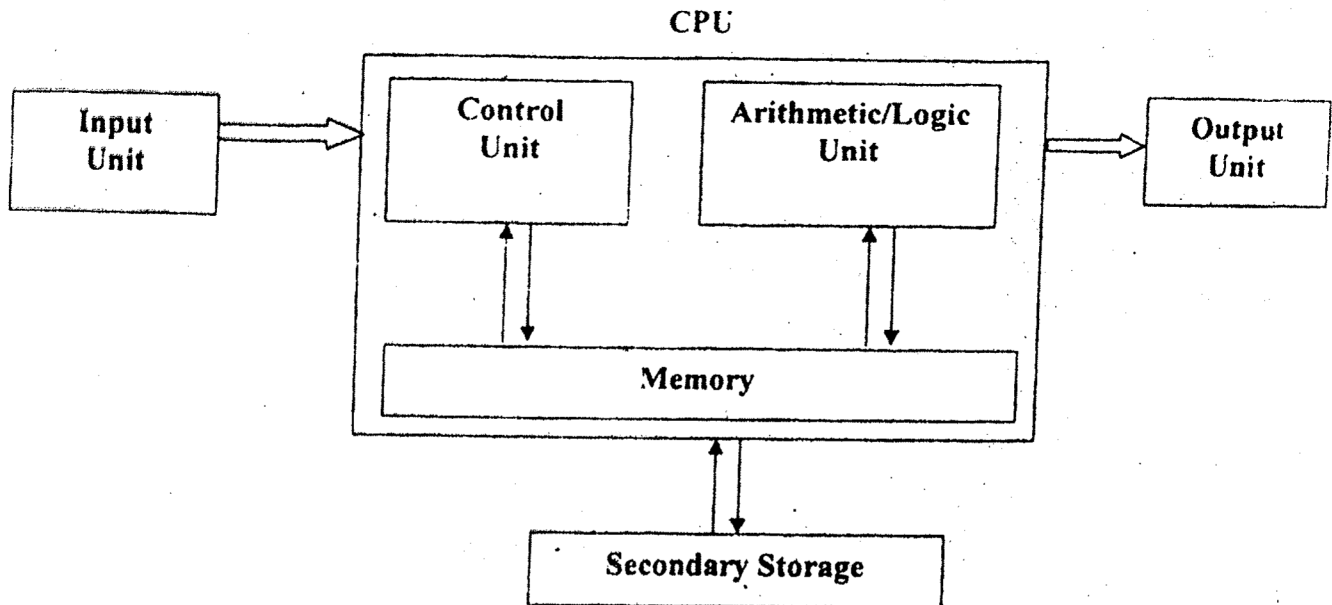


Fig. 1: Architecture of a computer.

3.1.1.3 Memory Unit: This is another important subunit of CPU. The memory unit can store or hold all data, instructions, results etc. The memory consists of numerous cells called storage locations each capable of storing one character or bit. The cells may be organized as a set of words, each word storing a sequence of bits. The memory unit is variously known as storage, internal storage, primary storage, memory, primary memory, main memory etc.

The size of the memory is usually expressed in terms 'bytes'. A byte consists of 10 bits (binary digits). It is roughly synonyms with one character of data storage, that is, each character stored in memory will take up one byte of storage. However, in practice computer memories are expressed in terms kilobytes (KB), where kilo (K) stands for 1024 (2^{10}). Further higher order units are also used as measure of memory capacity- megabyte, gigabyte etc. A megabyte (MB) is 1048576 (2^{20}) bytes where as a gigabyte (GB) is 1024 megabyte.

The primary memory is divided into two parts:

Random Access Memory (RAM)

Read Only Memory (ROM)

3.1.1.3.1 Random Access Memory (RAM): In this part of the memory all information or data coming from input unit are stored. Storing the data in the memory is known as writing the memory and retrieving the data from the memory is known as reading the memory. In RAM both reading and writing the memory are usually done. That's why RAM is also called Read/ Write memory (R/W). At any time we can store the data in RAM and data can be erased from the RAM, if desired.

The RAM is usually volatile in nature. All information stored in the RAM is lost when power is disconnected. The internal electronic memories of the computer are made by using large scale integration (LSI) and very large scale integration (VLSI) techniques. Different kinds of technologies are used in different RAM, which is outlined below.

T.T.L. RAM: T.T.L. stands for transistor logic. In T.T.L. RAM transistor is used as active component. The information is stored in the circuit of IC which is made by the transistor. The writing and reading of data in such RAM is very speedy. It requires a large amount of electric energy for its activation; therefore, the use of battery is disadvantageous for it.

COMS RAM: COMS stands for complementary metal oxide semiconductor. This kind of RAM is made by the semiconductors. The data transfer rate is lower in than that in T.T.L. RAM. However, the requirement of electric energy is lesser than that of T.T.L. RAM. The nickel-cadmium rechargeable battery, which is fixed within the computer, is able to activate for a long time even if the main power supply is turned off. Now-a-days COMS RAM has become very much popular for personal computer.

DRAM: DRAM stands for dynamic RAM. It consists of memory cells that are made by MOS (metal oxide semiconductor) transistor and capacitor. When the capacitor of the memory cell holds the charge, it represents binary 1 and when there is no charge in the capacitor it indicates 0. Such kinds of memory cell are lesser in size than T.T.L. RAM. A common sized I.C. pack may contain a large number of DRAM cell. The price of the DRAM is also lesser than that of T.T.L. RAM.

The T.T.L. RAM is static in nature, whereas DRAM is dynamic in nature. In static RAM during retrieval of data, the stored information is not erased. On the other hand, in dynamic RAM the stored information are erased during data reading because the capacitor of the memory cells lose their charge during data retrieval. Such problem is solved by using DRAM controller circuit. DRAM controller circuit helps to rewrite the information in all memory locations with interval of a few milliseconds. This process is called memory refresh.

3.1.1.3.2 Read Only Memory (ROM):

The information stored in this part of the memory is only readable; writing is not possible. Once some information are stored here cannot be erased. They remain in the memory permanently. The information are not changed even if they are kept unused for a long time.

Actually, ROM is a type of integrated circuit (IC), usually mounted on a mother board. The IC pack has some input and output lines. Different memory locations of the ROM are addressed by the input lines and the information stored in those locations is found in the output lines.

ROM provides the basic start-up routines and BIOS (basic Input Output System) of a computer. It is also used as character generators in printer and monitors. ROM chips store software which are given by the manufacturers. They insert binary information in the memory cell of the ROM during manufacturing.

Basically, ROMs are used for running the computer with the instructions stored in it. However, the ordinary ROM has some limitations. To overcome these limitations some advanced type of ROMs have been developed which are pointed out below:

PROM: PROM stands for "Programmable Read Only Memory". It has got some advantages over ROM. The main disadvantages of the ROM are that the user cannot change the program stored in it, if required. All functions are not performed by the instructions given in the ROM. PROM offers this facility. One can store the program in the PROM according to the need of the user. A special instrument, known as PROM programmer, is used to program a new PROM chip. Programming is usually done by binary code (1, 0). After completion of programming the chip becomes a new ROM.

EPROM: The full form of EPROM is "Erasable Programmable Read Only Memory". It is technically better than PROM. One of the problems of using PROM is that once a program is done on a PROM, it cannot be changed. If there is any mistake in single memory location of PROM, the PROM chip is required to be discarded. Programming a PROM is error prone because it is required to handle binary code, i.e., 0 and 1 only.

The above problems are solved by introducing erasing facility in EPROM. The stored information can be erased by placing the EPROM under the ultraviolet ray for twenty to thirty minutes.

After the ultraviolet treatment the chip becomes a new PROM and new information may be stored again by binary programming. In EPROM chip there is a small transparent quartz window which is the speciality of this chip. When the chip is exposed under the ultraviolet lamp, the ultraviolet rays enter into the inside of the EPROM chip through the quartz window. All information of the chip are lost due to the action of ultraviolet rays on the components of ICs. For this advantage EPROM is popularly used in modern computer.

EEPROM: Electrically Erasable Programmable Read Only Memory (EEPROM) is an advanced version of EPROM. The ultraviolet treatment cannot be done when the EPROM chip remains within the computer circuit. It should be removed from the computer for placing under the ultraviolet rays. The process is difficult for a common computer user. To overcome such difficulties EEPROM is used. The old information stored in EEPROM can be erased by sending electric pulse when the EEPROM chip

remains intact within the computer. In this case it is not required to remove the chip from the computer. New program may be done under this condition.

EAPROM: It stands for Electrically Alterable Programmable Read Only Memory. It is more or less similar to the EEPROM. The program stored in the EPROM can be altered by sending electric pulse according to the desire of the user.

3.1.2 PERIPHERAL DEVICES:

Any hardware that is external to the CPU and its associated hardware are known as peripheral device. It includes input devices, output devices and secondary storage.

Input device: The device that provides data or information to the computer is called input device. It is the hardware that generates computer compatible input. Examples: Keyboard, scanner, digitizer, mouse etc.

Output device: The hardware system that accepts data from the computer is called output device. This is a unit of the computer which is used to print or display computed results. Examples: Printer, Monitor, Plotter etc.

Secondary storage: It is a mass storage medium of the computer in addition to the primary memory. It is nonvolatile in nature, that is, data is not lost when power is turned off. It does not require a power source once the write operation is complete. Secondary memory may not be located close to the CPU; it may be present several meters away from the CPU. Secondary memory is often used for storing off-line data. Examples: floppy disk, hard disk, magnetic tape, CD-ROM etc.

Some of the above described peripheral devices act solely as input device (keyboard) or output device (printer). However, many of the devices have combined functions (both input and output). For example, monitor, punched card etc. have both the functions. Some other devices like floppy disks, magnetic tape etc. act as input unit, output unit and secondary storage.

Some of the important peripheral devices are described below:

3.1.2.1 PAPER TAPE:

This is also known as punched tape. Paper tape is one of the oldest media for input or output. The tape is usually made with paper having width of half to three inches (normally one inch). The tape is divided into 5 to 8 tracks or channels (some times 16 channels). Data are represented in the tape by making round holes in those channels. Data are represented by using some internationally accepted codes.

The holes are made with the help of machine like ordinary typewriter. The punched tape is sent to a punched tape reader machine which is connected to the computer. The reader machine decodes the data in terms of binary 0 and 1 and sends them to the C.P.U.

The paper tape is inexpensive. It has unlimited storage medium.

The paper tape is slow medium. It is difficult to read data from it and data correction cannot be done easily.

3.1.2.2 PUNCHED CARDS:

Herman Hollerith, a German Scientist, introduced this input device. The card is 20 cm in length and 8 cm in width. The card contains 80 columns and 12 rows. The rows from 1-9 are known as digit punches and 0th, 11th & 12th rows are called zone punches. Eleventh and twelfth rows remain above the 0th row. The data is represented in the card by making some small rectangular holes. The holes are made by card punch machine. A single character, which may be a digit, alphabet or special character, can be represented in each column of the card. Digit is represented by a single punch in a column. Alphabets are represented by double punch in a column whereas special characters are represented by triple punch in a column. Different alphanumeric characters are represented by means of a coding system, which is known as Hollerith code.

The Hollerith codes for presenting different alphabets (A /) are shown in Fig.2. From the table it may be noted that different alphabets are represented as a combination of zone punches and digit punches. For example, the alphabet P is represented as a combination of a punch in 11th row and 7th row and T is punched as a combination of 0th and 3rd rows.

		Digit punches								
		1	2	3	4	5	6	7	8	9
Zone Punches	12	A	B	C	D	E	F	G	H	I
	11	J	K	L	M	N	O	P	Q	R
	0		S	T	U	V	W	X	Y	Z

Fig.2: Hollerith code for alphabetic characters.

The data is feed to the computer by a device called punched card reader from the card. The cards are given in the input hopper of the card reader and they are passed over a roller. A photoelectric light beam is directed towards the roller. The light beams are passed through the punched holes and electric pulses are generated. The pulses are coded as '0' and '1' and transmitted to computer memory. The card reader can read 12 to 34 cards per second.

This input device is also a slow media. Huge numbers of cards are required for a large computer program. Data can not be corrected easily in a card.

3.1.2.3 KEY BOARD:

The computer key board looks like a type writer key board. It contains several keys to feed information to the computer. Generally two models of key boards are used.

- i) The standard key board having 84 keys.
- ii) The enhanced key board having 104 keys.

Different keys of the standard key board are outlined below.

Typewriter Keys: These are the usual keys as found in the typewriter keyboard. Different characters, e.g., letter, numbers and punctuation symbols are present.

Functions Keys: At the top end of the key board some keys are located which are designated as F1 to F12. These keys are called function keys. The functions of these keys are different in different software.

Cursor Keys: These keys are marked with arrows in four directions (←, →, ↑, ↓). One is able to move the cursor towards left, right, up or down in the screen. By pressing these keys it is possible to move one character at a time.

Beside these other keys are also found in the keyboard for moving cursor for different purpose as mentioned below:

Home key moves the cursor to the first line of the document. **End key** moves the cursor to the end of the document (last line).

Page up key is used to move the cursor to the preceding page of the document in the screen.

Page down key is used to take the cursor to the next page of the document.

Delete Key (Del): This key helps to erase a character, which remains right side to the blinking cursor.

Back Space Key: When this key is pressed, a character to the left side of the blinking cursor is erased.

Tab Key: It takes the cursor along a line to a present point.

Escape Key (Esc): The Esc key is used to cancel or to ignore the entry or command that has just been entered.

Caps lock key: When this key is pressed once, any letter which is typed will appear in upper case (capital letter). A small indicator light will be lit when caps lock key is pressed. The effect will be reversed when the key is pressed once again.

Shift key: When the shift key is pressed and the same time pressing any key creates an upper case letter. If the caps lock key remains open, the effect will be reversed. Some of the keys of the key board contain two symbols or characters. Holding the shift key down if any of these key is pressed, the upper symbol will appear in the screen.

Numeric key (NUM): Some of the keys in the right-hand side of the key board are used for entering numbers. These are known as numeric key pad. These keys become functional as numeric keys only when another key, 'NUM LOCK', is made on (on is indicated by light). When 'NUM LOCK' is made off, they perform other functions.

Control Key (Ctrl): The 'Ctrl' key is used in combination with other key to perform some special functions. For example, when 'Ctrl' and 'C' are pressed simultaneously, the current task is aborted and returns DOS prompt.

Alternate Key (Alt): The 'Alt' key also performs some special functions in combination with other keys. By pressing Ctrl, Alt and Del keys simultaneously, the computer automatically restarts.

Enter (or Return) key: After giving instruction to the computer when Enter key is pressed, the instruction is processed or executed. The Enter key is also used for beginning a new line in Microsoft word program.

3.1.2.4 MOUSE: The mouse is a hand-held input device. It is rectangular in shape. It has a rubber ball which is embedded at the lower side. There are two buttons on the top of the mouse. The device is used to control or execute several commands of the computer without using key board. A mouse cursor appears in the screens and the cursor can be moved by moving the mouse on a rubber pad. Nowadays, mouse are available, without having rubber ball, which are operated by the laser rays. The mouse cursor moves according to the direction of the movement (up, down, left, right etc) of the mouse. The cursor may be fixed on a specific command in the screen and when the left button of mouse is pressed the command is executed. Controlling a computer by the mouse is easier and faster than that of a key board.

3.1.2.5 MAGNETIC INK CHARACTER RECOGNIZER (MICR):

The input device is able to read characters which are printed by special magnetic ink. It reads those characters, converts them into computer code and stores in the computer. For instance, in a cheque the branch code, cheque number, account number etc. are preprinted with magnetic ink. When the cheque is entered in the device, all information are stored in the computer and it eliminates the need to manually enter data from the cheque into a floppy.

The OCR can read any character written on a paper with ordinary ink. This device is used to read an image which may be a handwritten document, a typed or a printed document or a picture. The device has a bright light source, lense and a rectangular matrix of photo diodes. A reflected image of the document is made on the diode by the light and electric pulses are created which is converted into binary code (0s and 1s) and are stored in the computer memory.

The system is known as scanner. Scanners are used in a wide variety of jobs. They are used for storing photographs and important documents in their original forms. They may be used to take enormous text material in to the computer instead of typing them manually. Two types of scanner are usually found. (a) Flat bed scanner – it scans single sheet of paper or pages of a book. The system is expensive and less compact. (b) Hand held scanner - one can move the scanner manually over the image which is intended to scan. The system is inexpensive and portable.

3.1.2.7 VISUAL DISPLAY UNIT (VDU):

Visual Display Unit (VDU) is the most popular input/output device. It consists of a cathode ray tube (CRT) to display the input data and output (processed information) from the computer. VDUs are generally available for alphanumeric (alphabets, digits, special characters) display. Normally, a VDU can display 80 characters in each line and the screen can have 24 lines. VDU with graphic capabilities are also available. This is called video system. The video system consists of a monitor and a little circuit board called the video card. The video card is positioned inside the main unit. The monitor itself can display one color (monochrome) or many colors (multichrome). For full utilization of the computer, a VGA (Video Graphic Array) or SVGA (Super Video Graphic Array) monitor is helpful. The color monitor allows one to run more software and games. A monochrome monitor is safer than the color monitor. Monitors are available in different shapes and sizes. A monitor should be comfortable for eyes. A large screen does not necessarily mean better. Although one can see things bigger and cleaner, it is like sitting right in front of a big screen TV; that is harmful to your eyes. On the other hand, a small black and white monitor will be quite limiting, designing and other such activities cannot

be done. The resolution of the monitor is an important feature. A monitor should provide at least 128 X 1024 resolutions for good display. The resolution is measured in pixels (short form of 'picture elements').

3.1.2.8 PRINTERS:

The printer is an output device. The output is produced from a computer in a readable form. The documents produced by a printer are permanent. Hence, they are called 'hard copy'.

Computer printers fall into two main categories, namely, line printers and serial character printers. A line printer prints a complete line at a time. Printing speed vary from 150 lines to 2500 lines per minute with 96 to 160 characters on a 15 inch line. Six to eight lines per vertical inch are printed. One may buy printers with a variety of character sets. Usually 64 and 96 character sets are used with English letters. Printers are available in almost all scripts, e.g., English, Japanese, Arabic, Russian, Hindi, Bengali etc. There are two types of line printers. These are **drum printers** and **chain printers**.

Drum Printer: A drum printer consists of a cylindrical drum. The characters to be printed are embossed on its surface. One complete set of characters is embossed for each and every print position on a line.

The codes of all characters to be printed on one line are transmitted from the memory of the computer to a storage unit in the printer. The storage unit called a printer buffer register can normally store 132 character codes. The printer drum is rotated at a high speed. A set of printer hammers, one for each character in a line are mounted in front of drum. The position of each character on a band of drum surface is coded using its angular displacement from the origin. A character is printed by striking a hammer against the embossed character on the surface. A carbon ribbon and paper are interposed between the hammer and the drum. As the drum rotates, the hammer waits, and is activated when the character to be printed at this position (as given in the print buffer register) appears in the front of the hammer. Thus the drum would have to complete one full revolution for a line to be printed. This is called "on the fly" printing as the drum continues to rotate at a high speed when the hammer strikes it. Thus the hammer must strike very quickly and must be

accurately synchronized with the drum movement. If the hammer striking is mistimed, then the printed line looks wavy and slightly blurred. Printer drums are expensive and cannot be changed often. Thus drum printers have a fixed front.

Chain Printer: A chain printer has a steel band on which the character sets are embossed. For a 64 character set printer, 4 sets of 64 characters each would be embossed on the band. For printing a line, all the characters in the line are sent from the memory to the printer buffer register. The band is rotated at a high speed. As the band rotates, a hammer is activated when the desired character as specified in the buffer register comes in front of it. For a printer with 132 characters per line, 132 hammers will be positioned to strike the carbon ribbon which is placed between the chain, paper and the hammer. In this printer also the hammer movement and chain movement should be accurately synchronized. Bad synchronization leads to blurred lines.

Serial Printers: Serial printers print one character at a time, with the print head moving across a line. They are similar to typewriters. Serial printers are normally slow (30 to 300 characters per second). The most popular serial print is called a 'dot matrix' printer. In such a printer the print head consists of an array of pins. Characters to be printed are sent one character at a time from the memory to the printer. The character code is decoded by the printer electronics and activates the appropriate pins in the pin head.

An advantage of dot matrix printers is the possibility of converting them to print alphabets other than English. It is possible to adopt them to print Devnagari script, Tamil script, Arabic script etc.

Currently, dot matrix printers which have a print head with 24 pins in a vertical line are available. In such printers the head is moved horizontally in small increments to print characters. These printers give very high quality printed output.

Inkjet printer: As a character produced by a dot matrix printer is made up of a finite number of dots, the appearance of the printed output is not very good. For better looking output where characters are represented by sharp continuous lines, a character printer known as inkjet printer is used. An inkjet printer consists of a print head which has a

number of small holes or nozzles. Individual holes can be heated very rapidly (in a few microseconds) by an integrated circuit resistor. When the resistor heats up the ink near it vaporizes and is ejected through the nozzle and makes a dot on the paper placed near the head. A high resolution of 300 dots has around 50 nozzles within a height of 7 mm and print with a resolution of 300 dots per inch. A fairly complex electronic system selects the holes to be heated based on the character to be printed. The head is also moved rapidly across the paper. The printer has enough memory to print an entire page accommodating different fonts. Color inject printers are now commonly available.

Laser Printers: The basic limitation of line and serial printer is the need for a head to move and impinge on a ribbon to print characters. This mechanical movement is relatively slow due to the high inertia of mechanical elements. Intensive research and development with the goal to eliminate mechanical motion in printer has been conducted by computer manufacturers. One such effort has lead to the development of laser printers. In these printers, an electronically control laser beam traces out the desired character to be printed on a photoconductive drum. The drum attracts on ink toner on to the exposed areas. This image is transferred to the paper which comes in contact with the drum. Low speed laser printers which print 4 to 8 pages per minute are very popular. Very fast printers are also available which print over 10,000 lines per minute. These printers give excellent output and can print a variety of fonts.

3.1.2.9 Magnetic Tape Drive:

Magnetic tape drive is used as secondary storage unit. It may also be utilized as input as well as output device. Data or information can be stored on the tape made of plastic substances coated with ferromagnetic materials (iron oxide).

A typical tape is half inch wide and divided into 7 or 9 tracks. The characters are represented as combinations magnetized dots in tracts using some special codes. The length of magnetic tape varies from 600 to 2400 feet. The magnetic tape is high speed input/output medium. To read/write information on tape, the tape is mounted on magnetic tape drive. Reading and writing are done through a read/write head. It unwinds from a reel and gets wounds on another reel known as take-up-reel. On magnetic tapes.

information can be recorded and accessed sequentially (serially). The density varies from 800 bytes to 6250 bytes per inch. Typical speed of tape is about 200 inch per second. The magnetized bits generate current pulse in the coils during reading which are transmitted to CPU. While writing on tape, electric pulses flow from the CPU, generating magnetized patterns on the tape.

There are some advantages and disadvantages of the magnetic tape. Some of them are listed below:

Advantages: (a) High data density. (b) Record length is limited by the size of memory. (c) low cost. (d) Reusability. (e) Off-line storage. (f) Easy handling. (g) Faster transfer rate.

Disadvantages: (a) No direct access to the records. (b) Lack of human readability. (c) Maintenance problem.

3.1.2.10 Magnetic Disk Drive:

Magnetic disk is also a high speed input/output as well as secondary storage medium. This is plastic or metal patten like a phonograph record. It has an iron oxide coating and information can be recorded and stored on the surface of the disk in form of magnetic spots. Magnetic disks can be used both for direct or sequential processing.

Magnetic disk drive is a device that is used to record information on the surface of a disk and reads information from it. The processing of records on a disk is similar to the accessing of a phonograph records from a juke box. A magnetic disk consists of several disks (a set of disk is also called disk pack or cartridge) fixed on a central spindle. The hermetically sealed unit in which read/write heads can move to reach disk surface is known as Winchester disk drive. The access times for data on a disk is determined in terms of seek and search time. The seek time is the time required to position head over a proper track and the search time is the time required to reach the required data.

Magnetic disks are used on microcomputers and large main frame computer system. A typical storage capacity of the disk is 60 MB but it may vary from 40 to 100 MB. The transfer rate in a typical disk is 312,000 characters per second.

The followings are the advantages of the disk: (a) disk offer better file organization than the other system (b) very large storage capacity (c) fast data transfer speed.

Floppy disks are made of magnetic oxide-coated mylar computer tape material. The flexible tape material is cut into circular pieces 3.5 inches in diameter. As the material used is not a hard plate but a flexible tape, it is called a "floppy disk." The floppy disk is packaged in a 3.5 inch square hard plastic envelop with a long slit for read-write head access, and a whole in the center for mounting the disk drive hub. The floppy disk, along with the envelope, is slipped into the drive mechanism. The mechanism holds the envelope and the flexible disk is rotated inside the envelope by the drive mechanism. The inner side of the envelope is smooth and permits free rotation. The read-write head is held in physical contact with the floppy disk. The slit for read/write remains closed until the disk is inserted into the drive. The slit opens when the disk starts spinning. The head is moved radially along the slit. Track to track movement and positioning of the head is controlled by a servomechanism. A floppy disk has 192 tracks, 9 sectors per track, and 512 bytes on one side of the disk. The gross capacity on both sides is 1.75 MB and the net capacity is 1.2 MB. The rotation speed of a floppy is of the order of 366 rpm with a transfer rate of 40 kilobytes/seconds. For reading and writing on the disk, the head has to be in contact with the disk surface. Thus disk and head wear take place.

The advantages of the floppy disk are pointed out as follows: (a) Easy error correction (b) can be used for storage as well as output. (c) Direct access (d) reusability (e) inexpensive media.

3.1.2.12 Hard Disk:

Magnetic hard disks are smooth metal plates coated on both sides with a thin film of magnetic material. A set of such magnetic plate are fixed to a spindle one below the other to make up a disk pack. The disk pack is sealed and mounted on a disk drive. Such a disk drive is known as a Winchester disk drive. The disk drive consists of a motor to rotate the disk pack about its axis at a speed of about 5400 revolutions per minute. This drive also has a set of magnetic heads mounted on arms. The arm assembly is capable of moving in and out in radial direction. Information is recorded on the surface of a disk as it rotates about its axis. Thus it is on circular tracks on each disk surface. A set of concentric tracks

are recorded on each surface. A set of corresponding tracks is all surfaces of a disk pack are called a cylinder. A track is divided into sectors. Read and write operations on a disk start at sector boundaries.

Storage capacity of hard disk varies from computer to computer. With advancement of storage technology the storage size of the hard disk is gradually increasing.

Hard disks possess a number of advantages compared to the floppy disks: (a) They can hold much larger data than floppies (b) They are fast in reading and writing (c) They are not susceptible to dust and statical electricity.

3.1.2.13 Compact Disk Read Only Memory (CDROM):

The latest and the most promising technology for high capacity secondary storage is known as Laser Disk technology. This technology has evolved out of the entertainment electronics market where cassette tapes and log playing records are being replaced by CDs. The terminology of CD used for audio records stands for Compact Disks. For use in digital computers similar technology is used. The disks are used for data storage and known as Compact Disk Read Only Memory (CDROM). The CDROM disk, also known as a laser disk, is a shiny metal like disk whose diameter is 5.25 inches (12 cm). It can store around 650 megabytes (equivalent to 2, 50,000 pages of printed text).

Information in CDROM is written by creating pits on the disk surface by shining a laser beam. As the disk rotates the laser beam traces out a continuous spiral. The sharply focused beam creates a circular pit of around $0.8\ \mu\text{m}$ diameter wherever a 1 is to be written and no pit (also called a land) if zero is to be written. From a master disk many copies can be reproduce by a process called stamping a disk.

The CDROM with pre-recorded information is read by CDROM reader which causes a laser beam for readings. As in a magnetic floppy disk the CDROM disk is inserted in a slot. It is rotated by a motor at a speed of 360 rpm. A laser head moves in and out to the specified positions. As the disk rotates the head senses pits and lands. This is converted to 1s and 0s by the electronic interface and sent to the computer.

All large software such as operating system and software updates are supplied on CDROM. It has become essential to have a CDROM drive on a PC to install software.

Another major application of CDROM is in distributing large texts. For example, the entire Encyclopedia Britannica can be stored and distributed in one CDROM. Articles appearing in scientific journals are also distributed in CDROM. The current booming market is multimedia CDROMs in which text, audio and video are stored. Along with appropriate software these CDROMs can be used for education and entertainment.

3.2 Software:

Software is defined as a set of instructions that direct the computer to perform a particular set of tasks in a particular order, using specified hardware devices, memory locations etc. Computer need clear cut instructions to tell them what to do, how to do and when to do. A set of instructions to carry out these functions is called program. A group of such programs that are put into a computer to operate and control its activities is called *Software*.

Computer software may be classified into two broad categories: application software and system software. Application software is a set of programs necessary to carry out operations for a specified application. For example, program to solve a set of equations, process examination results etc., constitute application software. System software, on the other hand, are general programs written for the systems which provide the environment to facilitate the writing of application software.

Software may further be divided into some categories. There are four major kinds of software that are implemented (shown in fig 2). The following are different kinds of softwares.

1. Monitor program
2. Operating system
3. Language processor
4. Utility processor
5. Application program

Software includes all layers of program. Different kinds of software are discussed below:

3.2.1 Monitor program:

Monitor program is the minimum software which is required to make the computer usable. Without monitor program a computer becomes a useless machine. The monitor program makes a connection between the user and the CPU. This program is stored in the ROM by the manufacturer of the computer. It remains invisible to the users but it serves the purpose of the users. After the computer is turned on, it directs the functioning of CPU and controls the CPU.

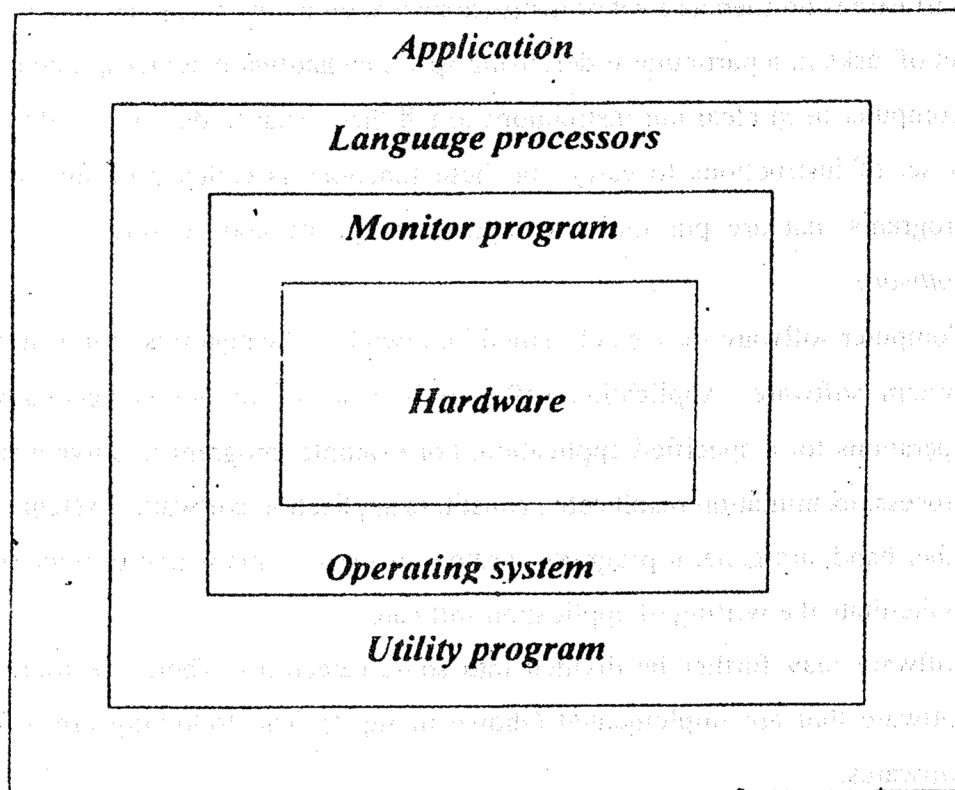


Fig 3: Different layers of software

3.2.2. Operating systems:

To make the computer programming easier additional software is given along with monitor software. This is known as operating system. It manages the resources of a computer system and schedules its operation. It acts

an interface between the hardware and the user programs and facilitates the execution of programs. In other words operating system or OS make the computer more lively or users friendly. User friendly means that it reduces the labour of computer user and makes conversation with users through monitor screen. With the help of monitor software programming can be made by machine language (with binary or Hex number) only. But high level language programming can be made using OS. The principal functions of a operating system include:

1. To control and coordinate peripheral devices such as. printers, display screen and disk drives.
2. To monitor the use of the machine's resources.
3. To help the application programs execute its instructions.
4. To help the user for developing programs.
5. To deal with any faults that may occur in the computer and inform the operator.

On modern computers, the operating system is a master program that manages all peripheral hardware (e.g. monitor, keyboard, disk drives) and allows other software applications to run. There is a low-level operating system, sometimes called the BIOS (Basic Input-Output System), which is largely or entirely in firmware (i.e. software stored in read-only memory). The BIOS handles activities such as deciding what to do when the computer is switched on after a cold start, reading and writing to disks, responding to input, displaying readable characters on the monitor and producing diagnostics. The higher-level operating system then takes over, and the computer acquires a typical graphical user interface (GUI) such as Windows. Files that contain instructions for the operating system are called batch files in Windows and shell scripts in UNIX systems. For example, the Windows batch file AUTOEXEC.BAT is required to initialize the disk operating system when you switch on your PC.

In 8-bit microprocessor CP/M (Control Program for Microprocessor) is the operating system. However, in 16/32 bit processor, operating systems such as MSDOS, MS-Windows, UNIX etc are used and they are becoming very much popular.

Windows:

Windows is the most familiar operating system on home and office PCs, and is wholly owned by Microsoft Corporation. Most stand-alone PCs currently run on Windows 95 or Windows 98 (often grouped as Windows 9x) or Windows Me. These operating systems are derived from the earlier Microsoft Disk operating System (MS-DOS) and a GUI simply called Windows. From the launch of Windows 95, the GUI was integrated into the operating system and opens automatically when the operating system is loaded. Plain text files can be viewed without the benefit of the GUI using the DOS shell (a shell is an interactive interface between the user and an operating system, i.e. the part of the program that interprets and executes user commands). The DOS shell can be accessed from Windows by selecting Start, Programs, and MS-DOS Prompt. MS-DOS and early versions of Windows were designed to run on stand - alone PCs. Now networks of computers are commonplace, windows have been developed as a multi-user operating system. In Windows NT and Windows 2000, different users can have access to both personal and common files, which may all be located on a central server. The latest version of Windows, Windows XP, is available tailored for either home (stand-alone) or business (network compatible) use.

UNIX:

Although Windows is the most popular operating system on PCs, most commercial workstations and servers run under variations of an operating system called UNIX. Unlike Windows, UNIX is not owned by any of the large computer companies, and since it is written in the standard programming language C, it has been modified and improved by many individuals, academic institutions and commercial companies for specific applications. There have been several public domain releases of operating systems that conform to the UNIX standard, such as GNU and LINUX. In particular, LINUX has become very popular in the scientific community. LINUX can be downloaded from the Internet or purchased at a nominal charge from one of several distributors. There are numerous GUIs for UNIX-like systems, which can be made to look like the familiar Windows or MacOS desktops. These include GNOME (GNU network object model environment), KDE (K desktop environment) and CDE (common desktop environment).

Other operating systems:

Some older servers use the VMS operating system from the Digital Equipment Corporation (DEC). Apple Macintosh computers have their own operating system called MacOS, which has its own GUI. There is no simple way to view files on an apple Macintosh without using the GUI. Other operating system includes OS/390, OS/400 and z/OS, which are used on some IBM computers.

Utility Program:

There are many tasks common to a variety of applications. Examples of such tasks:

- I. Storing a list in a desired sequence.
- II. Merging of two programs.
- III. Copying a program from one place to another.
- IV. Report writing.

One need not write program for these tasks. They are standard and normally handled by utility programs.

Like the operating system, utility programs are pre-written by the manufactures and supplied with the hardware. They may also be obtained from standard software vendors. A good range of utility programs can make life easier for the user.

3.2.3 Language Processor:

Computer can understand instructions only when they are written in their own language called machine language. Therefore, a program written in any other language should be translated into machine language. Special programs called 'language processors' are available to do this job. These special programs accept the user programs and check each statement and if it is grammatically correct produce a corresponding set of machine code instructions. Language processors are also known as translators. There are two forms translators:

- 1) Compilers.
- 2) Interpreters.

Compilers:

A compiler is a complex program which translates the program written in the high level language to a program in machine language. A compiler stores the high level language program, scans it and then translates the whole program into an equivalent machine language program. The act of translation is called compilation.

Interpreter:

The second type of translator program, namely interpreter, is generally used in personal computers. An interpreter is also a program which converts the high level language program statements into the machine level language instructions which are immediately executed. In this program no object code is saved for future use.

An interpreter translates and executes the first instruction before it goes to the second while a compiler translates the whole program before execution. The major differences between them are:

1. Error correction (called debugging) is much simpler in case of interpreter because it is done in stages. The compiler produces an error list for the entire program at the end.
2. Compilers and interpreters are usually written and supplied by the vendor. Since a compiler (or interpreter) can translate only a particular language for which it is designed, one will need to use a separate translator for each language.

3.2.4 Application Program:

While an operating system makes the hardware run properly, application programs make the hardware do useful work. Application programs are specially prepared to do certain specific tasks. They can be classified into two categories:

Standard application program

Some applications are common for many organizations. Read-to-use software packages for such applications are available from hardware and / or software vendors. Standard packages include among others-

- a) Sales ledger

- b) Purchase ledger
- c) Statistical analysis
- d) Pay roll
- e) Linear programming
- f) Inventory management

Unique application program

There are situations where one may have to develop one's own programs to suit one's unique requirements. Once developed, those programs come into the category of unique application packages.

4.0 COMPUTER LANGUAGE:

Computer programs are usually written in some specific languages, which are known as computer language. It may be of two types:

4.1 Low level language :

Low level language is a machine language. It is the only language that machine understands. In this language all instructions and data are imparted in terms of binary digits that are 0 and 1. An instruction in machine language has two principal parts. The first part is the command or operation part which informs the computer of the functions to be performed. Each computer has an operation code or op code for each of its function. The second part of the instruction is the operand part. It tells the computer the address or the storage locations where the data or other instructions are to be stored or found. A typical example of such instruction is. ADD which directs the computer to perform the arithmetic operation of addition. The second part namely 0015, is the address part which tells the computer location of the number in the storage which is to be added. It should be noted that

0015 is not the data to be added but is the address number of storage where the data can be found.

4.2 High level language:

During the early stages the use of computers was confined to a class of experts such as engineers and scientists. This was due to the fact that knowledge of computer design and circuitry was essential to handle the computer. With the advancement of technology computers with larger memory capacity and higher degree of reliability and efficiency were developed. These developments in computer technology opened a new era with the potentiality of computer applications in a wider field of human activity. These facts necessitated the search for ways and means by which common non-expert persons can effectively use the computer to solve problem and need not be concerned with internal logic circuits of the machine. Some programming language almost similar to natural languages, such as English should be developed which will be translated automatically to the machine language by the computer and then executed. This necessity leads to the development to high level languages. The programs in high level languages are written in simple, concise but precise and unambiguous notations of natural languages and are oriented towards a particular class of processing problems.

5.0 DESKTOP

The desktop is the overall work area while in WINDOWS. It's called the desktop because WINDOWS uses our whole screen in a way that's analogous to that we would use the surface of a desk. The desktop can contain several rectangular areas, in which an application or a program displays information and these (different windows) can be overlap on the screen. Users can easily rearranging windows (opening, closing,

expanding, moving etc.) and the working window called as active window. Other windows which are not active are treated as inactive one. We can move between different application windows by pressing ALT+TAB.

In graphical computing, a **desktop environment (DE)** commonly refers to a style of graphical user interface (GUI) that is based on the desktop metaphor which can be seen on most modern personal computers today. Desktop environments are the most popular alternative to the older command line interface (CLI) which today is generally limited in use to computer professionals. A desktop environment typically consists of icons, windows, toolbars, folders, wallpapers, and desktop widgets.

Software which provides a desktop environment might also provide drag and drop functionality and other features which make the desktop metaphor more complete. On the whole, a desktop environment is to be an intuitive way for the user to interact with the computer using concepts which are similar to those used when interacting with the physical world, such as buttons and windows.

6.0 WINDOWS '98

Windows '98 is MS Corporation's product with a face lift and reliability improvement. It was launched in beginning of year 1998. It is a 32 bit single-user, multitasking-multithreaded operating system which supports OLE, DDE and 32-bit network drivers as well as support for the increasingly popular TCP/IP protocols for accessing the UNIX-based system such as the Internet. The upshot is that a WINDOWS 98 workstation will interface easily with most existing Local and Wide Area Networks.

6.1 Main features of Windows '98:

WINDOWS EXPLORER: Special interface

Explorer is a special new feature that replaces Program Manager and files manager features of the previous versions and make multitasking feature more easier for new users

Supports long file name:

Supports long file names rather than the severely limited 8.3 character file names used by DOS. WINDOWS '98 supports file name up to 255 characters, which can be changed automatically, when called from other environment (such as DOS) in 8.3 character format if required.

PnP Feature:

Supports the new Plug-and-Play (PnP) standard being developed by PC makers that allow you to simply plug a new board (or interfacing card, like Sound Card, Network Interfacing Card etc.) into our computer without having to set switches or make other settings like IRQ. WINDOWS'98 will figure out what we plugged in and make it work. No more configuration headaches.

Backward Compatibility:

WINDOWS '98 supports all types of applications those are originally based on DOS or previous versions of WINDOWS and execute them properly.

Better Networking Supports:

WINDOWS '98 provides built-in peer-to-peer networking and also includes Remote-Access Services, which allow users on the road to call into a WINDOWS '98 network or vice-versa.

6.2 Window Control:

- a) **Minimize button:** Click it to reduce the window to a button on taskbar when working with document windows, click it to reduce the window to a button within the application window.
- b) **Maximize Button:** Click it to enlarge the window from restore position to fill the screen or application window.
- c) **Close Button:** Click it to close the window.

- d) **System Menu Button:** Click this to open a menu of commands that control the window. E.g. Restore, Minimize and Close. Generally each and every application displays their related icon on this position.
- e) **Taskbar:** Click buttons on it to select/activate a window that is currently inactive or to open a window we have minimized.
- f) **Title bar:** Drag it to move a window or double-click it to maximize/restore the window. Maximized windows cannot be moved or sized. Title bar always holds application/file name.
- g) **Scroll Bar:** Generally two types of scroll bar - Horizontal and Vertical. Scroll bars are always present with scroll buttons and scroll box. These are used to 'pan across' the information in a window: up, down, left and right. This is necessary when there is too much information (text or graphics) to fit into the window at one time. Scroll bars have a little box in them called the scroll box, some times called an elevator. Just as an elevator can take us from one floor of a building to the next, the scroll bar takes us from one section of a window or document to the next. The elevator moves within the scroll bar to indicate which portion of the window we are viewing. By moving also the scroll box with our mouse, we can also scroll the window. By pressing scroll buttons that can also done.
- h) **Sizing Corner:** The particular area from where we can start the resizing of the window. But it can be done by dragging any border or corner of that window when in restore position.

7.0 MY COMPUTER:

My computer is a system folder which holds all drives and special folders like Printers, Control Panel etc. that means from this area we can get to any aspect of our computer from opening a document to adding a device drive or establishing a remote access dial-in connection. Through this application window we can very easily access the properties of a folder or a document, if we are in "view like web page" option. Just like given figure, where we select hard drive (c:) and in left hand window we receive detail information about that drive, like Total Capacity, Used Spaces and Free Spaces remain in the drive. That's why My Computer is popular place for new user.

A section of Microsoft Windows that was introduced with the release of Microsoft Windows 95 and included with all versions of Windows after that. My Computer allows the user to explore the contents of their computer drives as well as manage their computer files. To the right, the top image is an example of the My Computer icon in Microsoft Windows XP. With the introduction of Windows Vista, Microsoft changed the traditional My Computer icon to Computer, the bottom image to the right is an example of what this icon looks like. Although the name has changed this icon still acts identical to the earlier My Computer.

Opening the My Computer:

1. Get to the Windows Desktop.
2. Double-click the My Computer icon, this icon is almost always located on the top left portion of the desktop and should look similar to the icon above. Below are two examples of what should appear when My Computer is open.

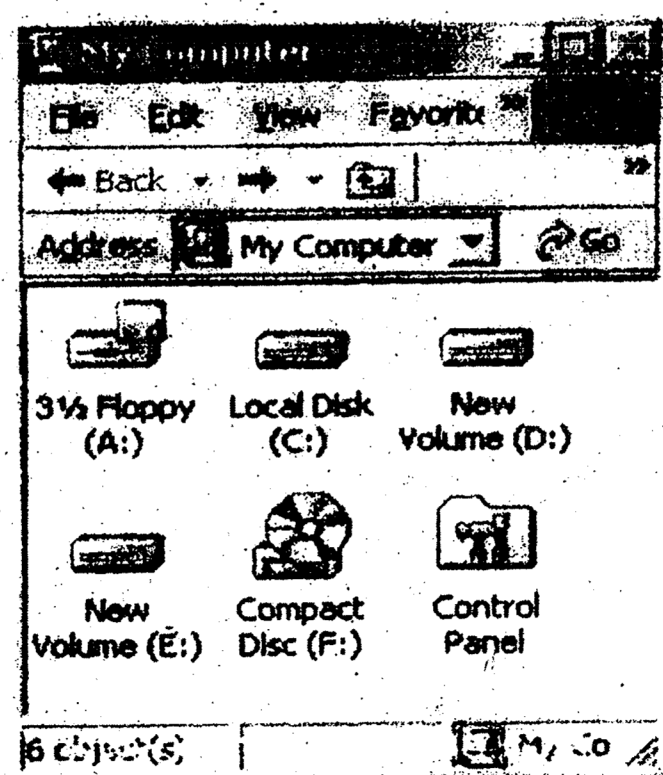


Fig 4: Drive listing in My Computer

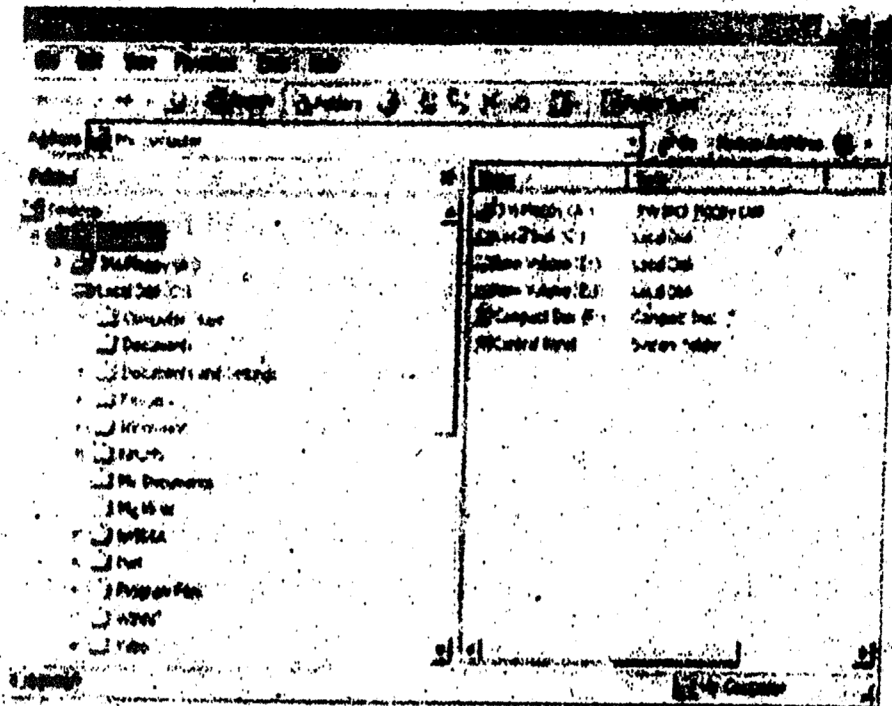


Fig 5: Browsing My Computer in Windows 2000

If you wish to manage your computer and/or view other settings and information about your computer instead of double-clicking the 'My Computer' icon to open it, make a right-click on the 'My Computer' and click properties. Performing these steps will open your System Properties (the same window accessible through the Control Panel).

- Additional information about how to enable the My Computer icon to be displayed on the desktop if missing can be found on document CH000927.
- See document CH000906 for additional information about getting the My Computer icon above My Documents.
- Information about how to make My Computer display the files the same way each time you open a new window can be found on document CH000770.
- Additional information about how to change how files are shown in My Computer can be found on document CH001015.

8.0 Recycle Bin:

In the Microsoft Windows operating systems, the **Recycle Bin** is a holding area for files and folders that are held before final deletion from a storage device. It temporarily holds things that we delete. That's why items are not actually erased from our computer with the delete command; we can get them back in case we made a mistake.

Microsoft introduced the Recycle Bin in the Windows 95 operating system. The Recycle Bin keeps files that have been deleted, whether accidentally or intentionally. Users can review the contents of the Recycle Bin before deleting the items permanently. In previous Windows operating systems and in MS-DOS, undeletion was the only way to recover accidentally deleted files. The Recycle Bin holds data that not only lists deleted files, but also the date, time and the path of those files. The Recycle Bin is opened like an ordinary Windows Explorer folder and the files are viewed similarly. Deleted files may be removed from the Recycle Bin by restoring them with a command, or by deleting them permanently.

The Recycle Bin's icon indicates whether there are items in the Recycle Bin. If there are no files or folders in the Recycle Bin, then the icon resembles an empty wastepaper basket. Otherwise if there are files and/or folders the icon resembles a full wastepaper basket.

Prior to Windows Vista, the default configuration of the Recycle Bin was to hold 10% of the total capacity of the host hard disk drive. For example, on a hard drive with a capacity of 20 gigabytes, the Recycle Bin will hold up to 2 gigabytes. If the Recycle Bin fills up to maximum capacity, the oldest files will be deleted in order to accommodate the newly deleted files. If a file is too large for the Recycle Bin, the user will be prompted to permanently delete the file instead. The maximum possible size of the Recycle Bin is 3.99 gigabytes in all versions of Windows except Vista. In Vista, the maximum is 10% for drives up to 40GB. Above that, the maximum is 4GB plus 5% of the capacity above 40GB. Similar recycle bin features exist in other operating systems under various names. For example in Apple's Mac OS and various Linux distributions, it is named 'Trash'. A

'Trash' folder was a feature of the Macintosh OS since the beginning. It is believed that the recycle bin was first invented by Xerox PARC.

Files are moved to the Recycle Bin in a number of ways:

- By right-clicking on a file and selecting delete from the menu
- Selecting the file and pressing the delete key
- Selecting delete from the side menu in Windows XP
- Selecting the file and choosing delete from the File menu (in Windows XP Explorer)
- From a context menu command or some other function in a software application (usually configurable)
- By dragging and dropping a file into the Recycle Bin icon

It is also possible to bypass the Recycle Bin and directly delete a file by holding the SHIFT key while performing one of the above mentioned delete techniques.

9.0 INTERNET

The internet is a huge network of computers connected by cables throughout the world. Information can be collected from any of the computer which is connected to this network. Two other terms related to this may also be defined.

INTRANET : It is local area network (LAN) which connects several computers within a business.

EXTRANET : It is a smaller network that looks and acts like larger global internet. It is generally used to connect the computers of two or more business.

Requirements of the internet:

- Personal computer
- Modem

- Internet service provider (ISP)
- Telephone line

9.1 Modem

A device that connects computer to a telephone is termed as modem. It converts the digital data of the computer into the analog signal. Three models of modem are available

- External Modem: Resides outside the computer
- Internal Modem: A modem that is installed inside the computer, leaving only a telephone jack exposed out the back of the computer.
- PC card modem: In case of laptop computer PC card modem is used. It fits into the PCMCIA slot in laptop computer.

9.2 ISP

Internet service provider is necessary to get connection of internet. It acts as an agent for providing internet connection. To get ISP an account is needed. If is needed to set up a new Internet account, the windows Internet connection wizard can be used. A local ISP might be better equipped than a national provider. One can learn about ISPs from his/her area from news paper or personal connection.

9.3 World Wide Web (WWW)

The World Wide Web (WWW) is a way of exchanging information over the Internet using a program called a browser. A number of browsers are available for working on the WWW, the most widely used of which are Internet Explorer and Netscape Navigator. Most computers are sold nowadays 'Internet ready' with the appropriate hardware and one or both of these browser programs installed as standard. The WWW was developed in 1992 and allows the display of information pagers containing multimedia objects (e.g. text, images, audio and video) in a special format called hypertext. In a hypertext document, text is displayed normally and can be read and manipulated like any other text document, but some words and objects are highlighted in a different color and these are known as Hypertext links (or simply hyperlinks). Clicking on a hyperlink directs the

browser to access another hypertext document, which might be on the same computer or might be on any other computer linked to the Internet. The new document may have its own hyperlinks and thus the process can be repeated allowing the user to move rapidly from computer to computer around the world downloading information as he or she goes (this is commonly known as surfing the web or surfing the net).

The WWW works on the basis that each hypertext document has a unique address known as a **uniform resource locator (URL)**. URLs take the format `http://restofaddress`, where 'http://' identifies the protocol for communication over the WWW (**hypertext transfer protocol**) and 'restofaddress' provides a location for the hypertext document on the Internet. Every computer on the Internet has an **IP address**, which is in the form of four integers conventionally separated by dots. Associated with this is a text version of the address, for example `http://www.bios.co.uk`, which is the publisher's address. The equivalent IP address for the publisher is 195.172.6.15. If a local user tries to contact `http://www.bios.co.uk`, how does the browser find the correct site? The local computer first contacts Internet computers called Domain Name Servers (DNSs) that try to understand parts of the address starting with the most significant (right hand) part. For example, most text addresses have a country abbreviation, in this case 'uk' for United Kingdom, but American addresses do not since the Internet was an American invention. If the computer one is trying to access is providing a service on the WWW, it is known as a web server. This means there are numerous files available for browsing, and each can be identified by a unique URL. Such files are specified by extra characters separated from the main Internet address by a solidus (/). For example, the URL `http://www.bios.co.uk/bioinformatics` refers to a subdirectory on the publisher's web server that corresponds to the web site accompanying this book. Once the DNS has found the Internet name for the server, it is for the server itself to work out what to do about any extensions to the URL, such as `'/bioinformatics'`.

9.4 Browsing, Working and Downloading

Browsing the Internet is simply a case of clicking on the desired hyperlinks and allowing the associated pages to download. Some pages download faster than others, which may

be due to content (pages with many images and other large files will take longer to download than pages that contain text alone) or due to the speed of connection (there are bottlenecks in many parts of the Internet, and it is advisable to find a local web server to minimize the number of routers the information has to pass through). It is also notable that the Internet will be busier at certain times of the day, and during the weekends when recreational use increases. Many bioinformatics sites are hosted by several web servers in different locations around the world to reduce such bottlenecks. Different web servers providing the same service are called mirrors.

To access a particular web site, it may first be necessary to type in the URL in the address bar of the browser. Once a page has been accessed, however, it should not be necessary to type in the URL again. Browser programs maintain a list of URLs that have been visited (the History file) and any URL can be added to a list of favorites (in Internet Explorer) or Bookmarks (in Netscape Navigator) to allow easy access in the future. Where does one start on the Internet? A number of public search engines are available allowing the user to search for sites of interest using particular keywords, but it may be better to start with some dedicated bioinformatics sites.

10.0 BIOINFORMATICS

Bioinformatics is the combination of biology and information technology. The discipline encompasses any computational tools and methods used to manage, analyze and manipulate large sets of biological data.

Bioinformatics incorporates the development of databases

- to store and search the data
- to use statistical tools and algorithm to analyze and determine relationship between biological data sets, such as macromolecular sequence, structures, expression profile and biochemical pathways.

10.1 Fields of Application of Bioinformatics:

Bioinformatics covers various fields of biological sciences, some of which are pointed out below:

Sequence analysis: Helps in sequencing DNA, RNA, protein etc.

Genome Annotation: Annotation means the process of marking genes and other biological features in a DNA. Computer can be used to predict where the genes are in genomic DNA

Gene Expression Analysis: Several techniques, viz., DNA micro array, Serial analysis of gene expression (SAGE), Massive parallel signature sequencing (MPSS), etc are applied. Clustering algorithms are used to determine gene co-expression.

Protein Sequence Analysis: Protein micro arrays & HT mass spectrometry (MS) provide information of protein in biological sample.

Structure Prediction: Structure prediction of different macromolecule is performed. Prediction of protein structure is made by comparative modeling and fold recognition.

Computational Evolutionary Biology: Origin and descent of species and their changes over time may be determined by suitable software. Phylogenetic analysis of protein and nucleic acid is also done. Development of phylogenetic tree, binary tree or cladogram, has been developed.

Modeling Biological Systems: Computer simulation of different system can be made. Different physiological processes, e.g., signal transduction, enzymatic pathway, gene regulatory network, have been simulated.

Preserving Biodiversity: Information of biodiversity, viz., name of species, description, distribution, population size, habitat etc. have been computerized.

Ex: Species 2000 project (internet based global research)

Biomedical Informatics: There are various applications in this field- image analysis, clinical image analysis and visualization, quantification of occlusion size, real time flow pattern, analysis of pictures of embryo & fate of cell clusters, analysis of mutation in cancer etc.

Dietary Analysis System: Applications in this field include computerized data analysis of dietary pattern, nutrient database (Ex: USDA, INFOODS, PDS).

Biological information is stored on many different computers around the world. The easiest way to access this information is for the computers to be joined together in a network. A computer network is a group of computers that can communicate, for example over a telephone system, therefore allowing data to be exchanged between

remote users. For transfer, biological data are first broken into small packets (units of information), which are sent independently and reassembled when they arrive at their destination. If information is sent from one computer (say, A) to another computer (say, C), it can travel via two different routes. In one case computer B acts as a router, and in the other case computers D and E both act as routers. The availability of different routes through the network means that communications can be maintained between computers A and C even if part of the network is unavailable, for example if computer B ceases to function.

Having got the feel of bioinformatics on the WWW, what are the merits and demerits of installing software locally, rather than using a WWW site? Although locally installed software will usually run faster than the same application used over the Internet, some software is difficult to install and might need expert help. There are advantages in having local copies of simple sequence alignment and other software if you are working 'at home' that is, limited by rates of data transmission on telephone lines. However, the use of locally installed databases can be disadvantageous because updates will be published less frequently than the WWW-based versions. Many academic institutions have an Intranet, that is, a local network that can be accessed only from computers within the institution. Such local networks may provide a number of bioinformatics tools and applications, which will usually run just as fast as locally installed software.

10.2 Searching the Internet:

Although the nine web sites provide some of the best starting points for bioinformatics on the WWW, there is a great deal of specialist biological data that cannot be accessed directly from these sites. Finding relevant data on the Internet is made simpler by the availability of general-purpose **search engines**, such as Google, Yahoo, Lycos, Alta Vista and Hotbot. These tools search the entire Internet for pages that contain particular keywords or phrases, and they can also be used to search for files of a particular type, such as image files or video files. For example, one might search the Internet using the phrase 'alcohol dehydrogenase' to find pages containing information about that enzyme. Alternatively, one might look for image files of a particular insect or flower, or video

files of frog development. Relevant sites are displayed as a list of hits, with hyperlinks allowing direct access to the page of interest. The problem with general-purpose search engines is that they have not been developed specifically with molecular biology in mind, and the information they provide can be irrelevant or misleading, especially if the search term used has other connotations.

As an alternative to search engines, the home pages of academic institutions or biotechnology companies can also be a good place to start. Many universities, for example, maintain comprehensive web sites with pages for staff to describe research projects and display data, and such sites often contain hyperlinks to sites of related interest.

On a general-purpose search engine it is probably better to start with a set of key words that is very restrictive and then remove some of the words if no hits are generated. If a search term is too broad (e.g. 'biochemistry') it will produce a ridiculous number of hits and it will be impossible to check all the listed pages. Search terms with known alternative uses are also best avoided. For example, searching the Internet with the word 'steroid' will likely hit more pages on body-building than molecular biology. A positive suggestion is to use a literature database on the WWW such as Pub Med to look for useful and appropriate keywords and phrases to use as search terms.

10.3 Software frame works for bioinformatics:

Software is a collective term for the various different programs that can run on computers. Software is distinguished from hardware, which refers to physical devices such as the processor, disk drives and monitor. On a stand-alone computer, software is divided into two categories: system software and application software. System software essentially comprises the computer's operating system and any other programs required to run applications, while application software is installed by the user for specific purposes (e.g. word processing, image analysis, etc). On networked computers, programs can also be run remotely. The same applies to computers attached to the Internet.

10.4 Programs and programming languages:

Computer programs can be written in a variety of programming languages, which are conventionally described in terms of three levels. The first level is machine code, which

is the binary code used by the computer's own processor. The second level includes a number of languages known as assembly languages. The third and subsequent levels are grouped as higher-level languages and include widely used programming languages such as Pascal and C programs written in assembly and for higher-level programming languages must be converted into machine code before they will run. For assembly languages, this process is known as assembly, and for higher-level languages it is known as compilation. In Windows, files in machine code are known as executable files and have the extension .exe. There are no rules or conventions for naming such files in UNIX systems, and they are known as executable images. These files can be created and stored in the computer's memory until the operating system is told to run them. Alternatively, assembly or compilation can be carried out 'on the fly' if programs are executed remotely, for example over the Internet.

10.4.1 Scripts and scripting languages:

Executable files (Windows) are executable images (UNIX systems) are written in machine code and are run by the computer's processor. Other program files are designed to be executed by another program, and such files are known as scripts. There are variety of scripting languages that can be used, including Microsoft Visual Basic, JavaScript and PERL. Script languages are easier work with than compiled languages, but take longer to process, so they are ideal for short programs.

10.4.2 Popular languages in bioinformatics:

A number of programming, scripting and markup languages are popular with bioinformatics because they are versatile and can integrate a wide variety of types of data either in a stand-alone environment or over the Internet. Some of these languages are discussed below.

HTML and JavaScript

HTML is hypertext markup language, a language used to specify the appearance of a hypertext document, including the positions of hyperlinks. Since HTML is not a programming language, basic hypertext documents are static. JavaScript is a popular

scripting language that adds to the functionality of hypertext documents, allowing web pages to include such features as pop-up windows, animations and objects that change in appearance when the mouse cursor moves over them.

Java

Java is a versatile and portable programming language that is designed to generate applications that can run on all hardware platforms, from large servers to individual PCs, without modification. The Java source code is based on C++ and can be run in a stand-alone fashion or from within a hypertext document, in which case it is called an applet (small application). When executed, a Java program is converted into an intermediate language called byte code, which is compiled into machine code as the program runs. Browsers must incorporate a Java plug-in interpreter called Java virtual machine for this purpose. Java applets may take a long time to download but the performance of the applet is not dictated by activities of the server. Java is a full programming language and is not the same as the JavaScript, which is a scripting language. The names are similar because both languages use a similar syntax. As discussed above, JavaScript is used primarily to enhance World Wide Web (WWW) pages, while Java has a much broader scope.

PERL

PERL (Practical Extraction and Reporting Language) is a versatile scripting language, which is widely used in bioinformatics for applications such as the analysis of sequence data. PERL is a free product, providing compatibility with Windows, UNIX or other operating systems. It has excellent facilities for file handling and uploading and downloading files over the WWW.

XML

XML stands for Extensible Markup Language. This is a new standard markup language that allows files to be described in terms of the types of data they contain. As a replacement for HTML, XML has the advantage of controlling not only how data are displayed on a WWW page, but also how the data is processed by another program or by a database management system.

10.5 Running programs over the Internet:

Software does not have to be downloaded and installed on local computers but can be run over the Internet. This can be achieved in two ways. If the programs are client-side, they are supplied for example as JavaScript or Java applets that are embedded in HTML within a hypertext document. The utility of these programs might be limited by the capacity of the local machine. Furthermore, although both Internet Explorer and Netscape Navigator support JavaScript, the script is interpreted in slightly different ways by the two browsers. There is currently no clean solution to this problem. The alternative is to use common gateway interface (CGI) programs or Java servlets, in which case the software is run on the server itself (the programs are server-side). Server side programs can be written in machine code or in a scripting language such as PERL or Java. It is easy to detect whether the software is running on a server because the URL will typically end with the extension .cgi. The performance of CGI programs is dependent on the number of current users (the server load). Some servers avoid bottlenecks by carrying out client instructions (e.g. homology searches) in their own time and then e-mailing the results to the client.

10.6 DATABASE MANAGEMENT IN BIOINFORMATICS:

10.6.1 Flat files and markup languages:

On a computer system, a file is a discrete collection of bytes that can be manipulated (moved, copied, deleted etc.) as a single entity. A file may either constitute a program or a data file that is processed by a program (e.g. a document that can be read by Microsoft Word). In the context of bioinformatics, files are used to store structured biological data. Most raw biological data can be stored in the form of text, for example nucleotide and protein sequences, protein structural coordinates and matrices of gene expression profiles. Text files can be handled by various software applications such as text editors (e.g. Simple Text), Internet browsers (e.g. Internet Explorer, Netscape Navigator) and word processor applications (e.g. Microsoft Word, Corel WordPerfect). Other types of biological data are stored as images, for example gene expression patterns and pictures of

two-dimensional-protein gels. In some cases, the raw data in the images are converted into numbers that can be stored in text files. For example, this is the case for micro array image data.

Most software applications that handle text include a markup language that specifies how the text should be displayed on screen or in a printed document. These instructions comprise hidden character sets, known as tags. In Microsoft Word, for example, the markup language controls the font, size, color, paragraph structure etc. of the text. Other familiar markup languages include HTML (hypertext markup language), which controls the display of text on WWW pages and enables hyperlinks to be inserted, and XML (extensible markup language), which allows the integral description of data objects. Such tags are often transparent, however, if the text is used by another software application, such as a sequence analysis program. Therefore, it is best to save biological data files in a simple format with no markup language. These text only files are known as flat files. Text editor programs such as SimpleText and NotePad are suitable for handling flat files, and flat files can be generated in word processor programs such as Microsoft Word by saving as text only.

10.6.2 Databases:

A database is a collection of structured information, often stored in the form of flat files in the case of bioinformatics data. Individual database entries are known as records and each record comprises the same set of fields (categories of data). For example, in a sequence database such as GenBank, each record represents a deposited sequence and fields include accession number, sequence name, taxonomy of source species, literature references, and the sequence itself. Computer databases are usually associated with software that allows the information to be accessed, amended and searched. This software is known as a database management system (DBMS) and also controls the security and integrity of the data. Searches are made possible by indexing the records, which in the case of flat files is achieved by looking for text strings in particular fields. In the case of a sequence database, the accession number could be used for index purposes. Several different types of database are used in bioinformatics,

A relational database is organized as tables, each table comprising a group of records (also known as tuples) with the same fields (known as attributes). This allows related data to be linked (reassembled) as required without reorganizing the original tables. For example, a sequence table might contain records with the attributes *accession number* and *protein sequence*, while a function table might contain records with the attributes *accession number* and *protein function*. Matching attributes in different tables can be joined to bring together related records, in this case linking *protein sequence* and *protein function*. The industry standard language used to interrogate and process data in a relational databases is SQL (symbolic query language). This is incorporated into familiar and widely used relational database management systems such as Microsoft Access and Oracle.

An object-oriented database has a more flexible organization, that is, it does not depend on the formal 'table, row and column' format of relational databases. Data are defined as objects, which have a class hierarchy, that is, they can be grouped into classes and subclasses etc. in a hierarchical manner. Properties attributable to classes of objects are inherited through the hierarchy; these may be general in the upper levels of the hierarchy but may become more specialized in the lower levels. Properties or procedures attributable to data objects are known as methods. The flexibility of the data organization in object orientated databases allows more complex relationships between datasets to be modeled than is possible with relational databases. Object-oriented database management systems are also capable of handling multimedia objects (pictures, videos and audio files) while relational DBMSs are often restricted to numbers, alphanumeric text and dates. Pure object-oriented DBMSs include Object Store and ONTOS DB. ACeDB (A *C. elegans* database) is an example of a customized object-oriented database. It is more common to see bioinformatics databases incorporating relational DBMS features in an object-oriented programming environment. Such object-relational DBMSs are generally accessed using a language based on SQL.

Developments in object-oriented programming have led to attempts to have object definitions that are common across different computer systems. This is useful for the integration of distributed databases, that is, databases that are physically stored on two or

more separate computer systems. An interface definition called CORBA (Common Object Request Brokering Architecture) has been developed which can be used to integrate large distributed bioinformatics databases. Software such as CORBA that functions as a conversion or translation layer in distributed systems is sometimes called middle ware.

11.0 CONCLUSION:

Computer is essential for every field of biological sciences, including botany. The students of biological science should learn different aspects of computer and extract benefit from it. The knowledge of hardware will help them to operate the computer in a better way and enable them to rectify minor hardware related problems. If they can learn software they may be able of writing application software for solving small computations. The usage of internet will be a great help for acquiring subject related information from different websites. Bioinformatics is comparatively new field; however, it has immense potential for guiding research and development.

12.0 SUMMARY:

Computer is an electronics device capable of computation with high degree of accuracy within a very time. The computer is composed two main components – hardware and software. The hardware consists of CPU and peripheral devices. The CPU has three subunits – control unit, arithmetic-logic unit, and primary memory. The primary memory consists of Read Only Memory (ROM) and Random Access Memory (RAM). There are several types of software – monitor program, operating system, utility program, language processor and application program. There different kinds of operating system, e.g., DOS, Windows, UNIX etc. Windows is the most popular operating system. The Windows system works with desktop, which contains different elements like, 'my computer', files, folders, title bar, scroll bar etc. Internet is global connection of computers through telephone lines. It requires a computer, modem, ISP, and telephone line. One can enter into the network through the WWW, the World Wide Web. Each website has an URL (uniform resource locator). The information can be accessed by means of a web browser like, Internet Explorer. Search engine, such as, Google or Yahoo search etc. Different

kinds of software help to down load desired information. Bioinformatics is the combination of informatics and biological sciences. The computer and internet are essential for bioinformatics. Different languages, viz., Java, PERL, XML, HTML, etc are popularly used in bioinformatics. Bioinformatics can be applied in various field including, sequence analysis of protein, DNA, RNA etc, prediction of structure of macromolecule, genome analysis, preservation of biodiversity, phylogenetic study and in other fields. Different kinds of databases are formed in bioinformatics and various types of software are used for the management of database.

13. BIBLIOGRAPHY:

1. Rajaraman V.: Fundamentals of Computers. Printice Hall of India Pvt. Ltd., New Delhi, 2001.
2. Westhead D.R., Parish J.H., Twyman R.M.: Bioinformatics. Viva Books Pvt. Ltd., New Delhi, 2003.
3. Gupta V.: Comdex computer course kit. Dreamtech Press. New Delhi. 2001.
4. Sen S., Sarkar. S.K., Sarkar S.C.: Foundation of computer science and programming. Grantha Bharati, Kolkata, 1996.
5. Sampat S. and Wasan S.K.: Basic programming Macmillan India Limited, Bangalore, 1988.
6. Stalling W.: Computer organization and architechture. Macmillan, New York, 1987.

14.0 MODEL QUESTIONS:

14.1 Short questions:

1. State the functions of ALU?
2. What is semiconductor memory?
3. Mention the function of optical character recognizer?
4. What is the advantage of using EPROM?
5. What are the components of bioinformatics?
6. What do you mean by hard disk?
7. What is search engine?.

8. What are the difference between source program and object program?
9. What is assembly language?
10. What are the requirements of internet?
11. What is modem? State the features of different types of modem.
12. How can you download data by FTP server?

14.2 Long questions:

1. Discuss the basic structure of the computer with a diagram.
2. What do you mean by RAM and Rom? Discuss different types of RAM and ROM.
3. What do mean by I/O devices? Described different kinds of input devices of a computer.
4. State the importance of printers. Discuss different types of computer printers mentioning their advantages and disadvantages.
5. What is meant by computer software? Discuss, in brief, different types of software with their importance.
6. Discuss in brief about different languages used for bioinformatics.
7. Discuss the database management in bioinformatics.
8. Discuss the importance of WWW for getting information from internet.

M. Sc. in Botany

Part -I

Paper - II

Module No . 18

Contributor :

Dr. M. M. Pal

Module No. - 18

Module Structure

Section 1: Microsoft Word 2000

- 1.1. Introduction
- 1.2. Starting word 2000
- 1.3. Menus and toolbars
- 1.4. Add or remove buttons menu
- 1.5. Arranging buttons on the toolbar
- 1.6. Creating, editing and saving a word document
- 1.7. Deleting text
- 1.8. Automatic spelling and grammar checking
- 1.9. Disabling the automatic spelling and grammar checker
- 1.10. Autocorrect feature
- 1.11. Formatting marks
- 1.12. Naming and saving document
- 1.13. Opening an existing document
- 1.14. Choosing a view
- 1.15. Navigating the document
- 1.16. Scrolling through a document by using the mouse
- 1.17. Scroll through a document using the keyboard
- 1.18. Go to a specific location or item using go to feature
- 1.19. Find text in a document
- 1.20. Cut, copy and paste text
- 1.21. Undoing and redoing changes
- 1.22. Keyboard shortcuts
- 1.23. Applying formatting to text
- 1.24. Additional text effect
- 1.25. Format painter
- 1.26. Changing paragraph alignment
- 1.27. Changing the paragraph spacing
- 1.28. Bulleting and numbering
- 1.29. Add borders and shading to a paragraph
- 1.30. Adding header and footer
- 1.31. Create special effects with text

- 1.32. Printing in multiple columns
- 1.33. Link document
- 1.34. Creating a table
- 1.35. Working with graphics
- 1.36. Mail merge
- 1.37. Previewing and printing a document

Section 2: Excel 2000

- 2.1. Introduction
- 2.2. Excel 2000
- 2.3. Opening of Excel 2000
- 2.4. Opening a file
- 2.5. Saving a Workbook
- 2.6. Workbook
- 2.7. Executing commands
- 2.8. Wizards
- 2.9. Creating and using templates
- 2.10. Working with Worksheets
- 2.11. Entering Date and Time
- 2.12. Entering data in a series
- 2.13. Manipulating cell contents
- 2.14. Formatting rows and columns
- 2.15. Password-Protecting a Workbook
- 2.16. Ranges
- 2.17. Quit from Excel
- 2.18. Formula
- 2.19. Entering a formula
- 2.20. Naming cells
- 2.21. Functions
- 2.22. Formatting
- 2.23. Printing a Worksheet

Section 3: Programming in C

- 3.1. Introduction
- 3.2. The C character set
- 3.3. Constant data
- 3.4. Variables and arrays
- 3.5. Declarations
- 3.6. Expressions
- 3.7. Mathematical functions
- 3.8. Assignment statement
- 3.9. Input/Output statements
- 3.10. Complete programs
- 3.11. Control statements
- 3.12. Decision making and branching
- 3.13. Worked out examples
- 3.14. User-defined function

Module summary

Self assessment questions

References

This module is divided into three sections to discuss three important topics - MS-Word, MS-Excel and C language. In section 1, the MS-Word is introduced in very simple way and discussed its most common topics. The Section 2 is devoted to MS-Excel. The basic features of Excel 2000 are discussed here. In last section, an introduction of C programming language is given. This section will help the learner to write simple programs.

Objectives

Go through this module you will learn three main topics MS-Word, MS-Excel and programming in C.

- Opening and working with MS-Word
- Formatting of a Word document
- Printing of a Word document
- Opening and working with MS-Excel
- Use of functions
- Formatting of a Excel workbook
- Printing of a Excel workbook
- Basic concept of C programming
- Input/output statements
- Branching statements
- Loop statements
- Writing simple programs

Key-words

MS-Word, MS-Excel, C programming language.

Section 1: Microsoft Word 2000

1.1. Introduction

Microsoft Word is one of the most popular amongst the word processing packages available in the market. Like any other word processor, the Word can create a new document such as letter, memo, article, report or can modify a document that is already created earlier. Microsoft Word may be used to type a text, edit existing text and format text to add emphasis, highlight ideas, arrange text attractively on the page and to insert graphics, tables and charts as well as check your document for spelling and grammatical mistakes. Also Microsoft Word can use to create and modify web pages that you can display on the Internet.

1.2. Starting word 2000

There are several ways to start Word 2000. A list of possible ways are given below:

- Click on the Office 2000 shortcut bar.
- Click the **Start** button on the Windows taskbar. The **Start** menu appears. On the **Start** menu, click **Programs** and then click on the **Microsoft Word** program to start Word 2000.
- If the shortcut for Word 2000 is created earlier, then double-click on it to start Word.
- If the shortcut for Word 2000 is created in the Quick Launch bar then click on it to start Word.

Once you have started Word 2000, the application opens a new blank document as shown in Figure 1.1.

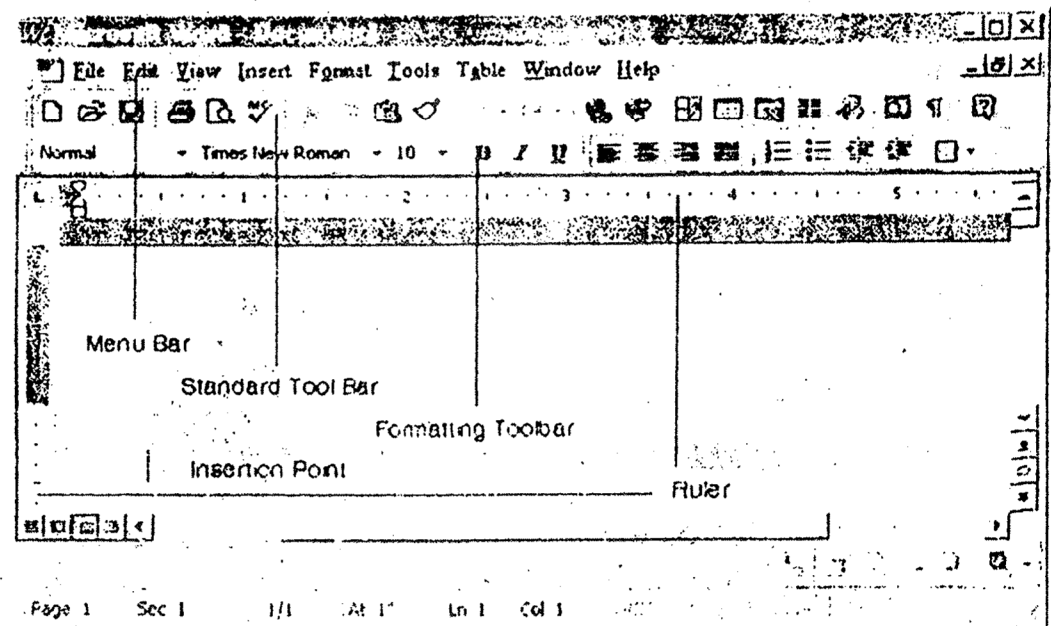


Figure 1.1: Word Opening Window

1.3. Menus and toolbars

The Word opening screen contains a blank document, a menu bar and a toolbar in addition to the usual Windows buttons. The menu bar organizes the commands in a logical manner, making it

easy for you to access the features which are needed. For example, the commands related to table creation and formatting are grouped under the Table menu. The menu bar as we have seen contains a list of commands. The toolbar contains buttons in a horizontal or vertical row across the top or on the side of the document window. Each button has a picture or icon on it corresponding to a command.

The Word 2000 default setting for toolbars displays the most commonly used buttons for the **Standard** and **Formatting** toolbars in a single row. By displaying only the frequently used buttons in a single row, Word makes more space available for typing document. Word 2000 is a very intelligent application. Once you use a command, its button is added to the toolbar, replacing another that is less often used.

For resetting the toolbars and menus do the following steps:

1. **View → Toolbars → Customize.** Click **Customize**. The **Customize** dialog box will appear.
2. Click the **Options** tab.
3. Click **Reset My Usage Data** button and click **Yes** when prompted.
4. Click close.

1.4. Add or remove buttons menu

If the button that you need is not available on the toolbar, you can find it by pressing the **More Buttons** drop-down arrow. After locating the button, click on it to add it to the toolbar. Similarly, you can remove a toolbar button from the toolbar. To do this follow the following steps:

1. Click on the **More Buttons** drop-down arrow.
2. Click **Add or Remove Buttons** button.
3. Uncheck the buttons you want to remove from the toolbar. These will be removed from the toolbar. Check the buttons you want to add to the toolbar. Such buttons will be added.
4. If an arrow appears at the bottom of the list, click on the arrow. Additional toolbar buttons will be displayed.
5. Click anywhere on the document to close the menu.

1.5. Arranging buttons on the toolbar

The order of the toolbar buttons can be changed. To arrange the toolbar buttons, do the following steps:

1. **View → Toolbar → Customize.** The customize dialog box appears.
2. If the **Customize** dialog box appears over the toolbar that you want to arrange, drag the dialog box away from the toolbar by clicking on the title bar at the top of the box and without releasing the mouse button, move the dialog box to the desired position and then release the mouse button.
3. On the toolbar drag the button you want to move and place it in the position you want. If the position is valid, then an I-beam will appear. If you drag the toolbar button to an area outside the acceptable positions, then instead of the I-beam an X mark will appear at the bottom of the mouse pointer.
4. After moving all the toolbar buttons to the desired positions, click the **Close** button.

1.6. Creating, editing and saving a Word document

The typing of any document is similar to typing on typewriter. The short vertical, blinking line at the top of the new document is the **insertion point**. It indicates where the text will be entered as you type. After typing a few lines or completing your typing you can move the insertion point (using either the mouse or by arrow) to edit the text anywhere in the document. One point should

be remember that when you typing a long sentences, do not press the Enter key at the end of each line. The line break will be done automatically. This feature is called word wrapping.

Now we will type a document, say, an invitation letter and it will be sent to all friends.

Suppose, a new blank document is open. If it is not open, press the **New** button on the **Standard Toolbar**. Alternatively, use the menu. Press **File → New → Blank Document** and press **OK**. Remember that the use of toolbar button is more faster.

Type **P.K.Sen** and press **Enter**. The insertion point moves to the next line.

Type **Department of Physics** and press **Enter**.

Type **Vidyasagar University, Midnapore – 721 102** and press **Enter**.

A red/green wavy line under a word means that the automatic **Spelling and Grammar Checker** is active. It identifies the possible spelling mistakes and puts a red wavy line under the word. Probable grammar mistake are indicated using a green wavy line under the words or sentence. For now ignore the red/green lines.

Then type the body of the letter.

1.7. Deleting text

Using the **Del** key or the **Backspace** key you can delete the text. The **Backspace** key deletes the text to the left of the insertion point and the **Del** key deletes the text to the right of the insertion point. So, it is important to position the insertion point at the appropriate position in the document. The mouse pointer may be used to move the insertion point.

1.8. Automatic spelling and grammar checking

The red/green wavy lines may occur under some portion of the text. These lines indicate possible spelling and grammatical errors. These lines occur when the automatic spelling and grammar checking feature is active. After finishing the typing you can go back to edit the document. To do this, right-click the underlined word or words. A shortcut menu is displayed, giving you a list of words that according to Word are possible alternatives. You may ignore the suggestions if you need. The words like names, that you frequently use are also underlined as spelling mistakes. In such cases you can either ignore the suggestion or add the word to the custom dictionary. Once if you added a word to the custom dictionary, Word will not underline it in the future. To add a word to the custom dictionary, click **Add** on the shortcut menu.

In the following, the automatic spelling and grammar checking feature of Word is described in details.

1. Press **ctrl + Home** to move to the top of the document. Or use mouse to go to the top of the document.
2. **Right click** on the incorrect word. Then **Automatic Spelling and Grammar Checker** shortcut menu will displayed.
3. Click **Spelling**. The Spelling dialog box displayed.
4. Click **Ignore**. The red wavy line disappears. The **Spelling and Grammar Checker** will ignore this occurrence of the word. The second occurrence will still be underline.
5. If a word that is underlined occurs more than once then click **Ignore All** in the shortcut menu. The red wavy line disappears from all the occurrences.

1.9. Disabling the Automatic Spelling and Grammar Checker

The Automatic Spelling and Grammar Checker feature can be turned off through the following steps:

1. **Tools → Options.** The Options dialog box will appear.
2. Click **Spelling and Grammar** tab.
3. In the Spelling area, uncheck **Check spelling as you type** check box.
4. In the Grammar area, uncheck **Check grammar as you type** check box.
5. Click **OK**.

1.10. AutoCorrect feature

Generally, when you type a document you might make many mistakes. But when you return for correcting them, many of them would be already corrected. This is due to the AutoCorrect feature of Word. This Word feature corrects most of the common typographical errors. Some of the common examples are 'adn' instead of 'and', 'teh' instead of 'the', etc. As soon as you type a space or begin a new paragraph after the misspelt word, Word recognizes the error and corrects it.

You can see the entries in the AutoCorrect feature by going to **Tools → AutoCorrect**.

1.11. Formatting marks

Whenever you create a document by typing, special characters called formatting marks are inserted into the document. The two most common formatting marks are the paragraph mark (§), which is placed in the document every time you press the Enter key and the space mark (.), which is inserted every time you press the space bar. The formatting marks can either be displayed or hidden on the screen, but when the document is printed these are not printed.

The formatting mark helps you in troubleshooting the document while you edit it. You can use these marks to identify extra lines between paragraphs, forced Enter keystrokes at the end of the line spaces between words and so on.

To display the formatting marks, click the **Show/Hide (§)** button. The formatting marks are displayed in your open document.

1.12. Naming and saving document

After finishing the work you have to save it for further use, such as, modification, printing etc. You can save the document to hard disk or floppy disk or to another computer's hard disk if the computer is connected to a network. You can save documents in a different formats – Word 2000 document, web page, text document and so on.

If you save your document for first time then you have to give a name to the document. If you are saving an existing document in another name, then also, you have to give a name to the document.

In the following the saving of a document is illustrated stepwise.

1. Click the **Save** button on the **Standard toolbar**. Since this is the first time that you are saving the document, the **Save As** dialog box appears.
Alternatively, you can use the menu as
File → Save or **File → Save As**.

2. In the **Save in** box, click on the drop-down arrow and select your hard drive.
3. In the list of folders, double-click on the folder that you want to save the document. The folder opens.
4. In the **File Name** box, type the file name you want. The file name can be up to 255 characters in length and can contain alphabets, numbers, spaces and other characters except the slash (/).
5. By default the document will be saved in Word 2000 - '*.doc' file. This can be seen in the **Save as type** box. If you want to change the format click on the drop-down arrow to see the different available formats.
6. Click **Save** to save the document.

1.13. Opening an existing document

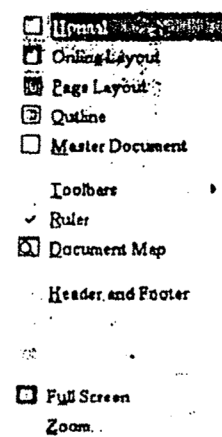
When you start Word, it opens a blank document. Now we will see how to open an existing document. We have saved the invitation letter to the hard disk. Suppose you have saved it in C: drive in a folder named **examples**. Let us assume that the name of the file is **invitation.doc**. We will see how to open that file and save it under a new name so that the original file remain unchanged.

To open and save the file, do the following steps:

1. Click the **Start** button on the **Windows taskbar**. The **Start** menu appears.
2. On the **Start** menu point to **Programs** and then click on **Microsoft Word**. Word opens with a blank document.
3. On the **Standard toolbar** click the **Open** button. The **Open** dialog box appears.
4. Click on the **Look in** drop-down arrow and then select (C:) drive. The folders in the C drive are listed.
5. Double-click on the folder named **examples**. The folder opens showing the word documents in that folder.
6. Double-click on the file named **invitation.doc** to open it. The document opens in the document window.
7. **File** → **Save As**. The **Save As** dialog box appears. The **examples** folder will be in the **Save** box. In the **File name** box, the name will be **invitation.doc**.
8. Select the file name if it is not already selected by clicking on it.
9. Type say, **invitation1**.
10. Click **Save**.

1.14. Choosing a view

Word provides a variety of views or display formats that you can use to view the document while working. The choice of the view depends on the type of document and the way want to work with it. The **Normal view** helps to focus on composition, text revisions and basic formatting such as font effects (bold, italic, etc.) without worrying too much about the layout of the page. When you are applying more sophisticated formatting or moving and copying text it will be better to work in the **Print Layout view**. This is also the default view. **Web Layout view** shows you how your web page will be displayed in a browser. **Outline view** lets you focus on document organization by highlighting headings and subheadings. **Full Screen view** fills the window with your document without displaying the toolbars or controls. All these views can be selected from the **View** menu.



1.2: The View Menu

We consider the **Print Layout** view in this note.

1.15. Navigating the document

Once the document has been created, then may want to navigate to a particular page in the document. Word has navigational and selection tools using which we can go to any page within the document regardless of how big the document is and the document can be edited.

The methods that are available in word for navigational and editing purposes are given below:

- Scroll through a document
- Go to specific item or location
- Find Text

1.16. Scrolling through a document by using the mouse

The horizontal scrollbar and the vertical scrollbar are one of the tools that are used for navigating within the document. Table 1.1 shows how to scroll through a document using the mouse.

To	Do this
Scroll up one line	Click the up scroll arrow
Scroll down one line	Click the down scroll arrow
Scroll to a specific page	Drag the scroll box
Scroll left	Click the left scroll arrow
Scroll right	Click the right scroll arrow
Scroll left, beyond the margin, in Normal view	Hold down SHIFT and click the left scroll arrow

Table 1.1

1.17. Scroll through a document using the keyboard

Apart from using the mouse for scrolling through the document, the keyboard can also be used. Table 1.2 lists the keys that are used for moving from character to character, word by word, line by line, paragraph to paragraph, window to window, screen to screen, etc.

Press	To move
LEFT ARROW	One character to the left
RIGHT ARROW	One character to the right
CTRL+LEFT ARROW	One word to the left
CTRL+RIGHT ARROW	One word to the right
UP ARROW	One character up
CTRL+DOWN ARROW	One paragraph down
SHIFT+TAB	One cell to the left (in a table)
TAB	One cell to the right (in a table)
UP ARROW	Up one line
DOWN ARROW	Down one line
END	To the end of a line
HOME	To the beginning of a line
ALT+CTRL+PAGE UP	To the top of the window
ALT+CTRL+PAGE DOWN	To the end of the window
PAGE UP	Up one screen (scrolling)
PAGE DOWN	Down one screen (scrolling)
CTRL+PAGE DOWN	To the top of the next page
CTRL+PAGE UP	To the top of the previous page
CTRL+END	To the end of a document
CTRL+HOME	To the beginning of a document

Table 1.2

1.18. Go to a specific location or item using GO TO feature

The **GO TO** feature is used to move in a document from one place to another place. This is called navigation. Generally, we use keyboard or mouse to move the cursor from one place to another place within the page. However, while working in a big document, using the keyboard or mouse will not be very efficient and in such cases **GO TO** feature will be very useful. This saves a lot of time and also makes the user's work easy and simple. This tool is specifically used for moving from page to page or to move to a couple of pages from the page, wherever in a document.

To display the **Go To** dialog box do the following steps:

1. On the **Edit** menu, click **Go To**.
2. A dialog box will appear on the screen and in the **Go To what** box, click the type of item.

3. Type the name or number of the item in the Enter box and then click **Go To**.
4. To go to the next or previous item of the same type, leave the **Enter** box empty and then click **Next** or **Previous**.

1.19. Find text in a document

Find and Replace feature is another navigational tool in which the user can navigate or search through pages based on some specific word or phrase. This tool is very useful when we want to search for some text and further change or replace that text with some other text. To find in a document, perform the following steps:

1. On the **Edit** menu, click **Find**. A menu appears.
2. Type the text to be searched in the **Find what** field.
3. Select any of the other options by clicking the **More** button.

1.20. Cut, Copy and Paste text

We want to make some changes to the letter. First, we want to list the names of the publishers in the alphabetical order. The sentence we want to change is given below:

We are distributed many important books of the leading publishers like McGraw-Hill, Addison Wesley, Prentice-Hall, Artech House, etc.

This can be done by moving Addison Wesley and then Artech House before McGraw-Hill. To do this follow the following steps:

1. Scroll down to the area where the words Addison Wesley appear.
2. Position the mouse pointer before the word Addison.
3. Drag the mouse pointer holding the left mouse button till you have selected the word Addison and Wesley and the comma. The selected words appear in a black background.
4. Click on the selection and drag (while holding the mouse button down) and move the pointer to the position before the word McGraw-Hill. When you click the selected section, the mouse pointer changes to a left pointing arrow instead of the I-beam. While moving a dotted rectangle appears on the tail of the mouse pointer and a dotted vertical line appears on the tip of the mouse pointer.
5. Release the mouse button when the dotted vertical line is positioned before the word McGraw-Hill. The words Addison Wesley and the comma will be shifted to a position before the word McGraw-Hill.
6. Repeat the above steps with the words Artech House and the comma. Now you will see the sentence changed as follows:

We are distributed many important books of the leading publishers like Addison Wesley, Artech House, McGraw-Hill, Prentice-Hall, etc.

Now we will use cut and paste for the same purpose. Do the following steps:

1. Scroll down to the area where the words Addison Wesley appear.
2. Position the mouse pointer before the word Addison.

3. Drag the mouse pointer holding the left mouse button till you have selected the word Addison and Wesley and the comma. The selected words appear in a black background.
4. Click the **Cut** button on the Standard toolbar. The words 'Addison Wesley', disappear. It is also copied to the **Clipboard**.
5. Move the insertion point to the position before the word McGraw-Hill.
6. Click the **Paste** button on the **Standard toolbar**. The words Addison Wesley and the comma will now appear before the word McGraw-Hill.
7. Repeat the above steps with the words Artech House and the comma.

If you want to use the same text matter more than once in a document, you can copy the text and then paste it whenever required. This will save a lot of time if you want to use a word many times in your document. To copy a word or sentence, select it and click on the **Copy** button on the **Standard toolbar**. The text is copied to the Clipboard and will be available until a new text is copied thereto. Whenever you want to type in the text that you have copied, place the insertion point where you want text to appear and click on the **Paste** button. It will appear in the new position.

To	Keyboard shortcuts
Cut	CTRL+X
Copy	CTRL+C
Paste	CTRL+V
Move	CTRL+X or CTRL+V
Delete	BACKSPACE or DEL
Undo	CTRL+Z
Redo	CTRL+Y

Table 1.3

1.21. Undoing and redoing changes

You can undo and redo changes after you make them by using the **Undo** and **Redo** buttons on the **Standard toolbar**. The **Undo** button reverses your last action. You use the **Redo** button to reverse an Undo action. For example, if you delete a word and then click Undo, the word reappears. If you then click Redo, it will be deleted again.

Word's undo and redo features are very powerful. You can reverse more than one action. When you click the Undo drop-down arrow, you can see a list of actions that you reversed. The list will have the most recent changes at the top and the previous changes below.

1.22. Keyboard shortcuts

By using keyboard shortcuts we can save an enormous amount of time. Working with the keyboard is faster than working with mouse. Keyboard shortcuts are keystrokes that activate a command directly and they bypass the menu. Keystrokes that are used to open a menu, such as ALT plus a key, are referred to as accelerator keys.

The shortcut key for each tool is displayed in the **ScreenTips** boxes that appears when the user points at an icon on a toolbar. If the **ScreenTips** box does not appear, it can be

activated by choosing **Tools** menu and by selecting **Customize** item. In the **Customize** dialog box, click the **Options** tab. In this tab, select the **Show Shortcut Keys** in **ScreenTips** check box.

<i>Task</i>	<i>Key combination</i>
Create a new document	CTRL+N
Open a document	CTRL+O
Save a document	CTRL+S
Help	F1
Make text bold	CTRL+B
Make text italics	CTRL+I
Underline text	CTRL+U
Left align text	CTRL+L
Center align text	CTRL+E
Right align text	CTRL+R
Justified text	CTRL+J
View normal	CTRL+ALT+N
View outline	CTRL+ALT+O
Print Preview	CTRL+ALT+I
Print	CTRL+P
Undo	CTRL+Z
Redo	CTRL+Y
Closing a document	CTRL+W

Table 1.4

1.23. Applying formatting to text

Using the formatting attributes you can change the appearance of text, which are available on the **Formatting toolbar**. Commonly used attributes include **bold**, **italic** and **underline**. You can change the font style and its size with the click of a button. A font is typeface applied to text, numbers and punctuation. There are many fonts that you can choose from font list.

In Word 2000 the font list is enhanced to display the name of the font in its own typeface so that you can preview it before selection. Word provides more complex formatting like drop cap, background, themes, etc. which are available from the **Format** menu.

In the following exercise we will apply some basic formatting to the **invitation.doc**. Do the following steps and show the changes in Figure 1.3.

1. Open the file **invitation.doc**.
2. Save it as **sample1.doc**.
3. In the first line select **Mr. P.K.Sen**.
4. On the **Formatting toolbar**, click the **Bold (B)** button.
5. Select the sentence **We are the distributors of many of the leading publishers like McGraw-Hill, Addison Wesley, Prentice Hall, Artech House, etc.**
6. Click on the *italic (I)* button on the **Formatting toolbar**.
7. Click on the Underline (U) button.

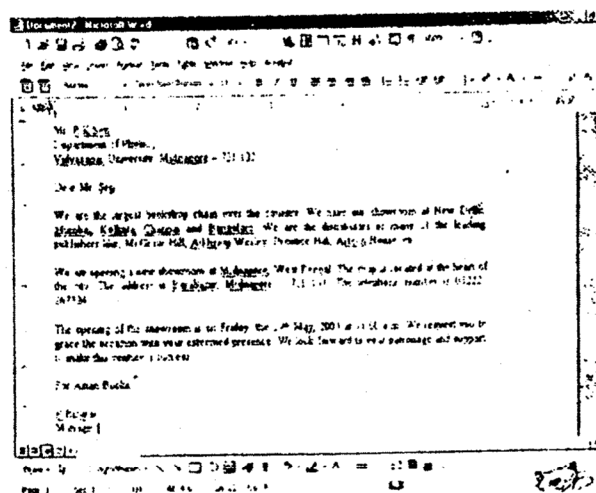


Figure 1.3: Basic Formatting

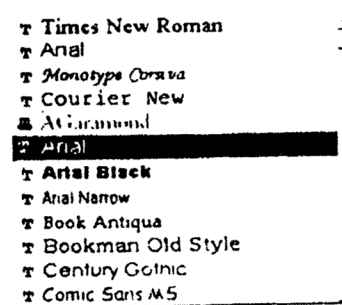


Figure 1.4: Font list

8. Select the second paragraph
9. Click the **Bold** button and then the *italic* button.
10. Select the last lines **P.Biswas** and **Manager**.
11. Click the **Bold** button.

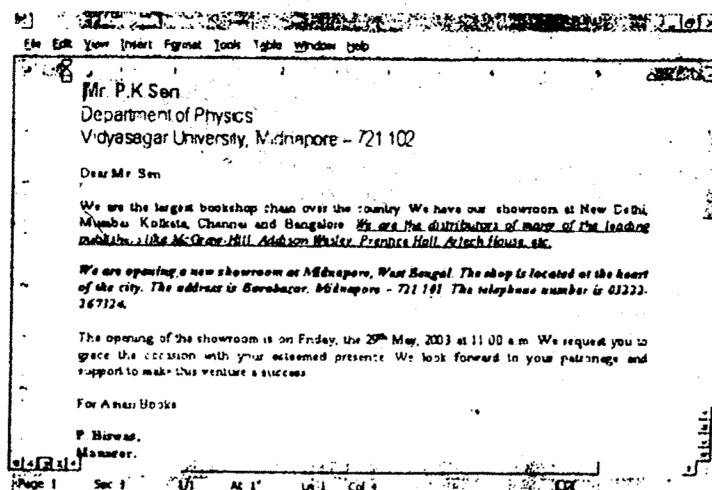


Figure 1.5: More formatting

In the following exercise we will change the font style and size of the document **invitation.doc**.

1. Open the document **invitation.doc**.
2. Save it as **sample1.doc**.
3. On the **Formatting toolbar** click the font drop-down arrow to see the list of fonts. The font list appears as shown in 4.
4. Browse through the font list to see the different options available.
5. Click outside the font list. The list will disappear.
6. Select the entire text of the document. You do it using mouse or using **Edit → Select All** or **ctrl + A**.
7. Again click on the font drop-down arrow.
8. Click on **Arial** from the list. The entire text changes to **Arial** font.
9. Click on the font size drop-down arrow and click 10. The size of entire text changes to 10 points.
10. Now select the first 3 lines of the text.
11. Change the font to **Arial Narrow**.
12. Click on the font size drop-down arrow and click 14. The size of the first 3 lines changes to 14 points.

Now your document becomes to Figure 1.5.

1.24. Additional text effect

Not all the formatting options are available on the **Formatting toolbar**. You can find additional effects on the **Format** menu. Effects used to animate text are also available in Word. These effects are very useful when you are designing document for the web. These effects are available from **Format → Font → Text Effects**. In the **Font** dialog box, go to the **Text Effects** tab. There you will find effects like

- (none)
- Blinking Background
- Las Vegas Lights
- Marching Black Ants
- Marching Red Ants

- Shimmer
- Sparkle Text.

If you click on each you can see how it looks like in the preview window.

In the **Font** dialog box, click the **Font** tab. There are a lot of special effects available like **Shadow**, **Outline**, **Small Caps**, etc. You can view these effects in the preview window.

1.25. Format painter

If you are using the same formatting several times in a document you can use the **Format Painter** instead of selecting the text and applying the formatting each and every time. To use the **Format Painter** first select the text that has the formatting you want to apply to other text selections. Once you activate the **Format Painter**, all the formatting attributes of the selected text will be attached to your pointer. Double click the **Format Painter** button, if you are going to copy formatting to several locations or just click the button if you are going to copy the formatting only once.

1.26. Changing paragraph alignment

Paragraph alignment is an important factor and it improves the look of a document. Paragraph alignment refers to the arrangement of the paragraph between the left and right margins. There are four ways to align a paragraph. Align left means that all lines on the left side of the paragraph are aligned with the left margin, while the lines on the right side end at different places. The default alignment of Word is aligning left. Align right is just opposite to align left. Align center means that the text is centered in the middle of the page. Text that is justified is aligned with both the left and right margins by spreading the words evenly between the margins.

If you change the alignment of a single paragraph you do not need select the entire paragraph. Word automatically recognizes that you want to apply an alignment style to that paragraph only. Just position the insertion point anywhere in the paragraph you want to align and click on the desired align button on the toolbar.

1.27. Changing the paragraph spacing

Changing the paragraph spacing the look of the document may be changed. You can choose from **single line spacing**, **1.5 lines**, **double line spacing** and so on.

To change the line spacing, follow the steps:

1. Open the document invitation.doc.
2. Save it under a new name, say, **sample2**.
3. Place the insertion point anywhere in the first paragraph.
4. **Format → Paragraph**. The **Paragraph** dialog box appears.
5. In the **Line spacing** area, click on the drop-down arrow and click on **Double**.
6. Click OK. The first paragraph spacing changes to double line spacing.
7. Select the three paragraphs.
8. **Format → Paragraph**. The **Paragraph** dialog box appears.

9. In the **Spacing** section, in the **Before** box, change the 0 pt to 18 pt either by typing it or clicking on the arrows.

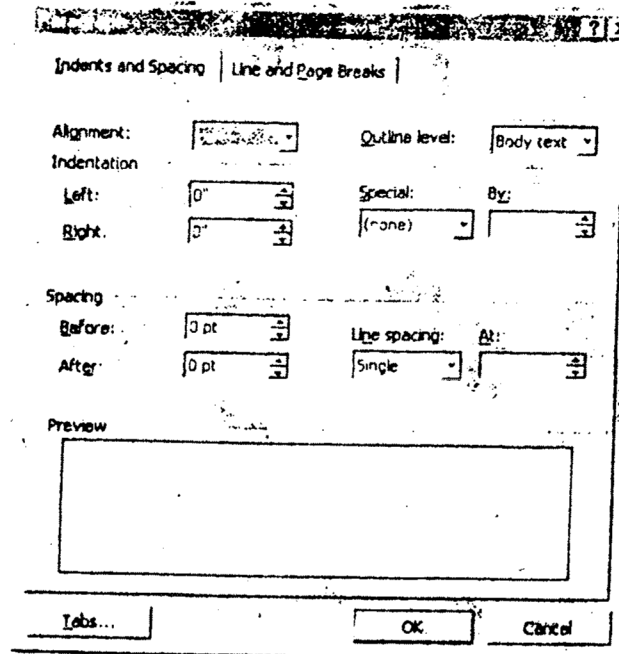


Figure 1.6: Paragraph Spacing

10. Click OK. The spacing between the paragraphs (spacing before each paragraph to be precise) increases to 18 points.

1.28. Bulleting and numbering

Bullets and numbers are another way of improving the readability of the text in a document. In Word, adding bullets and numbers is as clicking a button.

We will see how to use these features in the following:

1. Open **invitation.doc** and change it in another name.
2. Edit the text after first paragraph as follows:

We are the distributors of many of the leading publishers including:

**McGraw-Hill
Addison Wesley
Prentice Hall
Artech House**

We are official distributor of the standard of the following organizations

**IEEE
ISO
ANSI
BSI**

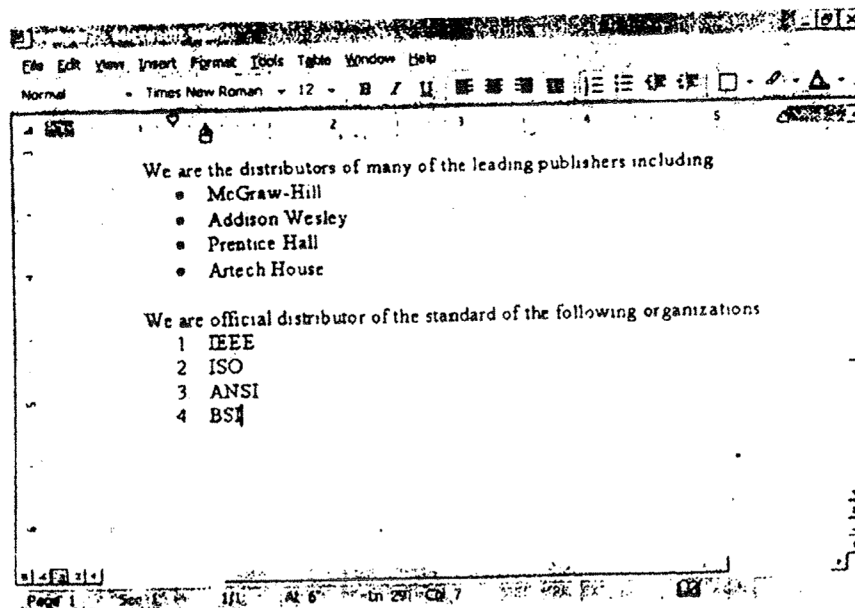


Figure 1.7: *Bullets and Numbering*

3. Now select the text from **McGraw-Hill** to **Artech House** (4 lines).
4. Click on the **Bullets** button on the **Formatting** toolbar. The selected text becomes a bullet list.
5. Select the text from **IEEE** to **BSI**.
6. Click on the **Numbers** button. The selected text becomes a numbered list (see Figure 1.7).

1.29. Add borders and shading to a paragraph

Once you have given the paragraphs the alignment you want and the spacing that you prefer, you can apply borders and shading. This formatting helps draw attention to the text. Word has more than 20 different border style you can choose from. Almost all the functions of adding borders and shading can be done from the **Tables and Borders** or **Format → Borders and Shadings...** toolbar. We will familiarize ourselves with the toolbar buttons.

The following steps demonstrate how to apply border and shading to paragraphs.

1. Open and rename **invitation.doc**.
2. Click at the beginning of the first paragraph.
3. Click the **Outside Border** button on the **Tables and Borders** toolbar. The paragraph is surrounded by a border.
4. Click at the beginning of the first paragraph.
5. On the **Tables and Borders** menu, click the **Line Style** drop-down arrow to see more options. Choose the last style in the drop-down box.
6. On the **Tables and Borders** menu, click on the **Line Weight** drop-down arrow to see more options. Choose 3 points.
7. On the **Tables and Borders** menu, click on the **Border Color** drop-down arrow to see the available colors. Choose Red.

8. Click the **Outside Border** button. The paragraph will be surrounded by a border that is red in color as shown in 1.8.
9. Click at the beginning of the first paragraph.
10. On the **Tables and Borders** menu, click on the **Shading Color** drop-down arrow to see the available colors.
11. Click Gray – 10%. The paragraph is filled with gray 10% shading. Your screen will like Figure 1.8.

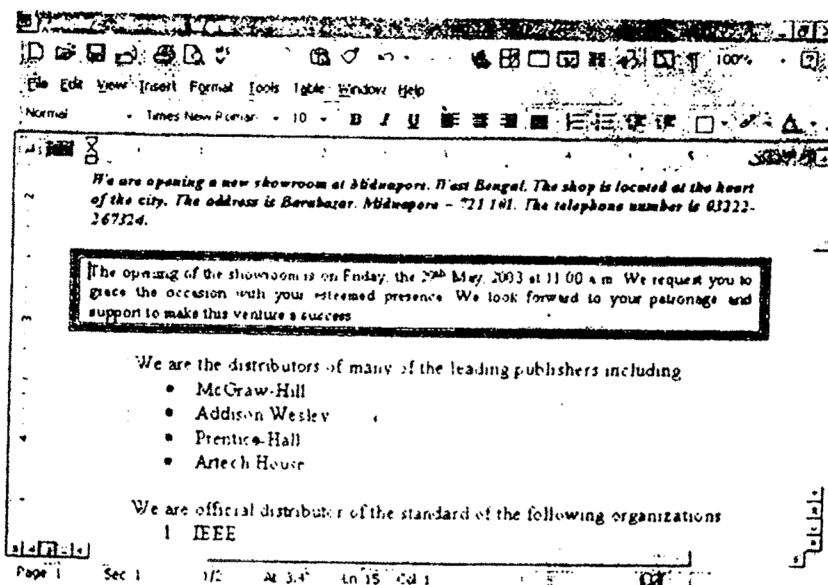


Figure 1.8: Border and Shading

1.30. Adding header and footer

To create header or footer following process is used.

1. On the **View** menu, click **Header and Footer**.
2. The document appears in **Print Layout** view with the header area indicated by a nonprinting dashed-line box. Notice that the regular document text is changed to gray and the header and footer toolbar is displayed.
3. To create a header, enter text or graphics in the header area or, click a button on the **Header and Footer** toolbar.
4. Use the toolbar buttons to add page numbers, date, time and so on.
5. To create a footer, click the **Switch Between Header and Footer** button to move to the footer area.
6. Choose **Close** on the **Header and Footer** toolbar to return to the document.

1.31. Create special effects with text

You can create dazzling special effects and formatting on the text so that it looks attractive and has a fancy look.

Follow the following steps and see what is displayed on your screen.

1. Create a new document and save it under the name **test1.doc**.
2. Type **Vidyasagar University**.
3. Select the text using the mouse pointer.
4. Click **Format → Font**. The Font dialog box appears (shown in Figure 1.9).
5. Choose **Arial Black** as the **Font**, **Bold** as the **Font Style**, **14** as the **Size** and **Red** as the **Font color**.
6. In the effects section, check **Emboss** and **Small caps**.
7. In the **Text effects** tab choose **Sparkle** text.
8. Click **OK**.
9. Click **Save** to save the changes.
10. Click **Close** to close the document.

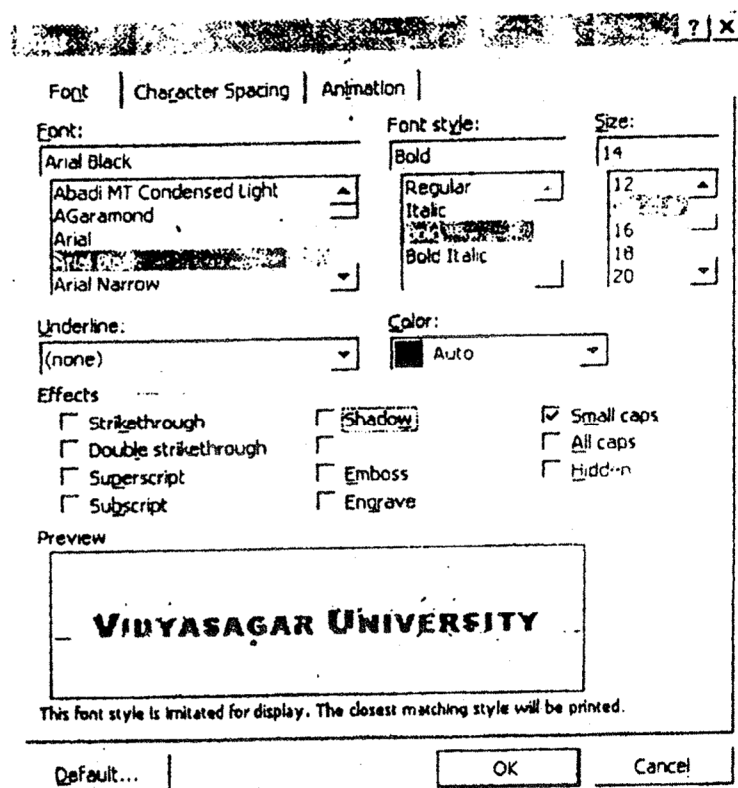


Figure 1.9: Special Effect with Text

1.32. Printing in multiple columns

Creating Columns: Columns can be created by clicking on Columns tool in the **Standard** toolbar or by choosing **Columns** from the **Format** menu. On doing this, a **Columns** dialog box appears, from which parameters such as the number of columns, width of each column and spacing between columns, etc. can be chosen.

Deleting Columns: To delete the columns created and convert text into a single column format, choose **Columns** from the **Format** menu, choose an option from the **Presets** box and click OK.

1.33. Link document

You have created a fancy looking text with some special effects. You want to place that at the top of your letter. In other words, you want to create a letterhead. You can do this by copying Vidyasagar University you have created in the previous section and pasting it at the top of your letter. You can use **Copy and Paste**. But a better option will be **Paste Special**. Using **Paste Special** you can paste the fancy text and simultaneously link the letter to the source document – the document from where you copied the text. Once you have linked the two documents, any changes you make to the text in the original document, will be reflected in your linked document as well.

1. Create a document named **letter.doc**.
2. Open the document named **head.doc**.
3. Copy the text (say your name) from the **head** document.
4. On the taskbar click the button for the **letter.doc** to display it.
5. In the **letter.doc**, place the pointer in the position you want to paste text.
6. **Edit → Paste Special..** **Paste Special** dialog box appears.
7. In the **Paste Special** dialog box, in the **As** list, select **Formatted Text (RTF)** to reflect the type of the text you are pasting from the source document.
8. Click the **Paste Link** radio button to create a link in your document to the source document.
9. Click OK.

The link be test as follows:

1. On the Windows taskbar click on the button for the **head.doc** to open it.
2. Select the name using the mouse pointer.
3. Change the color **Red** to **Blue** and the **Font** from **Arial Black** to **Arial Rounded MT Bold**.
4. The text in the **head.doc** changes.
5. Now click on the taskbar button to open the **letter.doc**.
6. You can see that the changes are reflected in that document also since we pasted it as a link. Whenever the source document is updated the document where the link pasted is also updated automatically.

1.34. Creating a table

A table consisting rows and columns of cells that you can fill with text and graphics. Tables are often used to organize and present information, but they have other uses as well. You can use tables to align numbers in columns and sort and perform calculations on them. You can also use tables to create interesting page layouts and arrange text and graphics.

Suppose you have to make a list of the name of book, author, publisher and price of some books of a book sellers. It can be done very efficiently using tables. Table features of Word are very powerful and extensive.

In the following we will create a table with *four* columns and *five* rows.

1. Create a new document as **book.doc**.
2. Click the **Insert Table** button on the Standard toolbar or click on **Table** on menu and then click on **Insert Table**.
3. Drag to select the number of rows and columns you need. In this case, select 4 columns and 5 rows.
4. Release the mouse button. A table with 4 columns and 5 rows is created as shown in Figure 1.10.

Book	Author	Publisher	Price (in Rs.)
Word 2000	K.P.Sinha	Central	160.00
Operating System	R.K.Rana	Asian Books	250.00
Fortran 77	M.Pal	Asian Books Pvt. Ltd.	185.00
Mathematical Physics	T.Dash	Universities Press	250.00

1.10: A table with 4 columns and 5 Rows

5. Fill the first row with the titles like Name, Author, Publisher and Price.
6. When the table is created like this all the columns have equal width. If you want to change the size of a column move the mouse over borderline. The mouse pointer changes to two horizontal lines pulled by two arrows. Click the mouse and drag in the direction you want to increase or decrease the width of the columns.
7. Now select the first row. To do this, move the mouse pointer by the left side of the table. The mouse pointer changes to a right pointing arrow. When it reaches the first row click. The entire row is selected.
8. Click on the **Bold** button on the toolbar. The text in the first row changes to bold.
9. Now select the last column. To do this, move the mouse from the top towards the top border of the last column. The mouse pointer changes to a back downward pointing arrow. Click the mouse button and the entire last column is selected.
10. Click the **Align Right** button on the toolbar. The contents of the last column are now right aligned.
11. To add a new row at the bottom of the table, position the mouse pointer at the bottom right corner of the table and press the **Tab** button. Another row is added.
12. To insert a new row, place the insertion point in the row above or below which you want it. **Table** → **Insert** and from the menu click on the appropriate button depending on whether you want to insert the row above or below the current row.

Book	Author	Publisher	Price (in Rs.)
Word 2000	K.P.Sinha	Central	160.00
Operating System	R.K.Rana	Asian Books	250.00
Fortran 77	M.Pal	Asian Books Pvt. Ltd.	185.00
Mathematical Physics	T.Dash	Universities Press	250.00

Figure 1.11: A table with 4 columns and 5 Rows

13. To insert a column, place the insertion point in the column right or left of which you want it. **Table** → **Insert** and from the menu click on the appropriate button depending on whether you want to insert the column to the right or left of the current column.
14. To delete a row, place the mouse pointer on the one that is to be deleted. **Tables** → **Delete** and then click on the **Rows**.

15. To delete a column, place the mouse pointer on the one that is to be deleted. **Tables → Delete** and then click on the **Columns**. The entire column is deleted.

Create a table to tabulate the names and the marks of the students in a class. The table should have columns for name, marks for English, Mathematics, Physics and Chemistry. The table should be of the following form.

<i>Tabulation Sheet of Marks of Students of Class XII</i>				
Name	<i>Subjects</i>			
	English	Mathematics	Physics	Chemistry
Raman Rao	50	67	87	45
Aniket Pal	60	78	90	46
Bikash Hota	56	23	56	34
Palash Saha	34	56	76	23

Figure 1.12: Sample Tabulation Sheet

1. Create a table with 5 columns and 7 rows.
2. The first row of the table should be merged. To do this, select the first row by clicking the mouse pointer and click on the **Merge Cells** button on the **Tables and Borders** toolbar (or use menu **Table → Merge Cells**).
3. Type the title **Tabulation Sheet of Marks of Students of Class XII** on the first row and press **Center** button on toolbar.
4. Select the second column to last column of second row by clicking the mouse pointer and then click on the **Merge Cells** button on the **Tables and Borders** toolbar.
5. Type **Subjects** on the second row and click **Center** button on toolbar.
6. Select second and third rows by clicking mouse pointer and click **Merge Cells** button. Then type **Name** on second row.
7. Then type English, Mathematics, Physics and Chemistry on the second to fifth columns.
8. To convert in italic font, select the respective rows and click **Italic** button on the toolbar.
9. If you need to arrange the tabulation according to the name of the students then, select the column containing the names and then click on **Sort Ascending** button of the toolbar. Then the table looks like as in Figure 1.12.

<i>Tabulation Sheet of Marks of Students of Class XII</i>				
Name	<i>Subjects</i>			
	English	Mathematics	Physics	Chemistry
Aniket Pal	60	78	90	46
Bikash Hota	56	23	56	34
Palash Saha	34	56	76	23
Raman Rao	50	67	87	45

Figure 1.13: Sample Tabulation Sheet, after sorting on name

To draw the table of the following form do

1. To insert 2003, insert a column at left. Then type 2003 and click on **Change Text Direction** button on **Tables and Borders** toolbar. To put it on the center of the vertical line click on **Center** button on toolbar.
2. To make it colorful select table and click on **Table AutoFormat** button on **Tables and Borders** toolbar (or use menu as **Table → Table AutoFormat → 3D effects**) and select any option according your choice.

Tabulation Sheet of Marks of Students of Class XII

2003	Name	Subjects			
		English	Mathematics	Physics	Chemistry
	Aniket Pal	60	78	90	46
	Bikash Hota	56	23	56	34
	Palash Saha	34	56	76	23
	Raman Rao	50	67	87	45

Figure 1.14: Sample Tabulation Sheet (with 3D effect)

1.35. Working with graphics

You can enhance the impact of document by incorporating images and other graphics elements into the text. Word 2000 provides thousands of images that will be suitable for any purpose. You can insert a graphics in five ways: from **Clip Art**, **From File**, **AutoShapes**, **Word Art** and **Chart**. To activate **ClipArt** do from menu **Insert → Picture → ClipArt** or from **Draw toolbar** **Insert ClipArt** button.

The steps are describe in the following to insert a graphics to your text.

1. Open a document **test2.doc**.
2. Click on the **Insert ClipArt** button on the **Draw toolbar** or **Insert → Picture → ClipArt..**
3. The **Insert ClipArt** dialog box appears.
4. Browse through the pictures to see if there is anything that interests you.
5. If not, click on the **Keep Looking** button at the bottom.
6. When you have found the clip that you want click on it. A callout appears with options to insert the clip, preview the document and so on.
7. If you want to see an enlarged view of the picture click on the **Preview Clip** button.
8. To insert the clip, click on the **Insert clip** button. The clip is inserted at the position where the insertion point is in the document.
9. After insertion you can change the position and size of the picture. To do this click on the picture and drag it.

1.36. Mail merge

Mail merging is a very useful and powerful feature of Word. You can create the same main document and send them to different people customizing it individually as if the mail was written for him/her. Of course, you have the addresses of all the people you want to send the letter.

Details of mail merge feature is illustrated in the following.

1. Open the document **invitation.doc** and save it as **temp.doc**.
2. Remove the address lines (the first three lines) for the document.
3. Remove the word after **Dear**.
4. Click **Tools → Mail Merge**. The **Mail Merge Helper** dialog box appears.
5. Click the **Create** button under the section **Main document**. In the drop-down list box choose **Form Letters**. The dialog box asking whether you want to make the active window as the main document appears. Click on **Active Window**. That is **Tools → Mail Merge → Create → Form Letters → Active Window**.
6. In the **Data source** section, click on the **Get Data** button. From the drop-down list choose **Create Data Source**. The **Create Data Source** dialog box appears (see Figure 1.14).
7. In the **Create Data Source** dialog box, Word provides a list of commonly used fields. You can change the order in which they are listed, add new fields, or remove existing ones.
8. Once you have customized the field names and their order, click **OK**.
9. You will be asked to save the data source. The **Save As** dialog box appears and prompts you to save the data source as a Word document. The data source is nothing but a Word document consisting of a table with columns that you have selected in the **Create Data Source** dialog box.
10. Once you have saved the blank data source, you will be asked to add records to the data source.



A mail merge data source is composed of rows of data. The first row is called the header row. Each of the columns in the header row begins with a field name.

Word provides commonly used field names in the list below. You can add or remove field names to customize the header row.

Field name:	Field names in header row:	
<input type="text"/>	FirstName LastName JobTitle Company Address1 Address2	Move
<input type="button" value="Remove Field Name"/>		
	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>

Figure 1.14: *Create Data Source Dialog Box*

11. You can press the **Edit Data Source** button to bring up the **Data Form** dialog box (see Figure 1.15).

Title:	Mr.	<input type="button" value="OK"/> <input type="button" value="Add New"/> <input type="button" value="Delete"/> <input type="button" value="Restore"/> <input type="button" value="Find..."/> <input type="button" value="View Source"/>
FirstName:	Aniket	
LastName:	Pal	
JobTitle:	Student	
Company:		
Address1:	Aurobindanagar (South)	
Address2:		
City:	Midnapore	
State:	West Bengal	

Record: 1

Figure 1.15: Data Form Dialog Box

12. Fill the form by entering addresses of all people. Use **Tab** to go to next field. After completion of each address press the **Add New** button. The record will be added to the data source. After completion data entry press **OK** to return to the main document.
13. Now the main document have an additional **Mail Merge** toolbar.
14. Position the insertion point at the beginning of the document. Click on the **Insert Merge Field** button. The field that you have created in the data source appears.
15. Click the field **Title**. The title field appears at the beginning of documents as <<Title>>. The symbol << >> indicates that it is a merger field.
16. Similarly, choose the fields that you want in the document. If you to put some punctuation mark then insert those. For example, after State put a hyphen before the Postal Code. Also put a comma after the First Name in the line **Dear <<First Name>>**.
17. Click on the **Merge ...** button and the **Merge** dialog box appears.
18. You can merge to a new document, printer or e-mail. You can also select the record to be merged.
19. Click on the **Don't print blank lines when data fields are empty** in the **When merging records** section. This will instruct not to leave a blank space if some fields are empty. For example, all the people will not have a Job Title and Company.
20. If you are merging to a new document, the letters will be created as a new document, where each individual letter will start in a new page. You can print this document later. Also you can see the merged letters before you actually print and correct mistakes, if there are any.
21. The merged document will look similar to Figure 1.16.

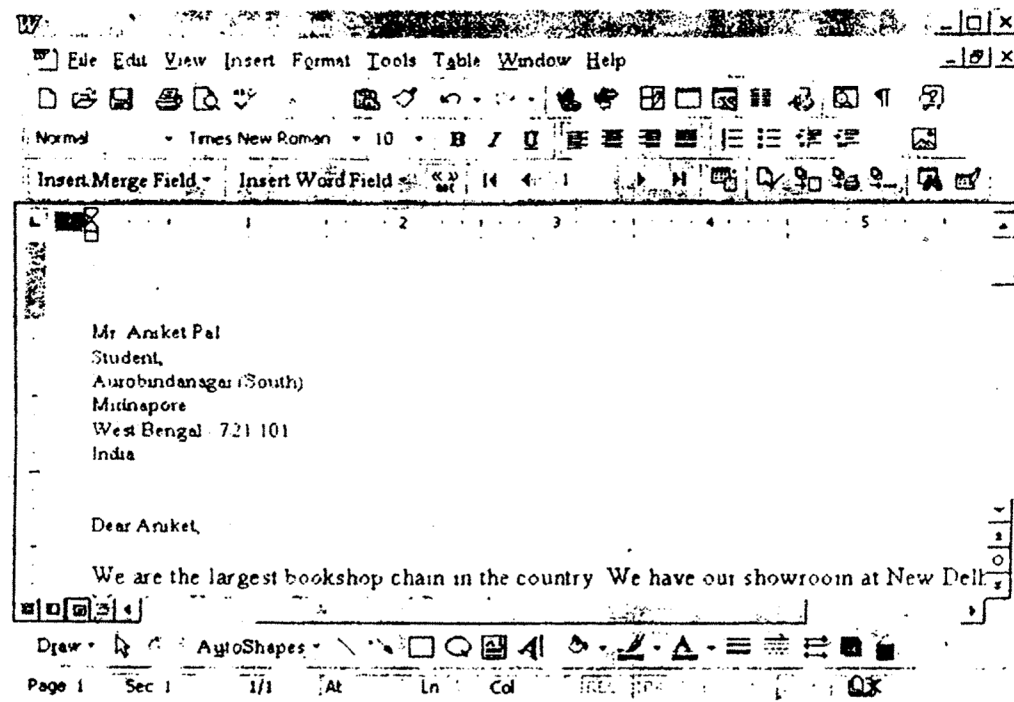


Figure 1.16: Merge document

You can merge envelopes, mailing labels, etc. using the same method. If the data source is available then you can use that data instead of creating a new one. Create a set of envelopes for sending the letters that we have just created.

1.37. Previewing and printing a document

The method by which you print a document is the same in all the Office 2000 applications. Here we see how to preview a document before printing, print whole or part of a document and control the options of the printer.

Print Preview

You may see how a document will look when printed by using the **Print Preview** feature. To preview a document before printing do the following:

1. Click the **Print Preview** button on the Standard toolbar. The **Print Preview** window will appear. The window in the different applications will be slightly different.
2. The **Print Preview** window toolbar contains buttons for printing, magnifying, see or page at a time, zoom (zoom to 10% to 500%, two pages, fit width, etc.) and a **Close** button. 1.17 shows a preview of this document.
3. Click on the zoom and choose the magnification that you want. You can choose 25% 27% etc.
4. To see more than one page at a time click on **Multiple Pages** button.
5. At end of previewed click the **Close** button.

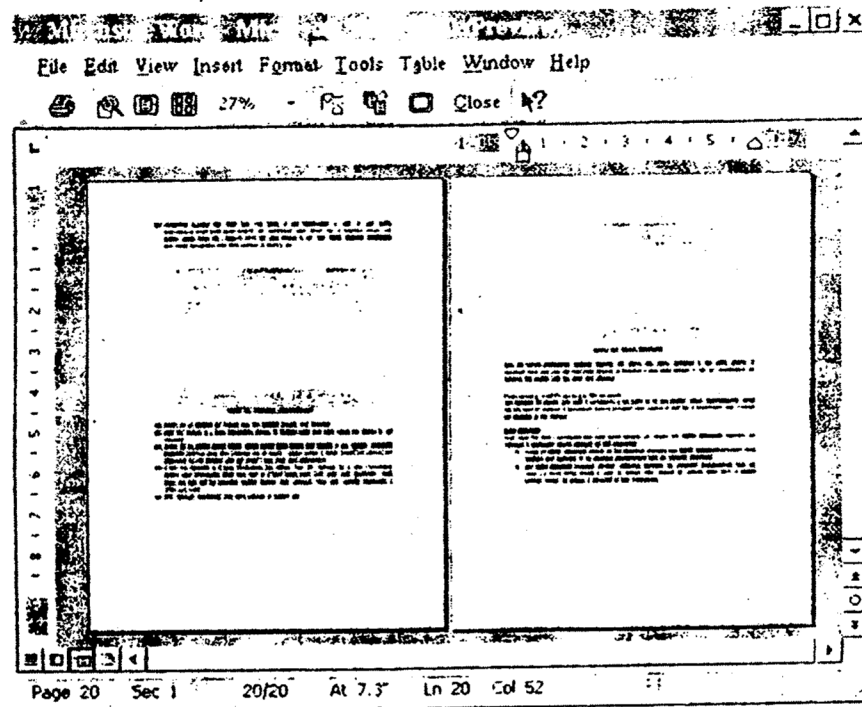


Figure 1.17: *Print Preview*

Printing a document

The print of a document can be taken in the following three ways:

1. Click on **Print** button on the **Print Preview** window toolbar.
2. From the main document, click on the **Print** button on the Standard toolbar.
3. From menu, click **File** → **Print**.

When you print a document using the toolbar buttons it will be printed on the default printer.

If you want to choose a different printer or if you want to exercise more control over the printing (like printing selected pages or a specific range of pages) then you have to print using the **Print** dialog box. The print dialog box is shown in Figure 1.18.

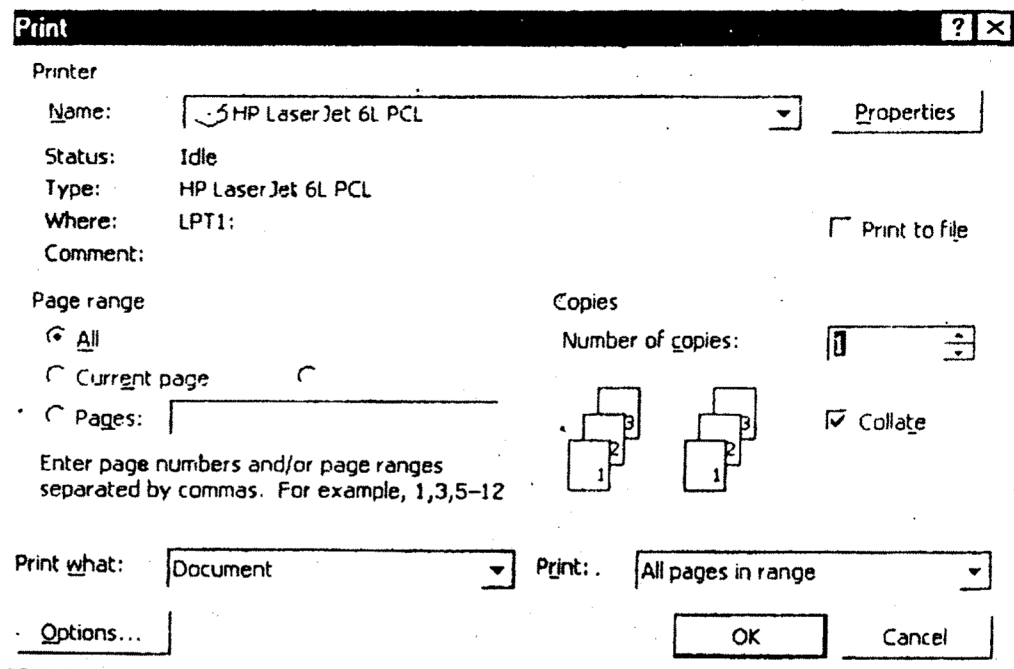


Figure 1.18: *The Print Dialog box*

In the **Print** dialog box you can choose the printer to print the document, the number of copies to be printed, to print the entire document or specific pages, to collate pages in the case of multiple copies and so on.

The **Properties** button in the **Print** dialog box allows you to control the unique features of your printer, such as font resolution, paper type, etc.

The **LaserJet 6L Properties** box has four tabs of information related to the unique features of this particular laser printer. The **Paper** tab lets you identify the kind of paper that is currently installed in the printer and select the orientation or direction you want text to appear on the page. The **Graphics** tab lets you select the font resolution and toner intensity, while the **Fonts** tab lets you choose the type of fonts you want to use. The **Device Options** tab allows you to control unique characteristics of your printer, including print quality and usage of printer's memory.

Once you have selected the required options and set the required parameters and properties then click the **OK** button in the **Print** dialog box to initiate printing. If you want to cancel the print job then click the **Cancel** button.

The contents of the tabs in the **Properties** dialog box will vary depending on the type of the printer.

Section 2: Excel 2000

2.1. Introduction

A spreadsheet is a highly interactive computer program that consists of a collection of rows and columns that are displayed on the screen in a scrollable window. It is a grid of rows and columns and is also called as a worksheet. Spreadsheet programs are developed to automate tasks such as technical calculations, inferential statistics, analyzing data, etc. It also has a powerful program for graphical preparation of numerical data. It is commonly used in production, planning, personnel management, marketing, payroll and accounting.

2.2. Excel 2000

Excel 2000 is a powerful spreadsheet application used for managing, analyzing and presenting data in a graphical manner. Microsoft Excel 2000 is developed on the GUI concept.

Excel performs three different classes of tasks mentioned below:

- It analyses and displays the text and numbers in the cells.
- It manipulates lists of information.
- It creates charts that help to present data in a graphical manner.

Excel 2000 is the most comprehensive spreadsheet application available in the market. It is not just a tool for calculating, manipulating and analyzing data, but also a versatile organizational tool for presenting information. The features of Excel 2000 are listed below.

Worksheet and Graphics: The worksheet and graphics feature includes extremely powerful calculating features. Apart from working with numbers and texts, it is also possible to present graphical data using Excel 2000.

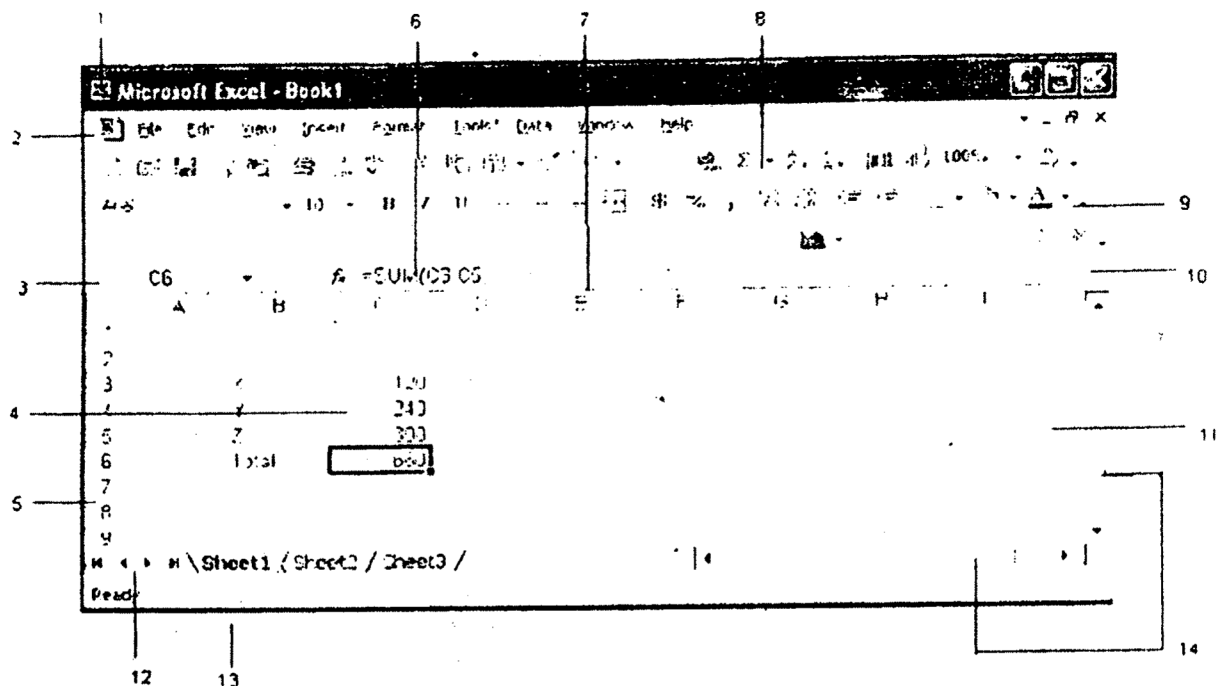
Data Lists and Databases: Database functions is an important feature of Excel. Several useful functions are available for working with data that are listed in a tabular form. Functions are also available for evaluating values, combining data and so on.

Data Exchange with other Applications: Excel takes advantage of the Windows environment. The Windows environment especially applies to the DDE (Dynamic Data Exchange) and OLE (Object Linking and Embedding) concepts within Excel and between Excel and other Windows applications.

Workbooks: Excel works with a consistent file concept. All data is gathered in workbooks. These workbooks store current status of the workspace, along with all currently opened files and the settings selected for them.

2.3. Opening of Excel 2000

To start Excel 2000, click the **Start** button and select **Microsoft Excel** from the **Program** option. On starting Excel, a blank workbook is opened. This workbook has several worksheets and by default, Sheet 1 is selected. Excel provides options in the menu and the user can select the appropriate option to perform an operation. Excel provides a toolbar with buttons, which graphically represent the frequently used commands. A typical Excel workbook is shown in Fig. 2.1.




- | | |
|------------------|------------------------------|
| 1. Title bar | 8. Standard toolbar |
| 2. Menu bar | 9. Formatting toolbar |
| 3. Name box | 10. Formula bar |
| 4. Cell contents | 11. Worksheet |
| 5. Rows | 12. Sheet tab scroll buttons |
| 6. Formula | 13. Sheet tab |
| 7. Columns | 14. Scrollbars |

Figure 2.1. Excel 2000 Window

When Excel 2000 is opened, two windows appear that are nested one within the other. The larger window is called the **Application Window**, which covers the entire screen. The application window is used to communicate with the Excel program. The smaller window is called the **Document Window** and is used to create and edit Excel worksheets and charts.

2.4. Opening a File

To open a new workbook, click **File** menu and choose the **New** option. In the New dialog box that appears, click **OK**. It will open a blank worksheet.

To open an existing workbook, click the  icon in the Standard toolbar or select **Open** from the **File** menu.

2.5. Saving a Workbook

To save a workbook after entering data, click **Save** from the **File** menu. When the file has to be saved for the first time then choose **Save As** option from the **File** menu. Type the file name and Excel will automatically give the ".xls" extension while saving the file.

A non-Excel file can also be opened in Excel.

2.6. Workbook

A workbook is an Excel file where the data is stored. A workbook consists of many worksheets. A worksheet is a page in the workbook where data can be entered. The current sheet is always highlighted in the sheet tab. Sheets belonging to a particular application can be stored in the same workbook. When the workbook is opened, all the worksheets contained in that workbook are automatically opened. Since each workbook contains many sheets, we can organize various types of related information in a single file.

By default a workbook contains three worksheets. To move from one sheet to another sheet, click the sheet tabs.

A worksheet consists of rows and columns. Rows run horizontally in a sheet and are identified by numbers. Columns run vertically in a sheet and are identified by letters. The intersection of a row and a column is called a **cell**. Cells are named by their position in the columns and rows. The column letter followed by the row number is called a **cell reference**.

2.7. Executing Commands

Excel commands can be given in one of the following ways.

- Choosing an option from the Menu bar.
- Choosing an option from the Shortcut menu.
- Selecting a tool from the Toolbar.
- Using Shortcut key combinations.

Menu Bars

Menus are the primary means of performing tasks such as opening a file, copying a group of cells, printing a worksheet or creating a chart. Menus can be invoked by using either the keyboard or the mouse. To invoke a command from the menu bar use the specific combination of keys.

Shortcut Menus

A shortcut menu is invoked by pressing the right mouse button. The shortcut menu gives direct access to the most commonly used commands. For example, clicking the right mouse button on the active cell displays a short-cut menu of editing and formatting options.

Clicking the right mouse button on the toolbar will display the toolbar shortcut menu as shown in Fig. 2.2

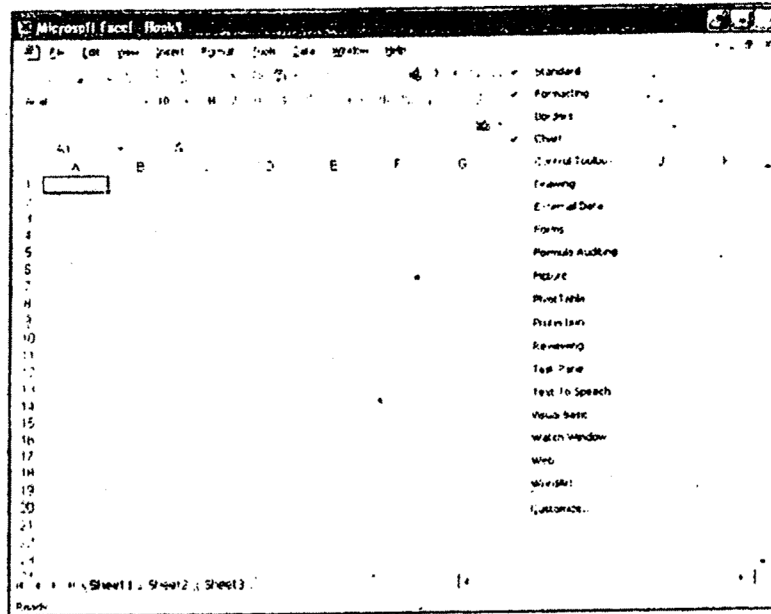


Figure 2.2.

Shortcut menus are also available for manipulating toolbars, rows, columns and sheets as well as for editing and formatting charts and databases.

2.8. Wizards

It provides guidelines and a series of step-by-step procedures to complete a process. Following are the different wizards available in Excel 2000.

- **Text Import Wizard**- it imports text files.
- **Chart Wizard**-It creates and edits charts.
- **Convert Text to Columns Wizard** – It separates contents in a cell into two different cells.
- **Function Wizard**-It inserts a function.
- **Tip Wizard**-It suggest easier ways of doing tasks.
- **PivotTable Wizard** - Builds an interactive table from data existing on sheets.

2.9. Creating and Using Templates

A template is like a pad of preprinted paper. Every time a template is opened a copy of the template is created. Templates can be extremely helpful when working with workbooks with identical formatting, labels, formulas and so on.

To save a workbook as template follow the steps given below.

- Setup the workbook to the desired format.
- Choose **Save As** from the **File** menu. The **Save As** dialog box appears.
- Click **Template** in the **Save As Type** box. Excel automatically opens the template's location.
- Specify a filename to the template and click **Save**. Excel gives a **.xlt** extension to the file.

To open a copy of the template, choose **New** from the **File** menu. In the **New** dialog box double click the icon to open the template.

2.10. Working with Worksheets

To add a new worksheet click **Insert** → **Worksheet** from the menu bar. To add multiple worksheets, click the **number of worksheet** tab to add in the open workbook by holding the SHIFT key. Then click **Worksheet** on the **Insert** menu.

Text, numbers and dates can be entered into any cell of a worksheet. Data can be entered in the active cell by clicking the enter box in the formula bar or by pressing F2. An entry can be cancelled by clicking the Cancel button **X** on the formula bar or by pressing the ESC key. Each cell can hold up to 255 characters. Text is left aligned and numbers are right aligned in Excel and this default alignment is called General alignment.

The column headings may not appear properly, if the column width is not large enough to contain them. To increase the column width, place the mouse on the right border of the column whose width has to be increased and click drag the mouse button.

2.11. Entering Date and Time

To enter today's date, move to a cell and type = today().

In Excel the time is added to the date serial number as a decimal fraction of a 24 hour day. Therefore, midnight is 0.000000, noon is 0.500000 and 11:59:59 PM is 0.999988. Whenever the date or time is to be changed, the cell format is automatically changed from the normal format to the appropriate date or time format.

2.12. Entering Data in a Series

Whenever the user wants to fill a cell range with data forms as a series (e.g. 1,2,3,4 or Jan, Feb or Mon, Tue) the data input can be automated. This can be achieved by using the fill handle. The fill handle is a black square located on the lower right corner of the selected cell. This is called **Autofill** feature.

For example, to generate month names in a range of cells, enter Jan into cell A1 and point to the fill handle with the mouse, the mouse pointer changes to a black cross. Click and drag through the cells A1 to A12 and then release the mouse button. The series of months from January to December will be filled into cells A1 through A12.

2.13. Manipulating Cell Contents

The cell contents can also be rearranged. Rearranging involves copying, moving, clearing cells or inserting and deleting rows. These are discussed below:

Copy Data: While copying or moving data, a copy of that data is placed in the clipboard. The clipboard provided by Office 2000 can store multiple bits or data (up to 12). To copy a range of data, select the range and press the key CTRL and the key-C or click the Copy button in the Standard toolbar. The icons are the same as in Word 2000. Select the destination cell and click the Paste button from the toolbar menu. The Copy and Paste commands can be easily accessed from the shortcut menu.

Move Data: Moving data is similar to copying, except that the data is removed from its original place. To move data choose the Cut button from the Standard toolbar.

Drag and Drop: The fastest way to copy is to drag and drop the data. To do this, select the cells to be copied, hold down the CTRL key and drag the border of the selected range. On releasing the mouse button the data is copied to the new location. The data gets moved to the new location if the border is dragged without holding down the CTRL key.

To copy data to a different sheet, press the keys **CTRL** and **ALT** while dragging the selection to the sheet's tab. Excel switches to that sheet, where the selection can be dropped in the appropriate location.

Delete Data: To delete data in a cell or range of cells, select the cell or the range and press the **DEL** key. The **Edit** → **Clear** command can be used to delete only the formatting of the cell(s). The **Clear** command can be used to clear the format, contents, comments or all of them.

The **Edit** → **Delete** command removes the cells and then shifts the surrounding cells to take over their place

Example 1. Prepare a worksheet showing employee code, employee name and designation of the software engineers working in a company ATLANTA. The employee code starts with 1000 increments by one for each engineer and ends with 1004. Insert today's date on top of the worksheet.

- From the **Start** menu, click on the program option and select **Microsoft Excel**. A worksheet will open.
- Place the insertion point in a cell in the top of the worksheet. Type **"=today()"** in the cell and press **ENTER**.
- Type **Employee code**, **Employee name**, **Designation** in the cells **B5**, **C5**, **D5** respectively.
- In the first cell of the employee code column, type 2000. In order to fill all the cells of the employee code column, select **Edit** → **Fill** → **Series**.
- Enter the step value as 1 and stop value as 2004, select **columns** option. Press **OK**. This will fill the column. Enter the employee names for each employee code.
- Type *Software Engineer* in the first cell of the *Designation* column. To have the same designation in all the following cells, select the cell, click and drag through the cells using the fill handle. This will have a copy of the first cell.

The resultant worksheet is shown in Fig. 2.3.

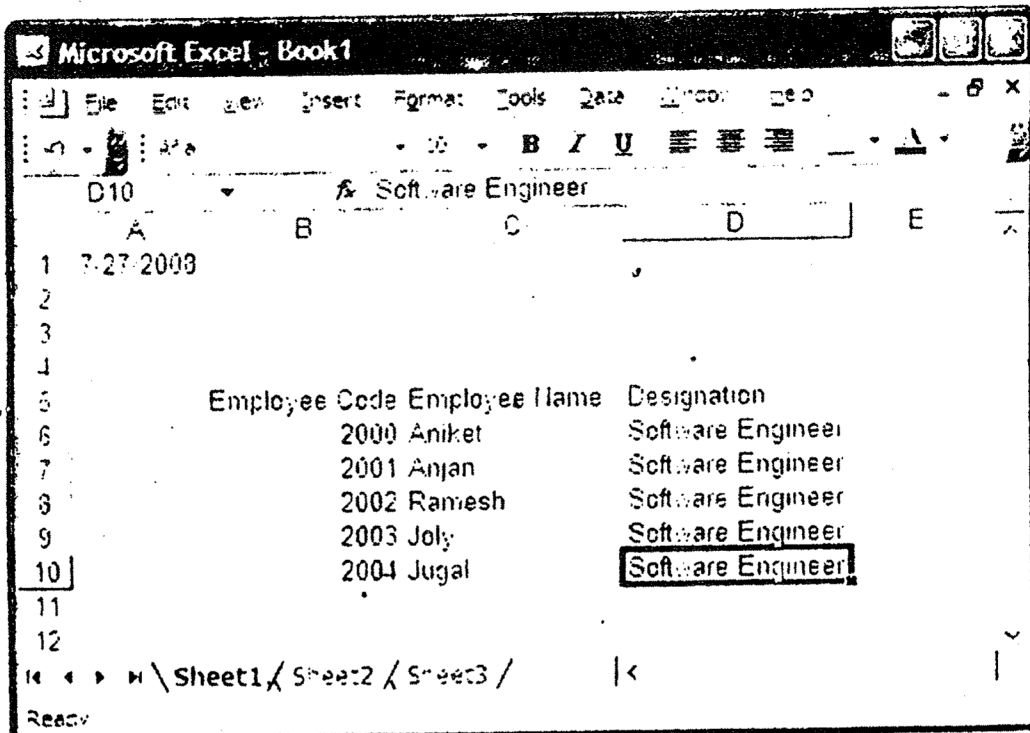


Figure 2.3: Manipulating Cell contents

2.14. Formatting Rows and Columns

Rows and Columns Insertion

To insert a row or column in the worksheet follow the following steps:

- Select the cell below which the row has to be inserted or the cell to the right of which the column has to be inserted.
- To insert multiple rows or columns, select the number of rows or columns equal to the number to be inserted by clicking and dragging the mouse over the worksheet.
- Select **Rows or Columns** from the **Insert** menu. Excel inserts row below the selection or columns to the left of the selection.

A row or column can also be inserted by right clicking the row number or column header and by clicking the **Insert...** from the pop-up menu.

Alternately, to insert a row or column, right-click a cell and from the pop-up menu, choose **Insert** option. An **Insert** dialog box will be displayed. Select the **Entire row or Entire column** option to insert a row or column in the worksheet.

Deleting Rows and Columns

When rows are deleted in the worksheet, the rows below the deleted row move up to fill the space. When columns are deleted, the columns to the right are shifted to the left.

To delete a row or column, click the row number or column letter on the row or column to be deleted and then select the **Delete** option from the **Edit** menu. To delete more than one row or column drag over the row numbers or column letters for selecting them. Alternately, right-click

the selection and choose **Delete** from the pop-up menu. The rows and columns are renumbered automatically.

Merging Cells

Excel allows merging data in one cell with adjacent cells (that are blank) to form a big cell. Merging cells is useful especially while creating a decorative title for the worksheet.

To create a title with merged cells, follow the steps given below.

- Enter the title in the upper-left cell of the range. To enter multi-line title, press ALT+ENTER to insert each new line.
- Select the range in which the title has to be placed.
- Click **Cells** from the **Format** menu. The **Format Cells** dialog box appears.
- Click the **Alignment** tab.
- Click the **Merge Cells** check box and click **OK**.

Inserting and Deleting Cells

Inserting cells will cause the data in the existing cells to shift down a row or over a column for creating space for the new cells. Follow the steps given below to insert a single cell or group of cells.

- Select the area where the new cell(s) are to be inserted. Excel will insert the same number of cells as selected.
- Choose **Cells** from **Insert** menu. The **Insert** dialog box appears.
- Select **Shift Cells Right** or **Shift Cells Down** and click **OK**.

To delete the cells completely, follow the steps given below.

- Select the cell or range of cells to be deleted.
- Choose **Delete** from the **Edit** menu. The **Delete** dialog box appears.
- Select **Shift Cells Left** or **Shift Cells Up** and click **OK**.

Note. When cells are deleted, they are removed from the worksheet and the surrounding cells are shifted to fill in the space. But when a cell is cleared, the cell remains in the worksheet while the contents, format or notes of the cells are cleared. To clear a cell, select the range of cells to be cleared and choose **Clear** from the **Edit** menu.

2.15. Password-Protecting a Workbook

If the workbook contains confidential information, a password can be applied to it. This will only allow the users with password to open and view the workbook. To protect a workbook, follow the steps given below.

- Open the workbook and select **Save As** from the **File** menu.
- In the **Save As** dialog box, click the **Tools** button to display the drop-down menu and choose **General Options**. The **Save Options** dialog box shown in Fig. 2.4 appears.
- Enter the password in the **Password to Open** box.
- To have a separate password for editing the file, enter the password in the **Password to Modify** box.
- Click **OK**.
- In the **Confirm Password** dialog box, re-enter the password for verification.
- Click **OK** to return to the **Save As** dialog box.

- Specify a filename and path for the file and click **Save**.
- If prompted, click **Yes** to replace the old version of the file with the new password-protected version.

If the **Read-only recommended** option is selected in the **Save Options** dialog box, the user is prompted whether he or she wants to open the file as read-only. On selecting **Yes**, the changes made to the file must be saved under a different filename.

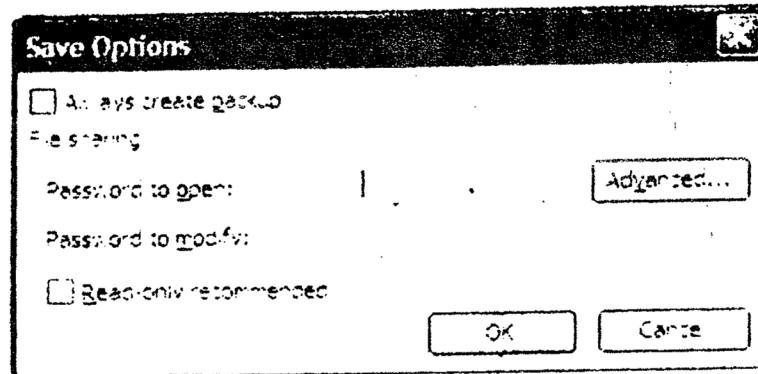


Figure 2.4: Password-Protecting a workbook

2.16. Ranges

A range is a rectangular group of cells. The smallest range is a single cell and the largest range includes all the cells in the worksheet. A range can include cells from same sheet or cells from adjacent sheets. Ranges are defined by the addresses of two opposite or diagonally paired corner cells separated by a colon or two dots.

Ranges can be set when only a part of the worksheet has to be printed. To print the selected range, follow the steps given below.

- Select the range of cells to be printed.
- In the **File** menu, point to the **Print Area** option and click **Set Print Area**.
The print area set is indicated by a dotted line around it in the worksheet.
- To check click the **Print Preview** button.

To print non-adjacent columns or rows in a table, hide the columns or rows not needed to print. To do this select the rows or columns and right-click the selection and select **Hide**.

2.17. Quit from Excel

The user can quit the Excel application by selecting **Exit** from the **File** menu. It is always recommended to save and close the file before quitting Excel. If a workbook has not been saved, Excel displays a confirmation box to confirm whether to save or quit without saving changes.

2.18. Formula

Excel formulae can be used to perform simple calculations on the data like addition, subtraction, multiplication and division. One of the significant features of a spreadsheet program is its ability to manipulate text and perform simple and complex calculations efficiently. Formulae usually consist of one or more cell addresses or values and a mathematical operator such as +, -, * and /.

Formulae are of the following three types.

Text Formulae: Uses text and may contain the text operator ampersand (&) which concatenates two numbers. For example, if we type the expression =123 & 456 in a cell, then the result 123456 will display in that cell.

Numeric Formulae: Contains arithmetic operators like +, -, *, /, A and %. For example, (A1+A2+A3+A4) or A1:A4.

Logical Formulae: Contains comparison operators like >, <, <=, >= and <>.

A formula can be up to 255 characters long. A formula must always begin with an '=', '+', or a '-' sign. Formulas do not accept spaces, except between sets of letters, numbers or symbols enclosed in quotation marks.

Order of Evaluation of Operators

Excel evaluates a formula in a particular order determined by the precedence number of the operators being used and the parentheses placed in the formula. This is listed in the following table.

Operator	Description	Precedence Number
:	Range of cells	1
Space	Intersection of cells	2
,	Union of cells	3
-	Negation	4
%	Percentage	5
A	Exponentiation	6
*	Multiplication	7
/	Division	7
+	Addition	8
-	Subtraction	8
&	Concatenation	9
=	Equal to	10
<	Lesser than	10
>	Greater than	10
<=	Lesser than or equal to	10
>=	Greater than or equal to	10
<>	Not equal to	10

2.19. Entering a Formula

A Formula can be entered in two ways, i.e. by typing the formula or by selecting cell references. To enter a formula, the following steps are performed.

- Select the cell in which the formula calculation has to appear.
- Type the equal sign "=" or click the Edit Formula button in the Formula bar. Type the formula and press ENTER.

Example 2. Open a new sheet and type the numbers 10, 45, 15 and 90 starting from cell A1 to A4. Now select the cell A5 and type =A1 +A2 +A3 +A4 and press ENTER. The sum 150 is displayed in the cell A5. To edit the formula select the cell A5 and press F2.

To enter a formula by selecting cell references, follow the steps given below.

- Select the cell in which the formula's result has to be displayed.
- Type the equal sign "=" or click the **Edit Formula** button in the **Formula** bar. Click the cell whose address has to appear first in the formula. We can also click a cell in a different worksheet or workbook. The cell address appears in the formula bar
- Type the mathematical operator after the value to indicate the next operation to be performed. The operator appears in the **Formula** bar.
- Continue to click the cells and type the operators till the formula is complete and press ENTER.

An error appears in a cell, if anyone of the following operations is performed. Division by zero, use of a blank cell as a divisor, referencing to a blank cell, deleting a cell used in a formula or including a reference to the cell in which the result appears.

Referencing Methods

A formula can be moved from one worksheet location to another. When a formula is moved, the cell addresses are automatically changed relative to the location to which they are moved. This is known as **Relative referencing**. By default, Excel does not treat cells included in the formula as a set location but considers them as a relative location. This type of referencing saves time for the user, since the same formula need not be created repeatedly. For example, **AutoSum** formulas are written with relative referencing.

In certain situations, it might be essential to refer to the same specific cell on the worksheet in every copy of the formula. The **Absolute** referencing method is used in such cases. Absolute references are denoted by dollar signs before the column and row addresses, for example \$A\$2.

In case some part of the address needs to be fixed, **Mixed** referencing is used. **Mixed** references contain both absolute and relative cell addresses, like \$A2 or C\$4. To quickly cycle through the reference types, click the formula bar and press F4. This will change the reference to \$A\$1, A\$1, \$A1, and A1 with each press.

AutoSum

SUM is one of the most commonly used functions and Excel provides a fast way to enter it. The **AutoSum** button on the Standard toolbar Σ automatically sums the rows or columns. The **AutoSum** tool automatically builds a SUM formula in the active cell based on a contiguous range of numbers, either above or to the left of the active cell.

AutoSum can be used in the following three ways.

- To locate and total rows or columns in the range nearest to the current cell.
- To total any selected range.
- To add grand totals to a range containing other totals.

To sum rows or columns for the nearest range click the **AutoSum** button and press ENTER or double click the **AutoSum** button. To find the sum of a specific range, select the range and click the **AutoSum** button. To add grand total to a range of cell values select the entire range, including sub-totals and click the **AutoSum** button.

2.20. Naming Cells

We can use different methods to name cells. The Name box, the Create Names dialog box and the

Define Name dialog box.

Using the Name Box

The **Name box** is the fastest way of naming a cell or range of cells if we are creating a single name. The **Name box** can also be used if the name is not already a label or table heading on the worksheet.

Follow the steps given below to name a cells or range of cells using a Name box.

- Select the cell or range of cells to be named.
- Click the Name box (Fig. 2.5).
- Type the name and press ENTER.

Hereafter, whenever the range is selected, the name appears in the Name box.

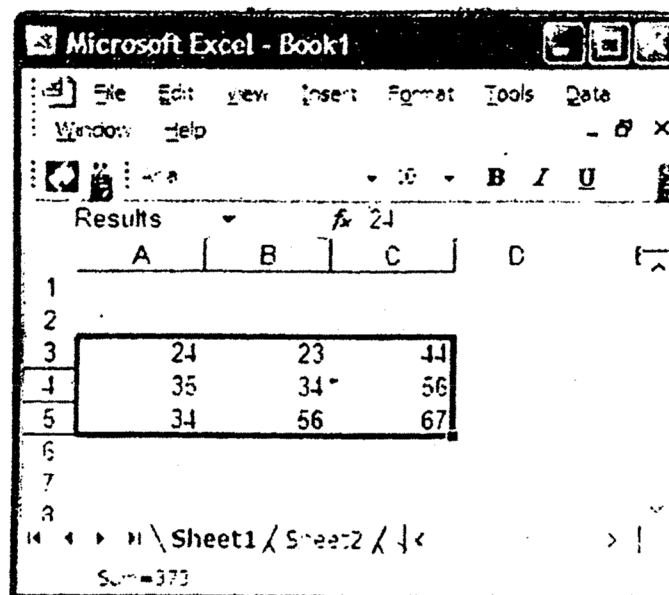


Figure 2.5. Naming cells

Writing Formulas with Named Cells

To use cell and range names in a formula follow the steps given below.

- Click the cell in which the result of the formula has to be displayed.
- Type the formula by pressing an "=" in the beginning.
- To insert the formula in a particular cell, type the name or select **Insert** → **Name** → **Paste** from the menu bar and click the name in the list.
- Complete the formula by pressing ENTER.

Example 4.

Payroll sheet of the employees is given in the following table:

Name	Pay rate	Hours
Pal	35	13

Karmakar	20	15
Nandi	22	20
Verma	40	23
Mondal	23	25

The gross pay for each employee, and the total number of hours worked are calculated using the following steps.

- Select the cell to the right of hours.
- Type the equal sign "=".
- Click the Payrate cell whose address has to appear first in the formula. The cell address appears in the formula bar.
- Type the "*" sign after the value. Click the hours cell whose address has to appear next in the formula. Press ENTER.
- Drag the fill handle to the rest of the cells in the column. The gross pay for other employees are displayed in the corresponding rows.
- Name the hours column by selecting the entire column. The name box is clicked and the column is named as "hours".
- Place the insertion point in the cell where the total number of days worked should be displayed. Click AutoSum button and select the entire hours column, press ENTER and the sum is displayed in the cell. The resultant sheet will be as shown in Fig. 2.6.

Pay Roll for the month of January				
Name	Payrate	Hours	Gross Pay	
Pal	35	13	455	
Karmakar	20	15	300	
Nandi	22	20	440	
Verma	40	23	920	
Mondal	23	25	575	
Total hours		96	Total Gross=	2690

Figure 2.6.

2.21. Functions

Functions are special pre-written formulae that take values and perform operations and then return a value to the cell in which they are entered. Functions simplify and shorten formulae in

the worksheet. For example, instead of using the formula =A1+A2+A3+A4+A5 we can use the function =sum(A1:A5). In certain instances, functions are the only way in which certain tasks can be carried out.

Specifying Arguments

In order to perform certain tasks, functions require specific information called **Arguments**. Arguments are values that are passed to the functions to perform operations. The number of arguments in a function varies between 0 and 14 and the length is restricted to 255 characters including quotation marks, if any. Arguments can be Constants, Cells or ranges, Range names or Functions.

Function Wizard

The **Function Wizard** can be used to select the Function and assemble the arguments correctly.

Although functions can be typed directly into a cell, the Function Wizard can be used to simplify the process. The Function Wizard helps in the process of inserting a function. The steps in the **Function Wizard** are the following.

- Select the cell in which the function has to be inserted.
- Type "=" or click the **Edit Formula** button on the **Formula** bar. The **Formula Palette** appears.
- Select the function that has to be inserted from the **Functions** list by clicking the arrow on the Functions drop down list. If a particular function is not listed in the **Functions** list, click the **More Functions** option at the bottom of the list.
- Enter the argument for the formula. To select a range of cells as an argument, click the **Collapse Dialog** button.
- After selecting the range, click the **Collapse Dialog** button again to return to the **Formula Palette**.
- Click **OK** to return to the worksheet.

Common Functions

Some of the commonly used functions and their purposes are shown in the following table.

<i>Functions</i>	<i>Purpose</i>
SUM	Add the values in the selected range.
MIN	Find the minimum value in the selected range.
MAX	Find the maximum value in the selected range.
AVERAGE	Average the values in a selected range.
COUNTIF	Count all values that meet specific criteria.
SUMIF	Add together all values that meet specific criteria.
VLOOKUP and HLOOKUP	Find a value in a table.
IF	Display a value that depends on a set criteria.
PMT	Calculate the payment for specific loan terms.
NOW	Return current date and time.
TODAY	Return the current date.
CONCATENATE	Join cell values together in a single cell.
LEFT and RIGHT	Return a specific number of characters from the left (or right) end of a cell's value.

2.22. Formatting

Formatting makes a worksheet look pleasant and makes the data more meaningful by visually

segregating data into groups. We can format cells, columns, number, text, border and alignment. Formatting can be done in the following three ways:

- Selecting tools from the Formatting toolbar
- Using the Format menu
- Using Automatic

Using the Formatting Toolbar

In order to display the Formatting toolbar, right-click any toolbar or menu bar and then select **Formatting** option from the shortcut menu. The most frequently used formats are available on formatting toolbar. Using the formatting toolbar makes the formatting work easier and quicker.

To find the hidden buttons in the toolbar, click the small arrow at the right end of the toolbar and click the **Add or Remove Buttons**. Click the buttons that have to be displayed in the toolbar.

Formatting Cells: After entering data, Excel's formatting features help us to make the worksheet and its contents fit together.

Aligning Entries: When data is entered, the text is automatically aligned to the left side of the cell and numbers to the right side. The **Align Left**, **Center** and **Align Right** buttons in the **Formatting** toolbar can be used to change the default alignment.

Rotate Cell Entries: Follow the steps given below to rotate the text entered in the cells.

- Select the cells that have to be rotated.
- Click **Cells** in the **Format** menu.
- On the **Alignment** tab, under **Orientation**, click in the half-circle to set a rotation angle as shown in Fig. 2.7.
- Click **OK**.

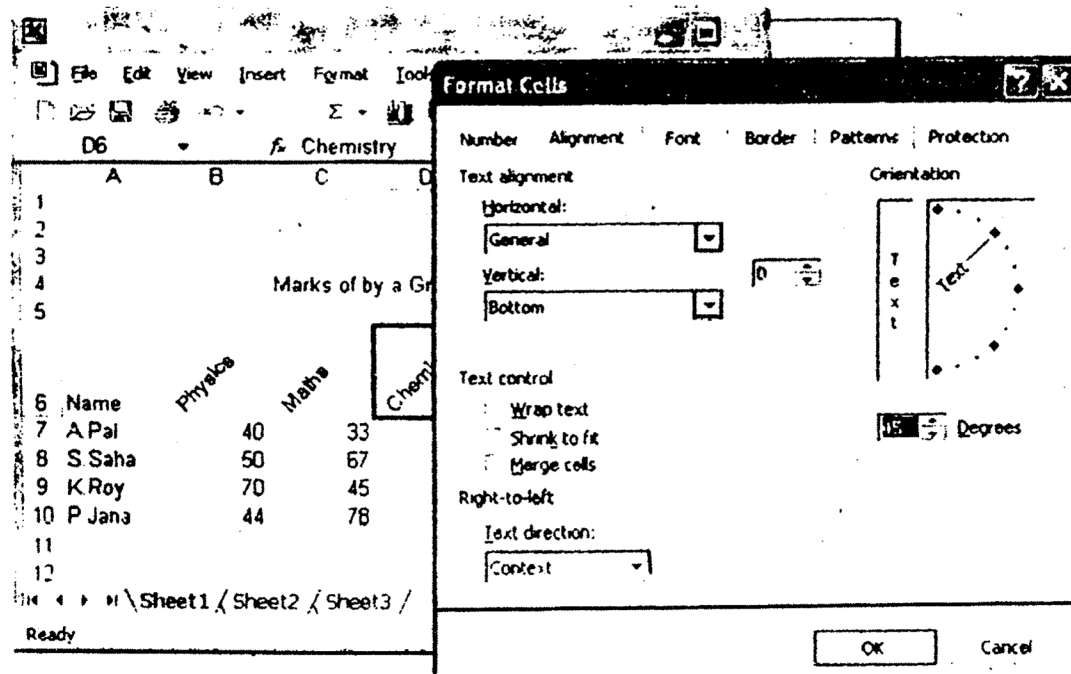


Figure 2.7.

Changing Column Width and Row Height: In most of the worksheets the column has to be

resized and row height adjusted. There are several ways to resize rows and columns.

Resize a Column Visually: Position the mouse pointer over the right border of the column selector for the column that has to be resized. When the mouse pointer becomes a two-headed arrow, drag the border to resize the column. Follow the same procedure to resize the rows as well, but in this case drag the bottom border of the row selector for the row that has to be resized.

Best-Fitting: When the worksheet is too large, we can save the trouble of guessing the appropriate width by "best-fitting" the column width. To best-fit columns, follow the steps given below.

- Position the mouse pointer over the right border of the column selector for the column that has to be best-fit.
- When the mouse pointer becomes a two-headed arrow, double-click the border to best-fit the column.

To best-fit a row, follow the same steps as above, but in this case double-click the bottom border of the row selector for the row to be resized.

Matching Precise Width or Height: We can match the measurements of rows and columns that are far apart in a worksheet or in a different worksheet. To achieve this, set the precise measurements for the columns and rows to be matched and they will match exactly. To set a precise column width or row height, follow the steps given below.

- Right-click a column or row selector.
- Click **Column Width** or **Row Height**.
- Enter a new measurement and click **OK**.

Hide or Unhide Rows and Columns

We can hide important but cluttered cells by hiding those particular columns and/or rows. The worksheet will still continue to function properly. In order to hide rows or columns, right-click the selector (row number or column letter) for the row or column that has to be hidden and click **Hide**. To hide multiple rows or columns, select the rows or columns and then select **Hide**.

To unhide rows or columns, select the rows or columns on both sides of the hidden column(s). Right-click the selection and select **Unhide** from the shortcut menu.

Shrinking Entries to Fit a Cell: In case there is a cell entry that does not fit and we also do not want to widen the column for that particular cell, we can shrink the entry so that it fits within the column width. To shrink an entry to fit the column width, follow the given steps.

- Select the cells whose entries have to be shrunk.
- Click **Cells** from the **Format** menu.
- Click **Shrink to fit** check box on the **Alignment** tab and click **OK**.

Wrap Text to Multiple Lines: A different method to fit long entries in a narrow column is to wrap the text into multiple lines. Wrapping increases the row height. To wrap an entry to multiple lines follow the given steps.

- Select the cells whose entries have to be wrapped.
- Click **Cells** from the **Format** menu.
- Click **Wrap text** check box on the **Alignment** tab and click **OK**.

Applying Number Formats: To format a number value, follow the steps given below.

- Select the cell or range of cells to format.
- Click **Cells** in the **Format** menu.

- Select a format category on the **Number** tab.
- Select and set the options available for the format category and click **OK**.

To format dates and times, select the **Date** or **Time** from the Category list box.

Formatting Borders: Border segregates information so that it is quickly understood. Since Excel does not print the worksheet gridlines, the borders have to be specified explicitly. To apply borders quickly, we use the Borders palette. To display the Borders palette, click the **Borders** button —

The **Format Cells** dialog box can also be used to apply borders. Select the **Borders** tab to set the appropriate border.

AutoFormat: To apply a format automatically, select a range and choose the **AutoFormat** command from the **Format** menu. Select one of the different table formats, including formats for financial and accounting data.

Adding Graphic Images to a Worksheet: It is possible to add a logo or graphic image to a worksheet. To paste a graphic image into a worksheet follow the steps given below.

- Make enough room for the image in the worksheet.
- Click the cell where the picture has to be pasted.
- On the **Insert** menu, point to **Picture** and then click **From File**.
- In the **Insert Picture** dialog box, double-click the image file. The picture will be inserted at the selected cell.

2.23. Printing a Worksheet

Once the data is entered, the calculations and formatting is completed the next step is to take a printout of the worksheet. Excel provides several options for customizing the printing job. To print a worksheet, click **Print** from the **File** menu. The **Print** dialog box will appear on the screen. The **Print** dialog box will give an overview of the available options.

To select various options such as Print quality, Paper orientation, Media and Media sizes click the **Properties** button. The **Preview** button gives a view of the document that is to be printed. To make adjustments (margins, page size, etc.) before printing the document, select the **Page Setup** option from the **File** menu. This is used to specify major facets of the page like page layout, margins, etc.

Setting and Removing a Print Area: If only a part of the worksheet has to be printed repeatedly, the print area can be set which allows printing the range quickly without selecting it first. To set a print area in the worksheet, follow the steps given below.

- Select the range of cells that has to be included in the print area
- On the **File** menu, point to **Print Area** and click **Set Print Area**. The print area will be set and a dashed line is displayed around it in the worksheet.

To remove a print area, point to **Print Area** in the **File** menu and select **Clear Print Area**.

Printing a Selected Range: To print a range of cell in a worksheet only once without setting print area follow the steps given below:

- Select the range of cells to be printed.
- Click **Print** from the **File** menu.
- In the **Print** dialog box, click **Selection** under **Print what** and click **OK**.

Programming in C

3.1. Introduction

C is a high level programming language was developed by Dennis Ritchie and was implemented at Bell Laboratories in 1972. This language running under different operating systems. At present it is the most useful programming language and it is used as general purpose programming language.

3.2. The C character set

A computer key board contains many symbols (characters), some of them are used to write a program in C. The valid C characters are listed below.

The C character set consists of alphabets A - Z (upper case), a - z (lower case), digits 0 - 9 and some special symbols listed below:

!	*	+	\	"	<
#	(=		{	>
%)	~	;	}	/
^	-	[:	,	?
&	_]	'	.	blank

C uses certain combination of these characters such as \b, \n and \t, to represent special conditions such as backspace, newline and horizontal tab, respectively. These characters combination are known as escape sequence.

3.3. Constant data

Any entry which remains unchanged during the execution of a C program is defined as a constant. It may be of *integer constants*, *floating point constants*, *character constants* and *string constant*.

Integer constant

An integer constant is a whole number that is written as a string of decimal digits without a decimal point or an exponent symbol. It can be either positive or negative. It is also called a fixed point constant.

Some valid integer constants:

25 0 -7 +12367

Long integer constant

Long integer constants may exceed the magnitude of ordinary integer constants, but require more memory within the computer.

Floating point constant

A signed or an unsigned whole number containing a decimal point or an exponent or both, is called a real or floating point constant.

Some valid real constants :

2.613 0.12 -0.563 0.000017 .123 1.

Character constant

A *character constant* is a single character, enclosed in apostrophes (single quotation marks).

Some character constants are 'x', 'a', 'B', '6', '+'

String constant

A *string constant* consists of any number of consecutive characters enclosed in double quotation marks.

Following are the valid string constants:

"raman" "Botany" "Rs. 2345" "Vidyasagar University"

3.4. Variables and arrays

A quantity that varies during program execution is called a variable. Each variable has a specific storage location in memory where its numerical value is stored. The following rules must be satisfied by every C variable.

- (i) The first character must be an alphabet.
- (ii) The variable name can have at most 32 characters.
- (iii) The variable name should not contain any special character other than underscore.
- (iv) Upper case and lower case letters are different, i.e., C is case sensitive.

The *array* is another kind of variable that is used extensively in C. An array is an identifier that refers to a collection of data items which all have the same name. The data items must all be of the same type (for example, all integers, all characters, etc.). The individual data items are represented by their corresponding *array element*, i.e., the first data item is represented by the first array element, and so on. The individual array elements are distinguished from one another by the value that is assigned to a *subscript*.

Suppose that x is an array having 10 elements. The first element is referred to as x[0], the second as x[1], and so on. The last element will be x[9].

The subscript associated with each element is shown in square brackets. Thus, the value of the subscript for the first element is 0, the value of the subscript for the second element is 1, and so on. For an n-element array, the subscripts always range from 0 to n-1.

There are several different ways to categorize arrays (e.g., integer arrays, character arrays, one-dimensional arrays, multidimensional arrays).

3.5. Declarations

A *declaration* associates a group of variables with a specific data type. All variables must be declared before they can appear in executable statements.

A declaration consists of a data type, followed by one or more variable names, ending with a semicolon. Each array variable must be followed by a pair of square brackets, containing a positive integer which specifies the size (i.e., the number of elements) of the array.

A C program may contain the following type declarations:

```
int a, b, c[20];
```

```
float x, y, z, u;  
char flag, name[25];
```

Thus a, b are declared to be integer variables and c is an 20-element integer array variables x, y, z and u are floating point variables, flag is a char-type variable and name is a 25-element char-type array.

3.6. Expressions

In C, an expression is a combination of variables, constants, operators and functions. An algebraic expression can not be entered to the computer directly. Before entering an algebraic expression in the computer it should be converted to a C expression. Three major types of expression namely (i) arithmetic expression, (ii) logical expression and (iii) character expression are used in C.

Expressions can also represent logical conditions that are either true or false. However, in C the conditions *true* and *false* are represented by the integer values 1 and 0, respectively. Hence, logical-type expressions really represent numerical quantities.

Arithmetic Operators

The following are the symbols for C operators corresponding to different arithmetic operators.

Operator	Purpose
+	Addition
-	Subtraction
*	Multiplication
/	Division
%	Remainder after integer division

The operator % is sometimes referred to as the *modulus operator*. There is no exponentiation operator in C. However, there is a *library function* (pow()) to carry out exponentiation.

Integer expression

The mathematical expression obtained by combining integer variables and integer constants with the help of arithmetic operators, is known as *integer expression*.

Note on integer division

When an integer is divided by another integer then the result may contain a fractional part. But in C, the fractional part is discarded during calculation. Let a=5, b=2. Then

a + b = 7
a - b = 3
a / b = 2 (not 2.5)
a % b = 1

The value of the expression 32/5*5+8/9 is calculated as follows:

$$32/5*5+8/9=6*5+0=30+0=30$$

Real expression

The mathematical expression obtained by combining real constants and real variables with the help of arithmetic operators, is known as a *real expression*.

Note:

All the exponents are evaluated first. After completion of exponents the expression is scanned from left to right, to complete execution of all divisions and multiplications in the order of their appearance. Finally, all additions and subtractions are performed again from the left of the expression. If there is any parentheses within the expression then the expression within the parentheses is evaluated first using the above rules.

Hierarchy of arithmetic operators

The hierarchy of operators is the order in which the arithmetic operations are executed. The hierarchy of arithmetic operators is shown in the following table.

Operator	Hierarchy
multiplication (*), division (/) and modulo (%)	First
addition (+) and subtraction (-)	Second

Relational operators

We often compare two quantities and depending on their relation, take certain decisions. For example, we may compare the age of two persons, or the price of two items and so on. These comparisons can be done with the help of *relational operators*. The value of a relational expression is either *one* or *zero*. It is one if the specified relation is *true* and zero if the relation is *false*. For example, $20 < 30$ is true while $40 < 30$ is false. The six relational operators and their meanings are shown below:

Operator	Meaning
<	Is less than
<=	Is less than or equal to
>	Is greater than
>=	Is greater than or equal to
==	Is equal to
!=	Is not equal to

Logical operators

In addition to the relational operators, C has the following three *logical operators*.

Operator	Meaning
&&	Logical AND
	Logical OR
!	Logical NOT

The logical operations && and || are used when we want to test more than one condition and make decisions. An example is : $x < y \ \&\& \ x < z$

An expression of this kind which combines two or more relational expressions is termed as a *logical expression* or a *compound relational expression*.

Increment and decrement operators

The language C has two very useful operators not generally found in other languages. These are the *increment* and *decrement* operators: ++ and --

The operator ++ adds 1 to the operand while -- subtracts 1. Both are unary operators. They take the following form: ++a or a++, --b or b--.

++a or a++ is equivalent to $a = a + 1$

-- a or a -- is equivalent to a=a-1

3.7. Mathematical Functions

There are certain mathematical functions such as trigonometric, exponential, logarithmic, etc. which are frequently used in programs, especially to solve scientific problems. These functions are generally referred as *library functions*, *built-in functions* or *intrinsic functions*. The general form of these functions is

Function name (argument)

The argument of a function can be a valid variable name or an expression. It can also be a real number, an integer, a character or a string. The argument should be written within parentheses. If a function has more than one argument then they are separated by comma. Some commonly used library functions are listed in the following table.

Function	Meaning
sqrt(x)	Square root of x, $x \geq 0$
pow(x,y)	x to the power y (x^y)
log(x)	Natural log of x, $x > 0$
log10(x)	Base 10 log of x, $x > 0$
exp(x)	the power x (e^x)
fabs(x)	absolute value of x
fmod(x,y)	remainder of x/y
ceil(x)	x rounded up to the nearest integer
floor(x)	x rounded down to the nearest integer
cos(x)	cosine of x
sin(x)	sine of x

All these functions are not available in C program unless we use the following statement at the beginning of the program.

#include<math.h>

3.8. Assignment Statement

An assignment statement is an executable statement in which a new value is assigned to a variable. Its general form is

$v=e$

where v is a variable and e may be a constant, another variable or an expression to which a value has been assigned previously or a formula which can be evaluated by the computer.

Some arithmetic assignment statements are

x=5.9

z=a

z=a+b*c.

3.9. Input/Output Statements

To enter the value of variables to the computer the input and output statements are used. C does not have any built-in input-output statements as part of its syntax. All input/output operations are carried out through function calls such as scanf() (input statement) and printf() (output

statement). These functions are collectively known as the *standard I/O library*. Each program that uses a standard input/output function must contain the statement

```
#include <stdio.h>
```

at the beginning. The file name `stdio.h` is an abbreviation for *standard input-output header file*. The instruction `#include <stdio.h>` tells the computer 'to search for a file named *stdio.h* and place its contents at this point in the program.' The contents of the header file become part of the source code when it is compiled.

Reading a character

The simplest of all input/output operations is reading a character from the standard input unit (usually the keyboard) and writing it to the standard output unit (usually the screen). Reading a single character can be done by using the function `getchar()`. The syntax of this function is

```
variablename=getchar()
```

`variablename` is a valid C variable that has been declared as `char` type. When this statement is executed, the computer waits until a key is pressed and then assigns this character as a value to `getchar()` function. For example,

```
char sex;  
sex=getchar();
```

If we press the key 'm' then the value of the variable becomes 'm'

Writing a character

Like `getchar()`, there is a similar function `putchar()` for writing character one at a time to the terminal. Its general form is

```
putchar(variablename);
```

where `variablename` is a type `char` variable containing a character. This statement displays the character contained in the `variablename` at the screen. For example,

```
sex='m';  
putchar(sex);
```

will display the character 'm' on the screen. The statement

```
putchar('\n');
```

would cause the cursor on the screen to move to the beginning of the next line.

The scanf() function

Input data can be entered into the computer from a standard input device by means of the C library function `scanf()`. This function can be used to enter any combination of numerical values, single character and string. The function returns the number of data items that have been entered successfully.

The general form is

```
scanf(control string, arg1, arg2, ..., argn);
```

where `control string` refers to a string containing certain required formatting information, and `arg1, arg2, ..., argn` are arguments that represent the individual input data items.

The `control string` comprises individual groups of characters, with one character group for each input data item. Each character group must start with a percent sign (%). In its simplest form, a single character group will consist of the percent sign, followed by a *conversion character* which indicates the type of the corresponding data item. Some commonly used conversion characters for data type is shown below:

Conversion character	Meaning
c	Data item is a single character
d	Data item is a decimal integer
f	Data item is a decimal floating-point value
i	Data item is a decimal integer, octal integer or hexadecimal integer
o	Data item is an octal integer
x	Data item is a hexadecimal integer
s	Data item is a string

```

char name[25];
int rollno;
float age;
scanf("%d %s %f",&rollno,name,&age);

```

Within the `scanf()` function, the control string is `"%d%s%f"`. It contains three character groups. The first character group, `%d`, indicates that the first argument (`&rollno`) represents a decimal integer value. The second character group, `%s`, indicates that the second argument (`name`) represents a string, and the third character group, `%f`, indicates that the third arguments (`&age`) represents a floating point value.

The printf() function

Output data can be written from the computer onto a standard output device using the library function `printf()`. This function can be used to output any combination of numerical values, single characters and strings. It is similar to the input function `scanf()`, except that its purpose is to display data rather than to enter data into computer. That is, the `printf()` function moves data from the computer's memory to the standard output device. The general form of `printf()` function is

```
printf(control string, arg1, arg2, . . . , argn);
```

where control string refers to a string that contains formatting information, and `arg1, arg2, . . . , argn` are arguments that represent the individual output data items. The arguments can be constants, single variable or array or more complex expressions. Function references may also be included. The control string is composed of individual groups of characters, with one character group for each output data item. Each character group must begin with a percent sign (%). In its simplest form, an individual character group will consist of the percent sign followed by a *conversion character* indicating the type of the corresponding data item.

The following C statements illustrate the use of the `printf()` function.

```

int i=10;
float x=23.923;
printf("The value of i=%3d and the value of x=%f",i,x);

```

These statements will print the following output
The value of i= 10 and the value of x=23.923000

The gets() and puts() functions

These functions accept a single argument. The argument must be a data item that represents a string (e.g., a character array). The string may include whitespace characters. In the case of `gets()`, the string will be entered from the key board and will terminate with a newline character (i.e., the string will end when the user presses the return key).

3.10. Complete programs

Now, from the ongoing discussions we have sufficient resources to write a complete program in C.

Example 1. Write a program to find Fahrenheit temperature from the corresponding Centigrade temperature.

Solution: We have the relation between Fahrenheit (F) and Centigrade (C) temperature as

$$\frac{C}{5} = \frac{F-32}{9} \quad \text{or} \quad F = \frac{9C+160}{5}$$

```
/* Conversion from Centigrade to Fahrenheit */
#include<stdio.h>
main()
{
    float c, f;
    scanf("%f", &c);
    f=(9*c+160)/5.0;
    printf("The Centigrade temperature is %f when Fahrenheit temperature is
%f\n", c, f);
}
```

Example 2. Write a program to find the sum and average of five real numbers.

Solution: We denote the variable as a, b, c, d, e. The sum of them is $\text{sum} = a+b+c+d+e$ and average is $\text{average} = (a+b+c+d+e)/5 = \text{sum}/5$.

```
/* Program to find sum and average of five numbers */
#include<stdio.h>
main()
{
    float a,b,c,d,e;
    scanf("%f%f%f%f%f", &a,&b,&c,&d,&e);
    sum=a+b+c+d+e;
    average=sum/5;
    printf("The sum=%f and average=%f\n", sum, average);
}
```

3.11. Control Statements

In most of C programs we have encountered so far, the instructions were executed in the same order in which they appeared in the program. Each instruction was executed once and only once. Programs of this type are very simple, since they do not include any logical control structures. Many programs require that a group of instructions be executed repeatedly, until some logical condition has been satisfied. This is known as *looping*. Sometimes the required number of repetitions will not be known in advance; the computation simply continues until the logical condition becomes true. In C the most useful looping statements are while, do-while and for.

The while statement

The while statement is used to carry out looping operations. The general form of the statement is

while (expression) statement

Or

```
while (expression)
{
    statements
}
```

The included *statement* will be executed repeatedly, as long as the value of *expression* is not zero. This *statement* can be simple or compound, though it is typically a compound statement. It must include some feature which eventually alters the value of *expression*, thus providing a stopping condition for the loop. Usually, the *expression* is a logical expression that is either true or false. Thus, the *statement* will continue to execute as long as the logical expression is true.

The following program will find the sum of first 20 natural numbers using while loop.

```
#include<stdio.h>
main()
{
    int x=1, sum=0;
    while (x<=20)
    {
        sum=sum+x;
        ++x;
    }
    printf("The sum of first 20 natural numbers is %d",&sum);
}
```

The do-while statement

When a loop is constructed using the while statement, the test for continuation of the loop is carried out at the beginning of each pass. Sometimes, it is desirable to have a loop with the test for continuation at the end of each pass. This can be done with do-while loop.

The general form of the do-while statement is

```
do {
    statement;
}
while(logical expression)
```

The included *statement* will be executed repeatedly, as long as the value of *expression* is not zero. Notice that *statement* will always be executed at least once, since the test for repetition does not occur until the end of the first pass through the loop. The *statement* can be either simple or compound, though most applications will require it to be a compound statement. Usually logical expression is a logical expression which is either true or false. The included *statement* will be repeated if the logical expression is true.

The for statement

The for statement is the third and perhaps the most commonly used looping statement in C. This statement includes an expression that specifies an initial value for an index, another expression that determines whether or not the loop is continued and a third expression that allows the index to be modified at the end of each pass.

The general form of the do statement is

```

for(expression1; expression2; expression3)
{
    statement;
}

```

where *expression1* is used to initialize some parameter (called an index) that controls the looping action, *expression2* represents a condition that must be satisfied for the loop to continue execution and *expression3* is used to alter the value of the parameter initially assigned by *expression1*. Generally, *expression1* is an assignment expression, *expression2* is a logical expression and *expression3* is a unary expression or an assignment expression.

When the for statement is executed, *expression2* is evaluated and tested before each pass through the loop, and *expression3* is evaluated at the end of each pass.

Example 3. Write a program to find the sum and average of n real numbers.

Solution. Let $x[0], x[1], x[2], \dots, x[n-1]$ be the set of n real numbers. Their sum and average are computed by the following formula

$$\text{sum} = \sum_{i=0}^{n-1} x[i] \text{ and average} = \text{sum}/n.$$

```

/* Program to find the sum and average of n real numbers */
#include<stdio.h>
main()
{
    int n,i;
    float x[50];
    float sum=0, average;
    printf("Enter the value of n ");
    scanf("%d",&n);
    printf("Enter data\n");
    for(i=0;i<n;i++) scanf("%f",&x[i]);
    for(i=0;i<n;i++)
        sum+=x[i];
    average=sum/n;
    printf("The sum=%f and average=%f",sum,average);
}

```

3.12. Decision making and branching

The C language possesses decision making capabilities and supports the following statements known as *control* or *decision making* statements.

- if statement
- switch statement
- Conditional operator statement
- goto statement

Here we shall discuss if and goto statements.

if statement

The if statement is a powerful decision making statement and is used to control the flow of execution of statements. It is basically a two-way decision statement and is used in connection with an expression. Its general form is

if (logical expression)

It allows the computer to evaluate the expression first and then, depending on whether the value of the expression is true or false, it transfer the control to a particular statement.

The if statement may be implemented in different forms depending on the nature of the condition to be tested.

- Simple if statement
- if...else statement
- Nested if...else statement

The simple if statement

The general form of a simple if statement is

```
if(logical expression)
{
    statement;
}
```

The *statement* may be a single statement or a group of statements. If the *logical expression* is true, the *statement* will be executed; otherwise the *statement* will be skipped and the execution will jump to the outside if block. It may be noted that if the *logical expression* is true then both the *statement* and the outside statement are executed in sequence.

Example 4. Write a program to find the maximum among n real numbers.

Solution. Let $x[0], x[1], x[2], \dots, x[n-1]$ be the set of n real numbers. To find the maximum, initially we set $\text{max}=x[0]$; for remaining data we check the condition $x[i]<\text{max}$. If it is true then max will be updated as $\text{max}=x[i]$ otherwise no action is required. The following program finds the maximum among n numbers.

```
/* Program to find the maximum among n real numbers */
#include<stdio.h>
main()
{
    int n,i;
    float x[50],max;
    printf("Enter the value of n ");
    scanf("%d",&n);
    printf("Enter data\n");
    for(i=0;i<n;i++) scanf("%f",&x[i]);
    max=x[0];
    for(i=0;i<n;i++)
        if(x[i]<max) max=x[i];
    printf("The maximum among the given numbers is %f",max);
}
```

The if...else statement

The **if...else** statement is an extension of the simple **if** statement. The general form is

```
if(logical expression)
{
    true-block statement;
}
else
{
    false-block statement;
}
```

If the *logical expression* is true, then the *true-block statement*, immediately following the **if** statement are executed; otherwise, the *false block statement* are executed. In either case, either *true block* or *false block* will be executed, not both.

Let us consider an example of counting the number of male and female students in a class. We use code 1 for male and 2 for female students. To find number of male and female students we may use the following program segment.

```
if(code==1)
{
    male++;
}
else
{
    female++;
}
```

The goto statement

Like other languages, C also support the **goto** statement to **branch unconditionally** from one point to another in the program. The **goto** statement requires a *label* in order to identify the place where the branch is to be made. A *label* is any valid variable name, and must be followed by a colon. The general form is shown below:

```
goto stop;
statement(s);
stop:
statement;
```

3.13. Worked out examples

Example 5. Write a program to find the maximum among three numbers.

Solution: The following program finds maximum among three numbers using simple **if** statement.

```
/* Program to find maximum among 3 numbers, using if statement */
#include<stdio.h>
main()
{
    float a,b,c,max;
    printf("Enter three numbers\n");
    scanf("%f%f%f",a,b,c);
```

```

max=a;
if(max<b) max=b;
if(max<c) max=c;
printf("The maximum number is %f",max);
}

```

Example 6. Write a program to find the standard deviation of a sample of size n .

Solution: Let $x[0], x[1], \dots, x[n-1]$ be a sample of size n . The standard deviation of this sample is

$$s = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} x_i^2 - \bar{x}^2}$$

where \bar{x} denotes the mean of the sample.

```

/* Program to find the standard deviation of a sample */
#include<stdio.h>
#include<math.h>
main()
{
    int i,n;
    float x[50],s,sum=0,sum2=0;
    printf("Enter sample size ");
    scanf("%d",&n);
    for(i=0;i<n;i++)
        scanf("%f",&x[i]);
    for(i=0;i<n;i++)
    {
        sum+=x[i];
        sum2+=x[i]*x[i];
    }
    xbar=sum/n;
    s=sqrt(sum2/n-xbar*xbar);
    printf("The standard deviation of the given sample is %f",s);
}

```

3.14. User-defined function

Like other programming languages, C also provide the facility of functions. Such facility is the most powerful and convenient way to solve a large problem. The programmer can divide a large problem into several smaller subprograms (blocks), which when reassembled constitutes the complete C program. C functions can be classified into two categories, namely, *library* functions and user-defined *functions*. The `main()` is an example of user-defined function. The functions `printf()` and `scanf()` are library functions. Also `sin()`, `cos()`, `log()`, etc are library functions. The main difference between these two categories is that library functions are not required to be written by user whereas a user-defined function has to be developed by the user at the time of writing a program. The function `main()` is a specially recognized function in C. Every program must have the `main()` function to indicate where the program has to begin its execution.

Example 7. Write a program to find the correlation coefficient of a bivariate sample of size n.

Solution. The correlation coefficient is denoted by r and defined by

$$r = \frac{\text{cov}(x, y)}{S_x S_y}, \quad \text{where } \text{cov}(x, y) = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=0}^{n-1} x_i y_i - \bar{x} \bar{y},$$

$$S_x = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} x_i^2 - \bar{x}^2} \quad \text{and} \quad S_y = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} y_i^2 - \bar{y}^2}$$

Let $\text{sum1}(x[i]) = x[0] + x[1] + \dots + x[n-1]$; $\text{sum2}(x[i]) = x[0]^2 + x[1]^2 + \dots + x[n-1]^2$; and similarly $\text{sum1}(y[i]) = y[0] + y[1] + \dots + y[n-1]$; $\text{sum2}(y[i]) = y[0]^2 + y[1]^2 + \dots + y[n-1]^2$ and $\text{sumxy}(x[i], y[i]) = x[0]y[0] + x[1]y[1] + \dots + x[n-1]y[n-1]$.

Therefore, $S_x = \sqrt{\text{sum2}(x)/n - (\text{sum1}(x)/n)^2}$ and similarly $S_y = \sqrt{\text{sum2}(y)/n - (\text{sum1}(y)/n)^2}$ and $\text{cov}(x, y) = \text{sumxy}(x, y)/n - (\text{sum1}(x)/n)(\text{sum1}(y)/n)$.

The C program is given below.

```
/* Program to find the correlation coefficient of a sample of size n */
#include<stdio.h>
#include<math.h>
main()
{
    int i, n;
    float x[50], y[50], r, sx, sy;
    float sum1(float x[]), sum2(float x[]), sumxy(float x[], float y[]);
    printf("Enter sample size ");
    scanf("%d", &n);
    printf("Enter the sample x[i], y[i]\n");
    for(i=0; i<n; i++)
        scanf("%f%f", &x[i], &y[i]);
    sx=sqrt(sum2(x)/n-(sum1(x)/n)*(sum1(x)/n));
    sy=sqrt(sum2(y)/n-(sum1(y)/n)*(sum1(y)/n));
    cov=sumxy(x, y)-(sum1(x)/n)*(sum1(y)/n);
    r=cov/(sx*sy);
    printf("The correlation coefficient of the given sample is %f", r);
}

float sum1(float a[])
{
    float s;
    int i;
    for(i=0; i<n; i++) s+=a[i];
    return(s);
}

float sum2(float a[])
{
    float s2;
    int i;
    for(i=0; i<n; i++) s2+=a[i]*a[i];
    return(s2);
}
```



```

float sumxy(float a[], float b[])
{
    float ss;
    int i;
    for(i=0;i<n;i++) ss+=a[i]*b[i];
    return(ss);
}

```

Module summary

In this module we discussed three topics MS-Word, MS-Excel and C programming language. All the three topics are introduced in very simple way. The frequently used features of MS-Word and MS-Excel are discussed. Open, create, delete, modify, save, load, formatting, printing of MS-Word and MS-Excel documents are discussed in this module. Most of the topics are explained with the help of simple examples. A self assessment exercise is supplied at the end of the module.

Self Assessment Questions

Section 1: Microsoft Word 2000

1. What are the different ways to create a Word document? Explain each.
2. What is Print Layout View?
3. What is the difference between Page Layout view and Print Layout view?
4. How would you find a word or phrase in a document? Explain the method briefly.
5. What are headers and footers? How would you go about creating a header or a footer?
6. What are the different types of pictures that can be inserted into Word documents?
7. Describe a method to create a table.
8. Is it possible to convert a table into text? If yes, how?
9. What is a cell in a table?
10. What is mail merge feature? Is it possible to create envelopes using mail merge?

Section 2: Excel 2000

1. What is a spreadsheet?
2. What are the advantages of using spreadsheet?
3. What do you mean by a cell in a spreadsheet?
4. What are the three components of Excel and what do they do?
5. What are the steps involved in building a workbook?
6. What is the use of fill handle? Discuss all of its uses.
7. What is a workbook and what is its use?
8. Explain insertion and deletion of rows and columns method in a spreadsheet.
9. How is a workbook protected? What is the advantage of protecting a file?
10. How do we create a formula in a spreadsheet?
11. Explain the types of formula in brief.
12. Why and how do we format a workbook?
13. What is the purpose of Function Wizard?
14. How is the AutoSum feature used and what are its advantages?

Section 3: Programming in C

1. Explain with example do-while loop in C.

2. What do you mean by branching and looping in a programming language?
3. Explain for loop of C programming language.
4. What are the uses of scanf() and printf() functions in C.
5. Explain if statement with example.
6. What are the purpose of gets() and puts() functions in C.
7. Write a program to find Fahrenheit temperature from the corresponding Centigrade temperature.
8. Write a program to find the largest among four given numbers.
9. Write a program to find the minimum among n given numbers.
10. Write a program to find the maximum and minimum among n given numbers.
11. Write a program to find the sum of n numbers.
12. Write a program to find the average among of n numbers.
13. Write a program to find the standard deviation of a sample of size n.
14. Write a program to find the correlation coefficient of a sample of size n.
15. Write a program to find the median of a simple ordered sample of size n.
16. Write a program to find the mode of a simple sample of size n.

References

1. Working with MS-Office 2000, CDG, Tata McGraw-Hill Publishing Company Limited, New Delhi, 2001.
2. Introduction to Computers with MS-Office 2000, A.Leon and M.Leon, Tata McGraw-Hill Publishing Company Limited, New Delhi, 2001.
3. Programming with C, E. Balagurusamy, Tata McGraw-Hill Publishing Company Limited, New Delhi.

M. Sc. in Botany

Paper - II

(2nd Half)

Module No . 18 (A)

Contributor :

Prof. P. C. Dhara

Module structure

- 1.0 Introduction
- 2.0 Aim of the module
- 3.0 Features of the BASIC language
 - 3.1 Elements of Basic
 - 3.1.1 The character set
 - 3.1.2 Constants and variables
 - 3.1.3 BASIC Expression
- 4.0 Construction of BASIC Program
 - 4.1 BASIC Statements :
 - 4.1.1 LET Statement
 - 4.1.2 INPUT Statement
 - 4.1.3 READ, DATA Statement
 - 4.1.4 PRINT Statement
 - 4.1.5 REM Statement
 - 4.1.6 END Statement
 - 4.1.7 RESTORE Statement
- 5.0 Some simple programs
- 6.0 BASIC Statements for branching operations
 - 6.1 GOTO Statement
 - 6.2 IF.....THEN Statement
- 7.0 Looping operation
 - 7.1 FOR.....NEXT Statement
 - 7.1.1 Nested Loop
- 8.0 Arrays
 - 8.1 Subscripted variables
 - 8.2 DIM Statement
- 9.0 Library Functions
- 10.0 Some application programs in botany
- 11.0 Running BASIC program in computer
 - 11.1 Entering into GWBASIC
 - 11.2 Editing the program
 - 11.3 Listing the program
 - 11.4 Renumbering lines
 - 11.5 Running the program
 - 11.6 Saving the program
 - 11.7 Starting a new program
 - 11.8 Loading a program
 - 11.9 Printing a program and output
 - 11.10 Deleting the statement
 - 11.11 Quitting the BASIC
- 12.0 Diagnosis of programming errors
- 13.0 Summary
- 14.0 Bibliography
- 15.0 Model Questions
 - 15.1 Long Questions
 - 15.2 Short Questions

PROGRAMMING IN BASIC

1.0 INTRODUCTION

Computer program is very much helpful for solving problems. Those can be used for computation of small, medium and large sized data. In biological sciences, including Botany, computation of data, which are gathered from the laboratory as well as from the field, is frequently required. A simple program may reduce the time and labour of the users. BASIC programs may a useful solution for this purpose. Basic is a very powerful language as a tool for the novice programmer.

BASIC is a high level language. BASIC Stands for Beginner's All-purpose Symbolic Instruction Code. The original BASIC language was designed in 1963 by John Kemeny and Thomas Kurtz and implemented by a team of Dartmouth students under their direction. BASIC was designed to allow students to write programs for the Dartmouth Time-Sharing System. It was intended to address the complexity issues of older languages with a new language design specifically for the new class of users that time-sharing systems allowed—that is, a less technical user who did not have the mathematical background of the more traditional users and was not interested in acquiring it. Being able to use a computer to support teaching and research was quite novel at the time. In the following years, as other dialects of BASIC appeared, Kemeny and Kurtz's original BASIC dialect became known as *Dartmouth BASIC*.

The BASIC language was designed to be conversational right from the start. This can put the programmer or user into direct communication with computer, usually through a teletype terminal. In this interactive mode, the user can enter his program statements directly into the computer memory and any errors in the statements will be immediately displayed. Thus, the user can correct his mistakes immediately. While running the program, the programmer can ask for results at intermediate points and check for the correctness of his program logic without having to wait for the computer to reach the end of the program.

These languages introduced many extensions to the original home computer BASIC, such as improved string manipulation and graphics support, access to the file system and additional data types. More important were the facilities for structured programming, including additional control structures and proper subroutines supporting local variables.

However, by the latter half of the 1980s newer computers were far more capable with more resources. At the same time, computers had progressed from a hobbyist interest to tools used primarily for applications written by others, and programming became less important for most users. BASIC started to recede in importance, though numerous versions remained available. Compiled BASIC or CBASIC is still used in many IBM 4690 OS point of sale systems.

2.0. AIM OF THE MODULE

You should learn a computer program. This will enable you to get confidence about the application of the computer. This will help you to learn other programming language easily. Not only that you can analyze your data very quickly and correctly. The main aim of this module is that you will be taught about the basics of the BASIC programming. From this module you will be aware of -

- elements of BASIC
- different statements used in the BASIC
- framing of the statements in the programming
- use of branching in the BASIC programming
- use of looping in the programming
- the way of writing application programs in Botany
- significance of different commands for running BASIC in the computer.

3.0 FEATURES OF BASIC LANGUAGE

The BASIC was developed on the following design principles:

1. Be easy for beginners to use.
2. Be a general-purpose programming language.
3. Allow advanced features to be added for experts (while keeping the language simple for beginners).
4. Be interactive.
5. Provide clear and friendly error messages.
6. Respond quickly for small programs.
7. Not to require an understanding of computer hardware.
8. Shield the user from the operating system

The designers of the language decided to make the compiler available free of charge so that the language would become widespread. They also made it available to high schools in the Dartmouth area and put a considerable amount of effort into promoting the language. As a result, knowledge of BASIC became relatively widespread (for a computer language) and BASIC was implemented by a number of manufacturers, becoming fairly popular on newer minicomputers like the DEC PDP series and the Data General Nova. The BASIC language was also central to the HP Time-Shared BASIC system in the late 1960s and early 1970s. In these instances the language tended to be implemented as an interpreter, instead of (or in addition to) a compiler.

The main features of the BASIC are as follows:

1. BASIC is a user oriented language and can be learnt easily. Any person with a little sense of logic can learn programming in BASIC. It can also be used as a stepping stone for learning other languages.
2. The flexibility of BASIC allows the programmer to write, run, modify and debug the program easily and quickly.

3. The BASIC programs. entering data is easy and programmers need not be confused about output formats because of facility for allowing more sophisticated formats for output.
4. The use of BASIC is not only restricted to elementary application but it is also useful for advanced applications.
5. It is relatively machine-independent language. Though there are minor differences between various versions of BASIC, programmers can run programs in different computers only with minor changes in the programs.

3.1 ELEMENTS OF BASIC:

Before entering the BASIC programming some elements of BASIC should be learnt. Important elements are discussed in the following subsection.

3.1.1 THE CHARACTER SET:

A letter, digit, punctuation mark, or specific Symbols used in programs are called characters. Like every language BASIC has its own character set. The following are the recognized character set of BASIC.

- i) Alphabets : A,B,C.....,Z.
- ii) Digits : 0,1,2,.....,9.
- iii) Special characters :

Addition	+
Subtraction	-
Multiplication	*
Division	/
Exponentiation	↑ or ^
Left parenthesis	(
Right parenthesis)
Equal to	=
Not equal to	≠
Less than	<
Less than equal to	<=
Greater than	>
Greater than equal to	>=
Comma	,

Colon	:
Semicolon	;
Quotation mark	"
Dollar Sign	\$
Exclamation point	!
Percent sign	%
Number sign	#
Blank	

3.1.2 CONSTANTS AND VARIABLES:

In BASIC, expression are used which may be constants or variables.

Constants: A quantity in a computer program which does not change its value during the execution of the program is called, a **Constant**. BASIC allows two types of constants.

- i) Numeric Constant
- ii) String Constant

Numeric Constant: A numeric constant is one that is formed by a sequence of digits, 0, 1, 2,..... 9. A numeric constant is also known as numbers. A numeric constant may or may not include a decimal point, i.e., it may be an integer or a real number. The numeric constant may be positive or negative. The sign + or - must precede the number. If no sign is indicated, the number is assumed to be positive. The followings are the valid numeric constants:

495 +781 0 -16.48

The numbers can be expressed in exponential notation. The exponential notation is used for representing very large and very small numbers. E represents base 10, and the signed integer following E is called **exponent**. The signed number (integer or real) preceding the letter E is called **mantissa**. The exponent can be positive or negative but cannot contain decimal point, the length of exponent cannot be more than two digits. There is no restriction on mantissa.

In the exponential format a given number is represented as a value between 1 and 10 multiplied by the power of 10. For example, if the given number is 12340000

then it is displayed as 1.234 (which is a number between 1 and 10) multiplied by 10 to the power of 7 (1.234×10^7) as 1.234E7 in GW-BASIC.

Example:

Number	Equivalent form	Exponential form
693215	6.93215×10^5	6.93215E05
-4423	-4.423×10^3	-4.423E+03
-0.0021	-2.1×10^{-3}	2.1 E-03

RULES:

- BASIC does not distinguish between an integer and a real.
- Commas are not allowed in a numeric constant.
- The limit on the number of digits that can be used in a numeric constant varies from computer to computer. Normally, a numeric constant can have up to a maximum of eight digits.

Invalid numeric constants

- 5522+ → plus sign is included in the right side of the number
- 2457, 86 → comma is not allowed within a number.
- 83 596 → between the digits 3 and 5 there is a blank
- 899E-0.5 → the exponent should not have decimal point.

String Constant: BASIC can handle non-numeric data called string constants. String constants are composed of valid basic characters (alphabets, digits, special characters, even blank) enclosed within quotation marks. The quotation marks help the computer to detect the beginning and the end of the string constant. String constant are used to represent non-numeric expression, such as, names, addresses, date, nature of the problem etc.

Rule: Mixture of numeric and string constants is not permissible.

In case of mixture of numeric and string GW-BASIC will display error message "Type mismatch".

Some examples of valid and invalid string constants are given below-

Valid String Constants

"Vidyasagar University"

"Mean Value ="

"98765" [it is string constant but 98765 is a numeric constant]

"*****"

Invalid String Constant

Botany [quotation mark absent]

"696 [right quotation absent]

"Rupees" 500 [Mixture of numeric and String constants]

Variables:

The quantity which may change its values during the execution of the program is called a **variable**. We use variables for the purpose of storing information (number of string) in the memory of the computer. It is actually a temporary storage area for a piece of information. When computer encounters a variables name in the program it immediately assigns a memory location by the variable name. The value of a variable can be changed by the program. BASIC accepts two types of variables:

- i) Numeric variable
- ii) String Variable

Numeric Variables:

Numeric variables are names that are used for assigning locations in computer memory for storing numeric constants. The value of a variable is either supplied by the programmer or by a result of computations made by the computer during the execution of the program.

Each of the numeric variables should have a name. The followings are the rules for presenting a variable name.

Rules:

- a) A numeric variable is represented by an alphabet (A, B, ..., Z), numbers (0,1,2,...9) and the decimal point.
- b) The first character in a variable name must be a letter. It may be represented by a letter followed by a digit (e.g., P5, 7.6 etc.), letter (e.g. abc, xy etc.) or decimal point (e.g. , Specis.1, p2.k, etc.).
- c) No space or punctuation mark is allowed.
- d) Reserved keywords, which are used in the BASIC programming, like READ, LET, and PRINT etc. cannot be used as variable names.

The variable names may be of any length, but only 40 characters are significant.

Valid numeric variables: A, T2, FLOWER1, FUNGI.500

Invalid numeric variables: 8k, D\$, A:B, ", 100SEEDS

String Variables:

A string variable is used to represent the value of character string. To store a string constant we need a 'string variable'. A string variable is the same as a numeric variable with one difference, string variables must end with a dollar (\$) symbol. If you put a dollar sign at the end of any numeric variable name it will become a string variable and can be used to store any string constant. As you cannot store a string constant into a numeric variable into a string variable, trying to store a number or numeric constant into a string variable will also give you "Type mismatch" error message.

Valid String variables: A\$, Y2\$, GENUSS, NAMES

Invalid String Variables: b, 7C\$, \$D2

3.1.3 BASIC EXPRESSIONS:

In BASIC two types of variables are used - arithmetic expression and logical expression.

Arithmetic Expression:

An arithmetic expression may be a numeric constant, a numeric variable, a function or a combination of numeric constants, variables and functions, connected by arithmetic operators to form a single value.

The following arithmetic operators are recognized by BASIC.

Operators	Operations
+	Addition
-	Subtraction
*	Multiplication
/	Division
↑ or ^	Exponentiation

In a particular arithmetic expression, there is a hierarchy for the order of execution. All exponentiation operations are performed first, then multiplication/division and the addition/subtraction operations are last to be carried out. Within a particular hierarchical group, the operations are executed from left to right. Normal hierarchy of operations can be altered by use of parentheses. The operations within the innermost parentheses are performed first and then the second innermost and so on.

Rules of Writing Arithmetic Expression:

- i) Two operations must not appear together. For example, $A + - B$, $X - / C$ etc. are not permitted.
- ii) Zero should not be raised to a negative point.

- iii) String constants and string variables should not be used in arithmetic expression. For example $A1 \div B\$$, $N + \text{"RAM"}$ are wrong.
- iv) When brackets are used they must be used in pairs, i.e., every left bracket must be matched with a right bracket.

The following table shows some examples of arithmetic expressions.

Table -1: BASIC equivalents of algebraic expressions

Algebraic Expression	BASIC Equivalent
$a + b - cd$	$A + B - C * D$
$(p + q)k + j$	$(p + q) * K + J$
$(x/y)z$	$(X/Y) * Z$
$B^2 - 4AC$	$B \wedge 2 - 4 * A * C$
$((x + y)z^2)/s$	$((X + Y) * Z \wedge 2) / S$
$-p^q$	$-P \wedge q$ or $-(P \wedge q)$

Logical Expression:

Logical expressions are used when it is necessary to compare two numerical or string quantities. A relational expression will have either the value **TRUE** or the value **FALSE**. For example, $X > Y$ will be true if X is greater than Y and false if X is less than or equal to Y . The result of the comparison **TRUE** or **FALSE** will determine the program flow.

Rules:

- i) When arithmetic and logical operators are both used in one expression, the arithmetic operations are performed first.
- ii) String comparison is alphabetical order and all string constants must be enclosed in a quotation mark.

The following are the examples of a few logical expressions:

$A = B$

$x + y > 100$

$P \leq Q - R$

$A \neq B$

$N\$ = \text{"Genetics"}$

4.0 CONSTRUCTION OF BASIC PROGRAM

A BASIC program consists of several statements called BASIC Statements. The statements are made up of **Keywords** and **line numbers** or **statement numbers**. The BASIC statement is either executable or non-executable. The format of BASIC statement is-

Line number # Statement

Keywords: Keywords are the words which have special significance for the interpreter. These words cannot be used as a variable name. For example, LET, INPUT, READ, DATA etc. are some keywords.

Line Numbers: A BASIC program consists of a set of statements put together in a logical manner. The statements are identified by line numbers. Line numbers tell the computer the order in which the statements are to be executed normally. The line number must be an integer between 1 and 99999. The upper limit varies from system to system. Line numbers must be taken in ascending order. It is customary to use the line numbers as 10, 20, 30, 40, 50.....etc, in order to leave room for the insertion of additional line, if required, at a later stage.

4.1 BASIC STATEMENTS:

The elementary BASIC statements used in a program are described below.

4.1.1. LET STATEMENT: The use of LET statement in BASIC is to assign a numeric or string value to a variable. It is also known as assignment statement. The general format of LET statement is

$Ln \text{ LET variable} = \text{constant/ variable/ expression}$

Where, Ln is the line number. The following are a few illustrations of let statements.

10 LET B=5

In the above statement, 10 is the line number and LET is the keyword. Here the variable B is assigned to the numeric constant value 5.

30 N\$ = "Taxonomy"

On line number 30, the string variable N\$ is assigned to a string constant Taxonomy. It should be noted that the string constant is within a quotation.

40 LET X = Y

A numeric variable X is assigned to the value of another numeric variable Y.
The statement

50 LET K\$ = D\$

Assigns the value of a string variable D\$ to another string variable K\$
The Statement

20 LET P = 2* A + 3*B

Assigns the value of expression $2A + 3b$ to the variable P.

Remarks:

- i) In the LET statement the variable on the left of equality sign must assign a value of the same type on the right side. If the variable on the left be a numeric variable then the value on the right of the equality sign must be a numeric value. In other words string value should not be assigned for numeric variable. If a numeric value is assigned for string variable or a string value is assigned for

numeric variable, the condition is called **type mismatch**. An error message "type mismatch" will appear on the screen during execution of the program.

The following statements are valid

```
200 LET K = 15.2
150 LET G$ = "Germ cell"
```

But the following statements are invalid because of type mismatch.

```
20 LET H = "Heredity"
30 LET P$ = 150
```

- ii) Two or more variables cannot be used for assignment purpose in the same LET statement separating each by a comma.

```
50 LET A = 10, B = 20, E$ = "Ecology" is an invalid statement.
```

- iii) In certain situation one can use multiple LET statements. If the same value is required to be assigned to different variables, then multiple equality signs can be used in a statement.

For Example

```
100 LET S1 = S2 = S3 = S4 = 0
```

Remark:

Optional use of the keyword LET

The use of keyword LET for the purpose of assigning value to a variable is optional. The equal sign is sufficient to do the same.

Examples:

```
10 B = 10
150 NS = "MENDEL"
160 X = Y
130 K = K+2
```

The above statements, in which the word LET has been omitted, are also valid.

4.1.2. INPUT Statement

The INPUT statement permits the entry of data during running of a program. The general format of INPUT statement is

Ln INPUT List of variables
where, Ln is the line number.

Examples:

```

30 INPUT A
40 INPUT TS

```

Rules:

- i) The INPUT statement may be associated with only one variable or more than one variable. If more than one variable are used, they must be separated by commas.

Example:

```

10  INPUT S
20  INPUT P,Q
40  INPUT P,Q,C$

```

When INPUT statement is encountered, the program execution stops and ? (Question mark) appears on VDU screen and the computer waits for users/ programmers to enter the value (data) corresponding to the variables listed in the INPUT statement. In case of more than one input variables the user should supply the same number of data as the number of input variables.

- ii). Different types of variables (both numeric and string) can be used in the same INPUT statement. The data to be entered in response to INPUT statement must match the type of variables used in the INPUT statement, that is, there should be no type mismatch.

Example:

```
50 INPUT X, BS, C
```

For above statement one may enter 60.5, "Phloem", 100; but, "Phloem", 60.5, 100 are not valid.

- iii) Entry of formula or expressions for the INPUT variables is not allowed.
- iv) In INPUT statement message with in the quotation may be used, following by a semicolon or comma and then the name of the variables are given.

Example:

```
80    INPUT "ENTER NAME"; N$
```

```
120   INPUT "PUT THE VALUE OF RADIUS"; R
```

During execution of program the message in the quotation will appear in the screen indicating the nature of data to be given.

- v) If the number of data supplied be less than the number of variables in the INPUT statement or data are supplied in wrong manner (Say, numeric for string), an error message "? Redo from start" will appear on the screen.

Disadvantage of INPUT Statement

- i) The INPUT statement is unsuitable for large number of data.
- ii) The data entered through INPUT statement cannot be stored for subsequent use and there is any mistake in entering data for a list of variables it cannot be corrected and data are to be supplied anew.

4.1.3. READ, DATA Statement:

The purpose of the READ Statement is to read data from the DATA Statement present in the program. The general format of READ statement is as follows:

```
Ln    READ List of Variables
```

Where, Ln is the line number.

Examples:

```
80 READ X
50 READ Y$
```

The DATA statement is used in BASIC programs to assign appropriate values (numeric and string) to the variable listed in the READ statement. The general format is –

Ln DATA List of constants

Where, Ln is the line number.

Examples:

```
100 DATA 23.15
200 DATA "Endodermis"
```

Rules:

- i) A READ statement may contain one or more than one variables and the variables may be numeric or string. In case the number of variables in a READ statement be more than one, each variable must be separated by comma.

Example:

```
50 READ P
70 READ P, Q$, R
```

- ii) The list of constants in a DATA statement must be separated by commas. No expression is allowed in DATA statement

Example:

```
20 DATA 5.2, 9, 2.71
110 DATA 20, "Petals"
```

- iii) A READ statement must accompany at least one DATA statement, i.e., a READ statement without a DATA statement is invalid and vice-versa. Because, the DATA statement assigns appropriate values to the variables of the READ statement. During execution of BASIC program when control encounters a READ statement it immediately searches for the DATA statement. There must be exact correspondence between the types of variables in READ statements and the types of values in DATA statements; otherwise DATA type mismatch will occur.

Example:

```
10 READ A, B, RP$  
30 DATA 10, 20, "Root pressure"
```

Assigns 10 to A, 20 to B, and "Root pressure" to RP\$.

For the following statements

```
10 READ A, B, RP$  
30 DATA "Root pressure", 20, 10
```

Data type mismatch will occur because "Root pressure" will be assigned to the numeric variable A.

- iv) Any number of data statements can be used in a program. A single READ statement may access one or more DATA statements in order. Several READ statements may access the same DATA statement.

Example:

```
50 READ A, B, C, DS  
.....  
.....  
100 DATA 10, 30,  
110 DATA 50, "CHLOROPHYLL"
```

Here a single READ statement access several DATA statements.

Example:

```
10 READ M$  
20 READ Y  
30 READ I, D  
.....  
.....  
150 DATA "MARCH", 1999, 1000, 50
```

Here several READ statements access a single DATA statement.

- v) It is not essential that DATA statement must follow READ statement. DATA statement can be given in the program before END statement.
- vi) If the number of variables in a READ statement become greater than the number of values in DATA list then "Out of data" will appear in the VDU.

4.1.4. PRINT Statement:

The Print statement is used to print output data on the printer/terminal. The general format of the print statement is

Ln PRINT List of items

Where, Ln is the line number.

The list contains constants, variables, formula whose values are to be printed.

Example:

```
10 PRINT A
20 PRINT B$
```

Rules:

- i) Each PRINT statement initiates output on a new line.

Example:

```
10 LET A = 5.2
20 LET B$ = "PLANTS"
30 PRINT A
40 PRINT B$
```

The print out of the results will be as follows:

```
5.2
PLANTS
```

- iii) The text can be taken as printout by the PRINT statement. All strings used as message must appear in a PRINT statement within quotation mark. When no item is mentioned in front of PRINT then a blank line will appear, i.e., nothing will be printed on the corresponding output line.

Example:

```
10 READ P, S
20 PRINT "Name of species"
30 PRINT
40 PRINT "Number of Petals="; P
50 PRINT "Number of Sepals="; S
60 DATA 5.4
```

The output of the above statements will be as follows:

Name of species

← blank line

Number of Petals = 5

Number of Sepals = 4

- iv) **Comma Control:** The output medium in BASIC (either paper page or visual display screen) is usually divided into equal zones. The width of each zone varies from 12 to 15 spaces with a space of one to four spaces between the zones. If the items in the PRINT statement be separated by comma, the output of the item following comma will be printed in the next zone. In case the output items do not fit in one line then two or more lines of output will be generated by the same PRINT statement. A pair of commas after an item causes a print zone to be skipped.

Example:

```
30 READ A, B, C, D, E, F
40 PRINT "NUM", "TEMP", "PRESSURE", , "Humidity"
50 PRINT A, B, C, D, E, F
60 DATA 5, 26.5, 76, 75.5, 10, 1.2
```

The output of the PRINT statement (Zone wise) will be follows:

Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
NUM	TEMP	PRESSURE		REMARK
5	26.5	76	5.5	10
1.2				

- v) **Semicolon Control:** The major limitation of the use of Comma is that we cannot print more than 5 items in a line. In many occasions it is advantageous to print more items in each line. And also, one may like to print information where the item name and its value are printed close to each other. These objectives can be met by using the semicolon in place of comma. The items which are separated by semicolons will be printed close together. The computer automatically provides one or two spaces between two items.

Example:

```
10 READ N$, R, M
20 PRINT "NAME"; "ROLL"; "MARKS"
30 PRINT N$; R; M
40 DATA "PUSPA", 25, 82
```


The output will be as follows:

N	A	M	E		R	O	L	L		M	A	R	K	S					
P	U	S	P	A		2	5		8	2									

If a semicolon is given at the end of a PRINT statement, this will suppress the carriage return and the output of the next PRINT statement will continue on the same line.

Example:

```

5    PRINT "DELHI"; "MUMBAI"; "PATNA"
10   READ A, B, C, D, E
20   DATA 111, 212, 313, 414, 515
30   PRINT A; B; C;
40   PRINT D; E

```

Output:

D	E	L	H	I		M	U	M	B	A	I		P	A	T	N	A		S
1	1	1		2	1	2		3	1	3		4	1	4		5	1		5

TAB Function: The TAB function provide additional flexibility for printing output of the items in the PRINT statement. TAB stands for the abbreviation of "Tabulate". It permits the programmer to specify the exact position of each output item in a PRINT statement. The general format of TAB function used in PRINT statement.

TAB (X); List

Where the argument X may be numeric constant, numeric variable or an arithmetic expression. If the value of X be a fraction, then it will be rounded off to the integer value.

Example:

```
60    PRINT TAB (10); X
```

It will cause the printing of the value of X from 10th column onward.

Example:

```
40    A = 10
50    PRINT TAB (5); I; TAB (15); J; TAB (20); K
60    PRINT TAB (A+2); X; TAB (A+3); Y; TAB (A+B); Z
70    PRINT "RUPEES"; TAB (7); R
```

The TAB function can position the cursor/ print head anywhere on the output line but it moves only from left to right. In the following statement

```
20 PRINT "NUMBER OF PLANTS"; TAB (5); N
```

Here, TAB (5) will not have any effect because for printing NUMBER OF PLANTS has already moved through 16 columns and since the argument 5 in TAB is less than 16, therefore, TAB will be ignored.

4.1.5. REM Statement:

The purpose of REM statement in BASIC program is to introduce remarks within the program which provides the information for programmer or any one else to understand the program, to define different variables used in the program, to highlight different segments of the program, to add comment, to make a heading of the program. This statement is not executed by the computer while running the program. The general format is

Ln REM Comments

Where, Ln is the line number.

Example:

```
10    REM Program No. 1
20    REM Calculation of Area
30    REM Roll No. VU/PG/BOT-II/05
```

4.1.6. END Statement:

The purpose of the END statement in a BASIC program is to terminate the program execution. So, END statement must be the last statement of the program. The general format is

Ln END

Where, Ln is the highest statement number of the program.

Example:

200 END

4.1.7. RESTORE Statement:

This statement is used to allow the same data to be read more than once. A situation may arise when the same has to be stored for other variables. The general format of RESTORE statement is

Ln RESTORE

Where, Ln is the line number.

Example:

```
10  READ A, B, C
20  DATA 10, 20, 30
.....
50  RESTORE
60  READ X, Y, Z
```

In the above example the values 10, 20 and 30 are assigned to the variables A, B and C respectively. After RESTORE statement the same values are assigned to the variables X, Y and Z. That is, X=10, Y=20 and Z=30.

5.0 SOME SIMPLE PROGRAMS

Program 1: Solving a simple equation. Suppose, you are given the values of A and B, you have to find out the value of x with the equation $x = B^2 - 2AB$

```
10  REM Solving Simple equation  $B^2 - 4AC$ 
20  INPUT A, B, C
30  LET X = B ^ 2 - 4 * A * C
40  PRINT X
50  END
```

Program 2: Program for finding square, cube and square root of a given number

```
10    REM computing square, cube and square root
20    READ X
30    SQ = X ^ 2
40    CUB = X ^ 3
50    SR = X ^ 0.5
60    PRINT SQ, CUB, SR
70    DATA 10
80    END
```

Program 3: Program for converting temperature given in degree Fahrenheit to degree Celsius.

```
10    REM CONVERSION OF TEMP FROM FAHRENHEIT
20    REM TO CELSIUS
30    INPUT "Give Temperature"; F
40    C = (F-32)*5/9
50    PRINT "TEMP IN CELSIUS="; C
50    END
```

Program 4: Program for student result [Suppose a student obtained different marks in three subjects, e.g., Taxonomy, Forestry, Physiology.] You have to find out total marks obtained in 3 subjects.

```
10    REM computing total marks for a student
20    READ N$, T, F, P
30    TM = T+F+P
40    PRINT "Name:"; N$
50    PRINT "Taxonomy:"; T
60    PRINT "Forestry:"; F
70    PRINT "Physiology:"; P
80    PRINT "Total:"; TM
90    DATA "MADHUMITA ROY", 68, 51, 60
100   END
```

Output of the program

```
Name      : MADHUMITA ROY
Taxonomy   : 68
Forestry   : 51
Physiology : 60
Total      : 179
```

6.0. BASIC STATEMENTS FOR BRANCHING OPERATIONS

The execution of instructions in a BASIC program generally takes place according to the ascending order of line numbers. But sometimes it becomes necessary to change the sequence of execution or to by-pass a few statements. The objective can be achieved by branching operations. Branching operation may be unconditional or conditional.

6.1. GOTO Statement:

GOTO statement causes unconditional transfer of control to a specified statement of the program. The general format of the GOTO statement is

Ln 1 GOTO Ln 2

Where, Ln denotes line numbers.

For example:

50 GOTO 20

Transfer the control from line number 50 to line number 20

GOTO statement can be referred to a line number which may be either executable or non-executable

Example

```
10  INPUT X, Y
20  LET C = X * Y
30  PRINT C
.....
.....
80  GOTO 10
.....
.....
10  END
```

6.2. IF.....THEN Statement:

The IF.....THEN statement is used for conditional transfer of control in a BASIC program. It is a decision making statement and depending upon the decision it can change

the order of execution of statements. It is always used in a conjugation with a relational expression. It allows the computer to check whether a specified relation is **TRUE** or **FALSE** and to perform the transfer of control to a particular statement if it is **TRUE**. This point of program has two branches to follow, one for **FALSE** condition and other for **TRUE** condition.

The general format of the statement is

Ln IF [Relational Expression] THEN Ln

Where, Ln indicates statement number.

For Example:

90 IF A>200 THEN 160

Will transfer the control to line number 160 if A>200, otherwise to next line number.

Example:

40 IF N\$ = "Sourav" THEN 250

Some other formats of IF.....THEN statements are also allowed.

a) Ln IF [Relational Expression] GOTO Ln
[Ln is the line number]

Example:

25 IF X >= Y THEN GOTO 15

b) Ln IF [Relational Expression] THEN Statement

Example:

60 IF P <> Q THEN S = P - Q

c) Ln IF [Relational Expression] THEN Ln/Statement ELSE Ln/Statement

Examples:

20 IF A > 100 THEN P = A-S ELSE P = A+S
 90 IF A >= 10 THEN 50 ELSE B = A

Program 5: A program to find out greater chromosome number between two plants

```
10    REM Greater chromosome number between two plants
20    READ cr1,cr2
30    IF cr1>cr2 THEN 60
40    PRINT cr2
50    GOTO 70
60    PRINT cr1
70    DATA 42, 56
80    END
```

.....
Output
.....

56

Program 6: Suppose you have measured the length of some leaves in inches. Now you have to write a program to convert the length in inch to length in centimeter.

```
10    REM Conversion of leaf length from inch to cm
20    PRINT "Type the length in inch"
30    INPUT L1
40    LC = L1 * 2.54
50    PRINT "Leaf length of leaf in cm.=", LC
60    PRINT "Do you want to continue?"
70    PRINT "If yes. type y, if no, type n"
80    INPUT A$
90    IF A$ = "y" THEN 20
100   END
```

Note: In the above statements from line number 60 to 90 provide an option for the user whether he/ she wishes to convert more data or not.

Program 7: A program for finding largest number of petals among three flowers

```
10    REM largest number of petals among three flowers
20    READ F1,F2, F3
30    IF F1 > F2 THEN 90
40    IF F2 > F3 THEN 70
50    PRINT "Largest no. of petals="; F3
60    GOTO 130
70    PRINT "Largest no. of petals="; F2
80    GOTO 130
90    IF F1 > F3 THEN 120
```

```

100 PRINT "Largest no. of petals="; F3
110 GOTO 130
120 PRINT "Largest no. of petals="; F1
130 DATA 4, 10, 6
140 END

```

.....
Output

Largest no. of petals= 10

Program 8: A program to calculate the sum of spikes in 5 plants

```

10 REM Sum of spikes in 5 plants
20 C = 0
30 SUM = 0
40 READ SP
50 SUM = SUM + SP
60 C = C + 1
70 IF C < 5 THEN 30
80 PRINT "SUM="; SUM
90 DATA 20, 28, 30, 31, 12
100 END

```

.....
Output

SUM = 121

7.0 LOOPING OPERATION:

In many BASIC programs a statement or a number of statements are required to be repeated a number of times. This type of operations can be effectively handled with the help of loops. A GOTO statement can be used to set up a loop because it can cause a jump to a statement that was executed earlier. Loops can also be formed by using FOR.....NEXT Statement.

7.1 FOR.....NEXT Statement:

FOR.....NEXT statements are used when a loop consisting of several statements is to be executed repeatedly in a program. FOR and NEXT statements are used in conjugation. The general format of FOR.....NEXT statement is:


```

Ln1 FOR I = J TO K [STEP L]
.....
Loop .....
.....
Ln2 NEXT I

```

In this case, all the statements between line number 1 to line number 2 are executed while varying the variable I from J to K at the increment of L each time. The variable I is known as **control variable** or **index variable** whose initial value is J and the final value is K. The STEP in this format is optional and in case it is absent, L is assumed to be 1, i.e., increment is taken as 1. All the statements between Line number 1 and Line number 2 are said to form the body of the FOR.....NEXT loop.

Example:

```

10  FOR I = 1 to 10
.....
.....
.....
50  NEXT I

```

} Statements

Example:

```

20  FOR J = 2 TO N STEP 2
.....
.....
.....
80  NEXT J

```

} Statements

REMARKS:

1. The loop execution is terminated when the value of control variable exceeds the end value.
2. If the loop control variable's initial value under some condition has exceeded the end value then loop is not executed.

7.1.1 Nested Loop:

One or more FOR.....NEXT loop can be inserted within another loop. Such loops are called nested loops. There can be several levels of nested loops in a specific program.

Rules:

1. Each loop within the nest must have its own FOR and NEXT statements. The loops must not overlap. The inner loops must lie within outer loops as shown in Fig. 1.

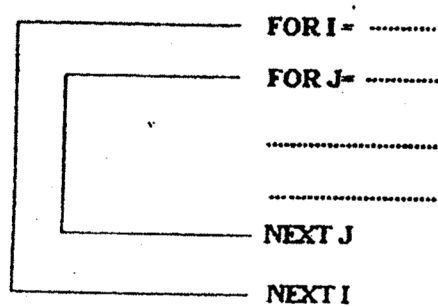
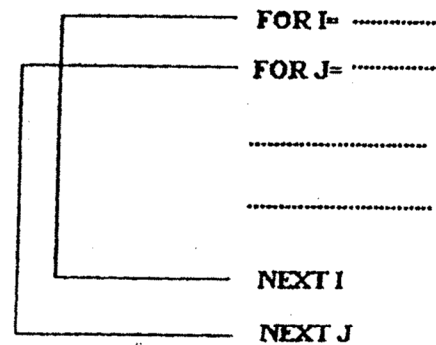


Fig. 1 a) Valid Nesting



b) Invalid Nesting

2. The outer loop and the inner loop must have different variables.
3. One can enter FOR.....NEXT loop only through FOR statement; however, branching out is possible as shown in Fig. 2.

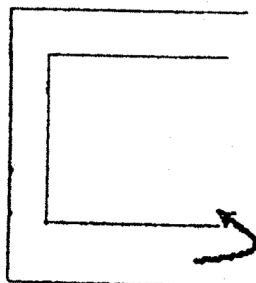
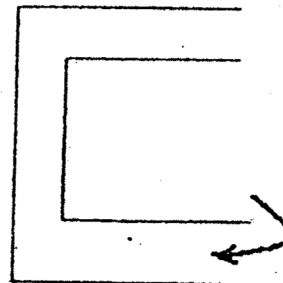


Fig. 2 a) Invalid Branching in



b) Valid Branching out

Program 9: A program for Using FORNEXT loop

```
10    REM printing natural numbers from 1 to 10
20    FOR I = 1 TO 10
30    PRINT I
40    NEXT I
50    END
```

.....
Output
.....

1
2
3
4
5
6
7
8
9
10

Program 10: Program for finding sum of cube of even-numbers between 1-50.

```
10    REM sum of cubes of even numbers between 1-50
20    S = 0
30    FOR I = 2 TO 50 STEP 2
40    S = S + I 3
50    NEXT I
60    PRINT "RESULT="; S
70    END
```

Program 11: A program for listing the characteristics of 5 flowers.

```
10    REM Listing the floral characteristics
20    REM ** No. of Sepals: S , .No. of Petals: P
30    REM ** No. of Carpels: C , Ovary type: OT$
40    REM ** Superior ovary: SUP , Inferior ovary: INF
50    PRINT TAB(5); "Flower No."; TAB(20); "Sepals"; TAB(31); "Petals";
    TAB(42); "Carpel"; TAB(55); "Ovary Type"
60    PRINT
70    FOR I = 1 TO 5
80    READ S, P, C, OT$
90    PRINT TAB(8); I; TAB(23); S; TAB(34); P; TAB(45); C; TAB(60); OT$
```

```

100 NEXT I
110 DATA 5,5,5, "SUP", 4,3,3, "INF", 6,8,6, "INF", 3,4,5, "SUP", 4,4,4,
    "SUP"
120 END

```

Output

Flower No.	Sepals	Petals	Carpels	Ovary Type
1	5	5	5	SUP
2	4	3	3	INF
3	6	8	6	INF
4	3	4	5	SUP
5	4	4	4	SUP

8.0. ARRAYS:

An orderly arrangement of data may be called an **array**. Arrays are used when many different data-items are required to share the same type of variable name. They are also useful in program segments when one requires to store some values or intermediate results for further use.

8.1. Subscripted Variables:

An ordered set of variables is called **subscripted variables**. For example, A(1), A(2), B(5), C(10) etc are subscripted variables. The numbers 1, 2, 5, 10 are **arguments**. The argument of a subscripted variable may be a variable expression but, when referred, the value of argument will be evaluated first and it should be an integer value. Set of subscripted variables, say, A(1), A(2), A(3),.....is said to be form an array. In fact this is a one dimensional array as there is a single argument. There are also two-dimensional or multi-dimensional arrays. Two-dimensional arrays are known as **matrices**. One-dimensional arrays are called **vectors**. For example, S1, S2, S3, S4, S5 may represent total sales of 5 salesman. In BASIC we shall write these variables as S(1), S(2), S(3), S(4), S(5). The use of arrays and subscripted variables helps us in processing large quantities of data.

REMARKS:

1. In case of subscripted variables. the subscripts may be a constant, variable or an expression.
2. A subscripted variable may consists of several letters and numbers.

Examples of subscripted variables

P(5)
K(I)
X(A+B)
A(P*Q)

8.2. DIM Statement:

In using subscripted variables in programs we have to declare the size of array to the computers. In other words, some memory locations should be reserved for the value of subscripted variable. This is attained by using DIM statement which is a short form of dimension. The general format of DIM statement is

Ln DIM variable (argument)

Where Ln is the line number

For example,

20 DIM A(100)

Set aside 100 storage locations for the variable A. That means that array A can contain up to 100 elements.

Rules:

1. The maximum value of the subscript should not exceed the value declared in DIM statement.
2. DIM statement can be placed any where in a program. However, it is good practice to place the DIM statement at the beginning of the program.

3. A single DIM statement is not necessary when the number of subscripted variables are less than equal to 11 for lists. However, it is good practice to use DIM statement in program which uses subscripted variables, whether it is necessary or not.
4. A single DIM statement may be used to declare the dimensions of several arrays.

Example:

```
10 DIM X(100), Y(200), Z(150)
```

9.0 LIBRARY FUNCTIONS

BASIC has a facility of some in-built functions. They are called library functions. These functions help the users to evaluate many complicated mathematical functions and carryout logical operations. The general format of library function is

Name (argument)

The followings are some of the important library functions used in BASIC:

ABS: This function retains the absolute value of a number. Absolute value is the actual value of a number without sign part.

ABS(X)

Where, X must be a numeric constant or a numeric variable or a numeric expression.

Example:

```
A=ABS(-9)
```

The value of A will be assigned as 9.

EXP: The function gives the exponential value of a given number X, i.e., e to the power X, where $e = 2.718282$ is the base of the natural logarithm.

The general format is -

EXP(X)

Where, X must be a numeric constant or a numeric variable or a numeric expression and must be less than 88.02969.

Example:

```
100 PRINT EXP(5) J
```

148.4132

OK

FIX: the function is used to set up the integer part of a given number. It does not care about decimal part of a give number, be it positive or negative. The FIX function just removes the decimal part and returns the integer portion of the number.

The general format is-

FIX(X)

Where, X has the significances as mentioned before.

Example:

```
20 PRINT FIX(3.8) J
```

3

OK

```
40 PRINT FIX(-3.8) J
```

3

OK

INT: This function gives the whole number part of a given number. In other words it gives the largest integer, not greater than the given number. The difference between INT and FIX comes when the given number, INT(-5.9) will give -6 whereas FIX(-5.9) will give -5.

The general format is -

INT(X)

Example:

```
5 PRINT INT(15.6) J
```

15

OK

```
10 PRINT INT(-5.8) J
```

-6

OK

```
130 PRINT INT(10/30) J
```

3

OK

LOG: This function returns the natural logarithm of any number X, where the X should be greater than 0.

The general format is-

LOG(X)

Example:

K= LOG(10)

The value of K will be 2.302585

SQR: This function is used for making square root of a given number. The general format is –

SQR(X)

Where X is a numeric value and must be greater than or equal to zero.

Example:

```
60 PRINT SQR(64) J
```

64

OK

Trigonometric Functions:

The following functions are used to get the results of trigonometric functions.

SIN, COS, TAN

The general formats are –

SIN(X), COS(X), TAN(X)

Example:

```
10 PRINT COS(45*(3.141593/180)) ↓
```

0.7071067

OK

10.0 SOME APPLICATION PROGRAMS IN BOTANY

Program 12: Suppose. you have measured the growth rate of different plants by an auxanometer. Write a program in BASIC to find the mean value of growth rate of a number of plants.

```
10 REM Finding mean growth rate of n number of plants
20 DIM GR (50)
30 S = 0
40 PRINT "Growth rate of plants:"
50 READ N
60 FOR I = 1 TO N
70 READ GR (I)
80 S = S + GR (I)
90 PRINT GR (I)
100 NEXT I
110 MGR = S/N
120 PRINT "Mean Growth Rate="; MGR
130 DATA .....
140 DATA .....
150 END
```

Program 13: You are given some data of length and diameter of some cylindrical logs. You have to write a program in BASIC for determining girths and volumes of n number of logs and listing them in tabular form.

```
10 REM Determination of girths and volumes of n logs
20 REM D: diameter, L: Length, G: Girth, V: volume
30 DIM D(100), L(100), G(100), V(100)
40 READ N
50 PRINT "No."; TAB(7); "DIAMETER"; TAB(20); "LENGTH"; TAB(31);
  "GIRTH"; TAB(41); "VOLUME"
60 PRINT TAB(9); "(cm)"; TAB(21); "(cm)"; TAB(32); "(cm)"; TAB(41);
  "(sq.m.)"
```

```

70   FOR I = 1 TO N
80   READ D(I), L(I)
90   G(I) = 3.1416 * D(I)
100  V(I) = 3.1416 * L(I) * ((D(I) ↑ 2)/4)
110  PRINT I; TAB(7); D(I); TAB(20); L(I); TAB(31); G(I); TAB(42); V(I)
120  NEXT I
130  DATA .....
140  DATA .....
150  END

```

Program 14: In an experiment root pressures of some plants have been measured. Write a program in BASIC to find out highest value of root pressure among n number of plants.

```

10   REM Finding highest value of root pressure
20   REM Root Pressure: RP; Highest root pressure: HPR
30   DIM RP(50)
40   READ N
50   READ RP(1)
60   HPR = RP(1)
70   PRINT "The values of Root Pressure are:"
75   PRINT RP(1)
80   FOR J = 2 TO N
90   READ RP(J)
100  PRINT RP(J)
110  IF HPR > RP(J) THEN 130
120  HPR = RP(J)
130  NEXT J
140  PRINT "Highest Root Pressure="; HPR; "mm Hg"
150  DATA .....
160  DATA .....
170  END

```

Note: The lowest value of root pressure (or any other data) can be computed by the same program with a slight alteration. The line number 110 should be altered like this-

```

110  IF HPR < RP(J) THEN 130

```

The variable names can be altered (say LPR in place of HPR) , if desired.

Program 15: In an experimental study the length of a number of seedlings were measured and they were divided into 3 groups: 2-5 cm, 6-9 cm and 10-15 cm. Now, write a program to determine the total count and percentage of seedling length in each group.

```

10  REM Finding Total count and percentage of seedling lengths of 3 groups
20  REM * * * seedling length: SL
30  DIM SL(200)
40  READ N
50  L1 = L2 = L3 = 0
60  PRINT "Measured Seedling Lengths:"
70  FOR K = 1 TO N
80  READ SL(K)
90  PRINT SL(K)
100 IF SL(K) > 15 THEN 190
110 IF SL(K) < 2 THEN 190
120 IF SL(K) >= 10 THEN 180
130 IF SL(K) <= 6 THEN 160
140 L1 = L1 + 1
150 GOTO 190
160 L2 = L2 + 1
170 GOTO 190
180 L3 = L3 + 1
190 NEXT K
200 PL1 = (L1/N) * 100
210 PL2 = (L2/N) * 100
220 PL3 = (L3/N) * 100
230 PRINT TAB(30); "SEEDLING LENGTH"
240 PRINT TAB(15); "....."
250 PRINT TAB(15); "2-5 cm"; TAB(26); "6-9 cm"; TAB(37); "10-15 cm"
260 PRINT: PRINT
270 PRINT "COUNT"; TAB(17); L1; TAB(28); L2; TAB(39); L3
280 PRINT "PERCENTAGE"; TAB(17); PL1; TAB(28); PL2; TAB(39); PL3
290 DATA .....
300 DATA .....
310 END

```

Output: The print out will be found in the following format:

		Measured seedling lengths:	
.....
.....

SEEDLING LENGTHS

	2-5 cm	6-9 cm	10-15 cm
COUNT
PERCENTAGE

Program 16: A program for computing **Standard Deviation** of mean of leaf lengths

```

10  REM computing standard deviation of mean of n number of leaf lengths
20  DIM L(100)
30  SL = 0
40  PRINT "Measured Leaf lengths are:-"
50  READ N
60  FOR I = 1 TO N
70  READ L(I)
80  PRINT L(I)
90  SL = SL + L(I)
100 NEXT I
110 ML = SL/N
120 V = 0
130 FOR I = 1 TO N
140 V = V + (M - L(I)) ^ 2
150 NEXT I
160 SD = SQR(V/N)
170 PRINT "Mean leaf length ="; ML; "cm"
180 PRINT "Standard Deviation ="; SD; "cm"
190 DATA .....
200 DATA .....
210 END

```

Program 17: Suppose, in a forest the heights of a particular group of trees have been measured. Write a program in BASIC to find out the mean height of the trees and to arrange the heights of the trees in ascending order.

```

10  REM finding mean height of trees and arranging them in ascending order
20  DIM HT(500)
30  READ N
40  PRINT "Height of the trees in Meter:-"
50  SHT = 0
60  FOR P = 1 TO N
70  READ HT(P)
80  PRINT HT(P)
90  SHT = SHT + HT(P)

```

```

100 NEXT P
110 MHT = SHT/N
120 PRINT "MEAN HEIGHT OF THE TREES="; MHT "Meter"
130 REM Arrangement of heights in ascending order
140 FOR K = 1 TO N-1
150 FOR M = 1 TO N-K
160 IF HT (M+1) >= HT(M) THEN 200
170 X = HT(M)
180 HT(M) = HT(M+1)
190 HT(M+1) = X
200 NEXT M
210 NEXT K
220 PRINT
230 PRINT "Heights of the Trees in Ascending order (in meter):"
240 FOR P = 1 TO N
250 PRINT HT(P);
260 NEXT P
270 DATA .....
280 DATA .....
290 END

```

Program 18: In a biochemical analysis percentage of sucrose has been determined in two types of fruits. In type A fruit n number of samples and in type B fruit m number of samples were taken. Write a program in BASIC to compute t-value for finding significance of difference in sucrose percentage between two groups of fruits.

```

10 REM * * computing t-value for two groups of sucrose percentage.
20 DIM PSA(100), PSB(100)
30 READ N, M
40 PRINT "Percentage of Sucrose in Type A fruit="
50 SA = 0
60 FOR I = 1 TO N
70 READ PSA(I)
80 PRINT PSA(I)
90 SA = SA + PSA(I)
100 NEXT I
110 MA = SA/N
120 PRINT "Percentage of Sucrose in Type B fruit="
130 SB = 0
140 FOR J = 1 TO M
150 READ PSB(J)
160 PRINT PSB(J)
170 SB = SB + PSB(J)
180 NEXT J

```

```

190 MB = SBM
200 VA = 0
210 FOR I = 1 TO N
220 VA = VA + (MA - PSA(I)) ↑ 2
230 NEXT I
240 VB = 0
250 FOR J = 1 TO M
260 VB = VB + (MB - PSB(J)) ↑ 2
270 NEXT J
280 SDA = SQR(VA/N)
290 SDB = SQR(VB/M)
300 X = MA-MB
310 P = (SDA↑2/N) + (SDB↑2/M)
320 T = X/SQR(P)
330 PRINT "Mean of Group A ="; MA
340 PRINT "S.D. of Group A ="; SDA
350 PRINT "Mean of Group B ="; MB
360 PRINT "S.D. of Group B ="; SDB
370 PRINT "t-value ="; T
380 DATA .....
390 DATA .....
400 END

```

Program 19: In a field survey, distribution of plants of different families was studied in an area. On the basis of field survey data, write a program in BASIC to find out frequency and percentage of distribution of plants belonged to the following families –
 (1) Solanaceae (2) Leguminosae (3) Gramineae (4) Labiatae

```

10 REM Finding percentage of distribution of 4 families
20 REM Solanaceae: SL; Leguminosae: LG; Gramineae: GR; Labiatae: LB
30 DIM SL$(500), LG$(500), GR$(500), LB$(500)
40 READ N
50 SL = LG = GR = LB = 0
60 FOR I = 1 TO N
70 READ F$(I)
80 IF F$(I) = "SL" THEN 130
90 IF F$(I) = "LG" THEN 150
100 IF F$(I) = "GR" THEN 170
110 IF F$(I) = "LB" THEN 190
120 GOTO 200
130 SL = SL + 1
140 GOTO 200
150 LG = LG + 1
160 GOTO 200

```

```

170 GR = GR + 1
180 GOTO 200
190 LB = LB + 1
200 NEXT I
210 REM Calculation of percentage
220 PSL = (SL/N) * 100
230 PLG = (LG/N) * 100
240 PGR = (GR/N) * 100
250 PLB = (LB/N) * 100
260 PRINT TAB(15); "SOLANACEAE"; TAB(30); "LEGUMINOSAE";
    TAB(45); "GRAMINEAE"; TAB(58); "LABIATAE"
270 PRINT : PRINT
280 PRINT "FREQUENCY": TAB(18); SL; TAB(33); LG; TAB(48); GR;
    TAB(60); LB
290 PRINT
300 PRINT "PERCENTAGE"; TAB(18); PSL; TAB(33); PLG; TAB(48);
    PGR; TAB(60); PLB
310 PRINT "....."
320 PRINT "TOTAL NUMBER OF PLANTS STUDIED ="; N
330 DATA .....
340 DATA .....
350 END

```

11.0. RUNNING BASIC PROGRAM IN COMPUTER:

The process of starting the computer is known as booting. To boot the computer the following step may be taken. The switch of the computer should be turned on and after some time DOS prompt will appear in VDU screen. In case of Windows system one can enter into DOS through the sub menus of 'All Program' options of the 'Start' menu. These steps may be varied from computer to computer. After entering into DOS the following will appear in the screen:

C:\>

Here C denotes the hard disk drive. The symbol is used to mean the blinking cursor. The presence of DOS prompt indicates that DOS invites you for using commands.

To get into BASIC environment a version of BASIC is required. There are many versions of BASIC, e.g., GWBASIC, QBASIC, BASICA etc. We shall discuss the operation of the version GWBASIC.

11.1. Entering into GWBASIC:

Type GWBASIC after DOS prompt as

```
C:\>GWBASIC ↵
```

And press enter key. The VDU will display

```
OK
```

Prompt on the screen instead of DOS prompt. Now you are in GWBASIC environment and can start typing your program.

From the program sheet the steps of the program may be typed and after each step enter key (↵) should be pressed. Care must be taken to include punctuation marks, spaces etc. during typing the program.

To save the time the line numbers can be generated by using AUTO command.

```
AUTO Ln ↵
```

Where Ln is the starting line number.

Example:

```
AUTO 10 ↵
```

Generates line number 10, 20, 30, and so on.

The arrow symbol ↵ denotes the enter key.

11.2. Editing the program:

A program may have a number of mistakes. In order to correct them EDIT command is used

```
EDIT Ln ↵
```

Where Ln is a line number to be edited.

Example:

EDIT 50 J

The statement having the line number 50 will appear under the command. The statement can be altered or corrections can be made by shifting the cursor in the desired position. The enter key should be pressed after this.

11.3. Listing the program:

LIST command is used to list all or part of the program on the screen. The LIST command can be typed from the key board or the function Key (F1) may be pressed to get the LIST command on the screen.

LIST Ln - Lm

Where Lm and Ln are line numbers.

Example:

Command	Functions
LIST	Lists the entire program on the screen
LIST 60	Lists the lines starting from 60 and onwards
LIST 10-200	Lists lines from 10 to 200

11.4. Renumbering Lines:

The lines of the program can be renumbered by using RENUM command.

Example:

RENUM
Renumbers the program starting with line number 10 and increment 10

RENUM 100, 90, 5
Replaces the line number 90 by 100 and subsequent line numbers are increased by 5.

11.5. Running the program:

To run the program RUN command is used. The command can be typed from the keyboard or the function key 2 (F2) may be pressed to get the word RUN on the screen and the result of the program (output) will be displayed.

11.6. Saving the program:

The SAVE command saves the program in the memory (for further use) under the specific name. The SAVE command can be typed from the key board or the function Key (F4) may be pressed to get the command displayed on the screen. The file name should be within the quotation marks. However, closing quotation mark is optional.

SAVE "File Name ↵

Example:

SAVE "PS2 ↵

The program will be saved under the file name "PS2".

11.7. Starting a new program:

After saving the program, if another new program is to be started the command NEW should be used.

NEW ↵

The OK prompt will appear on the screen.

11.8. Loading a program:

The LOAD command allows us to load a previously saved program in computer memory. The LOAD command can be typed from the key board or the function Key (F3) may be pressed to get the LIST command displayed on the screen.

LOAD "File Name

In the first step, function key no.3 (F3) is pressed and in second step, file name is typed after the command LOAD.

Example:

LOAD "PS2 J

This will load the program which is stored under the file name PS2. The OK prompt will appear on the screen. After this, if F1 key is pressed the list of the program will be displayed.

11.9. Printing a program and output:

The LIST and PRINT commands are used to list the program and print the output on the screen. On the other hand the hard copy of the output may also be taken from the printer. The following commands are used to get the same in a printer.

LLIST J

The printer will print the whole program and printed copy of the program will be obtained.

To get a printed copy of the output (results) all the PRINT statements should be replaced by LPRINT. After preparing the program with LPRINT statement instead of PRINT when the program is run (by pressing F2 key) a printed output will be obtained from the printer.

11.10. Deleting the statement

Sometimes one or more statements of a program are required to be deleted. The DELETE command is used for this purpose. To delete the statement number should be given after the command.

DELETE 70 J

The statement number 70 will be deleted from the program

DELETE 70 - 100 ↵

All the statements from 70 to 100 will be deleted from the program.

11.11 Quitting the BASIC:

After completing your work with BASIC, you should quit the GWBASIC environment and return to MS-DOS. To do this following command is used-

SYSTEM ↵

The display will show

C:/>

Which confirms DOS command mode.

12.0 DIAGNOSIS OF PROGRAMMING ERRORS

Programming errors often remain undetected until an attempt is made to execute the program. Once the RUN command (pressing F2 key) has been given, the presence of certain errors will become readily apparent, since such errors will prevent the program from being interpreted, i.e., transformed into a machine language program. Some particularly common errors of this type are a reference to an undefined variable or an undefined statement number, right- and left-hand parentheses that do not balance failure to terminate the program with an END statement, etc. Such errors are called grammatical or syntactical errors. The BASIC will generate a diagnostic message when grammatical error has been detected. They are helpful in identifying the nature and location of the error.

Grammatical and typing errors are usually very obvious when they occur. Much more insidious are logical errors. Here the program correctly conveys the programmer's instructions, free of grammatical and typing errors, but the programmer has supplied the computer with a logically incorrect set of instructions.

Sometimes a logical error will result in a condition that can be recognized by the computer. Such a situation might result from the generation of excessively large quantity or for an attempt to compute the square root of a negative number, etc. Diagnostic messages will be generated in situations of this type, making it easy to identify and correct the errors. These diagnostic messages are called execution diagnostics to distinguish them from the interpretation diagnostics.

Logical Debugging: The first step in attacking logical errors to find out if they are present. It can sometimes be accomplished by testing a new program with data that will yield a new answer. If the correct results are not obtained, then the program obviously contains errors. Even if the correct results are obtained, one cannot absolutely certain that the program is error free, since some errors cause incorrect results only under certain circumstances (as, for example, with certain values of the input data or with certain program options). Therefore a new program should receive thorough testing before it is considered to be debugged. This is especially true for complicated programs or programs that will be used extensively by others.

As a rule, a calculation will have to be carried out by hand, with the aid of a calculator, in order to obtain a known answer. For some problems, however, the amount of work in carrying out a hand calculation is prohibitive. (A problem requiring a few seconds of time on a large computer may require several weeks to solve by hand.). Therefore a sample problem cannot always be developed to test a new program. Though logical debugging of such programs can be particularly difficult, the programmer can often detect logical errors by studying the computed results carefully to see if they are reasonable.

Correcting Errors: Error detection should always begin with the programmer carefully reviewing each logical group of statements in the program. Armed with the knowledge that an error exists somewhere, the programmer can often spot the error by such careful study. If the error cannot be found, it sometimes helps to set the program aside for a while. It is not unusual for an overly intent programmer to miss an obvious error for the first time around.

If an error has not been found after repeated inspection of the program, then the programmer should proceed to rerun the program, printing out a large quantity of intermediate output. This is referred to as tracing. Often the source of error will become evident once the intermediate calculations have been carefully examined.

13.0. SUMMARY

BASIC (Beginners All Purpose Symbolic Instruction Code) is high level language. This is a simple to use computer language to write a program. It recognizes three categories of character sets- i) alphabets – A, B, ..., Z and a, b, ..., z ii) digits – 0, 1, 2, ..., 9 and iii) special characters like, @ # \$ % & * (), > < etc. For writing program in BASIC variables and constants are used. The BASIC constants are of two types – numeric and string constants. The numeric constant deals the numbers, e.g., 2345, -203, 0.098 etc. It can also be represented in exponent form, viz., 123E-04. The string constant represents non-numeric characters like, name, address, etc. It is usually kept under the quotation marks. The examples of the string constants are- "algae", "University", "234", etc. The BASIC variables are also divided into numeric and string variables. The numeric variables are used to store numeric constants in the memory. The numeric variable name is given by using alphabets and digits; the first character of the variable name should be an alphabet. For example – A, XY, B3, Flower1, M15K, etc. The string variable represents the string constants. The nomenclature of the string variable is done in the same way as done in numeric variable but a dollar sign is followed by the numeric variable name. Example- A\$, d25\$, Genus\$, etc. In BASIC programming both arithmetic expression (e.g., $A+B^2$) and logical expression (e.g., $X>Y$) are used.

The format of the BASIC program is a line number followed by BASIC statement. The line number is given in ascending order. The BASIC statements are consisted of key word(s) having special significance. The common keywords are LET, INPUT, PRINT, READ, DATA, END, REM etc. The LET statement is used for assigning values (e.g., 50 LET A=4). The INPUT statement is used for providing data to the computer during execution of the program. The READ and DATA statement are also used for the same purpose. However, data are given in the DATA statement before the execution of the

program. The PRINT statement is used for getting output of the program in the monitor. The END statement declares the physical end of the program. The REM is a non-executable statement, which is used for putting the comments of the users. With the help of those statements simple programs can be framed. Some of the examples are – solving small equations, conversion of temperature from Fahrenheit to Celsius, preparation of student's results, etc. In those programs the flow of control moves in ascending direction. However, branching of the flow may be caused where the flow is moved in both ascending and descending directions with skipping some steps. The branching may of two types- unconditional and conditional. For unconditional branching GOTO statement (e.g., 70 GOTO 30) is used. For conditional branching IF ... THEN statement (60 IF GR> 15 THEN 110) can be used. By those statements some programs like, finding larger between two numbers or largest among three numbers, conversion of length from inch to centimeter of a number of leafs, computing sum of numbers etc. may be written. When some steps of the program are executed repeatedly for more than one times, the process is called looping. FOR ... NEXT statements are used for this purpose. The FOR statement (e.g., 80 FOR I = 1 to 10) is the beginning statement and the NEXT statement (e.g., 150 NEXT I) is the closing statement of the loop. Sometimes, one FOR ... NEXT loop may contain another loop. This is known as nested FOR NEXT loop. Using this loop printing of natural numbers, floral characteristics of a number of flowers can be listed. An orderly arrangement of data may be called an array. Arrays are used when many different data-items are required to share the same type of variable name. An ordered set of variables is called subscripted variables. For example, A(1), A(2), B(5), C(10) etc are subscripted variables. When the subscripted variable is used in the program a DIM statement should be used. This statement (e.g., DIM A (100), X5(50), F\$(200)) is employed for assigning space in the memory location. Using all the above statements a number of application programs can be written. For example, programs for computing the mean value of leaf length, determining the girth and volume of a number of cylindrical logs and their listing, finding highest value of root pressure, arranging the tree length in ascending order, standard deviation of mean of leaf lengths, t-test for finding the significance of difference between the sucrose% of two groups fruits, listing the percentage distribution of plants in different families etc. may be written.

For running the BASIC program in the computer several commands, viz., LIST, RUN, LOAD, SAVE, EDIT, DELETE etc. are used. To run the program perfectly, correction of errors in the program is needed. This is called debugging. The syntactical errors can be corrected from the error list given during execution of program. However, the correction of logical should be made by careful examination of each of the steps in the program.

14. BIBLIOGRAPHY:

1. Dhara P. C.: Computer in Biological Sciences. Academic Publisher, Kolkata.
2. Balagurusamy E.: Programming in BASIC. Tata McGraw-Hill Publishing Co. Ltd., New Delhi.
3. Sampath S. and Wasan S.K.: BASIC programming. Macmillan India Ltd., Bangalore.
4. Rajaraman V.: Fundamentals of computers. Printice-Hall of India Pvt. Ltd., New Delhi.
5. Lotia M., Nair P. and Lotia P.: Modern all about GW-BASIC. BPB Publication, New Delhi.
6. Gottfried B.S.: Programming with BASIC. Tata McGraw-Hill Publishing Co. Ltd., New Delhi.

15.0. MODEL QUESTIONS:

15.1 Long Questions:

1. What do you mean by READ, DATA Statements of BASIC? How do they differ from INPUT statement. Explain with example.
2. What is FOR – NEXT loop? What do you understand by nested FOR – NEXT loop. Illustrate with examples.
3. What is the function of PRINT statement? Discuss the rules of PRINT statement. What is TAB function?
4. What is subscripted variable? Write a program for finding the highest value of transpiration rate of different leaves.
5. Write a program in BASIC to compute stomatal index of 'n' number of leaves of a plant.

M.Sc. in Botany

Part-I :: Paper -II (Second Half)

Module No. 19

M.Sc. Botany
Part - I Paper - II (2nd half)
Module No. - 19
BIOMETRY AND STATISTICS

What is Statistics ?

Statistics is a field of study concerned with

- 1) the collection, organisation and summarisation of data, and,
- 2) the drawing of inferences about a body of data when only a part of the data are observed.

The concepts and methods necessary for achieving the first objective are presented under the heading of Descriptive Statistics and the second objective is reached by the study of Inferential Statistics.

The word statistics has multiple meanings :

- 1) Statistics is synonymous with numerical facts or data,
- 2) Statistics is the science dealing with the designing of experiments leading to the collection of numerical facts and method of analyzing, interpreting and presenting these numerical facts.

Characteristics and limitations of statistics :

Characteristics :

1. In statistics, all available informations are to be expressed in quantitative terms. Even in the study of a quality like intelligence of a group of students we require scores or marks secured in a test.
2. Statistics deals with a collection of facts, not an individual happening.
3. Statistical data are collected with a definite object in view i.e. there must be a definite field of enquiry.
4. Statistics are affected by a multiplicity of causes. In all fields or enquiry, the observed data are the result of a large number of factors, each of which contributes to the final figure.
5. Statistics is not an exact science. Conclusions are usually derived from samples and hence exactness can not be guaranteed.
6. Statistics should be capable of being related to each other, so that some cause and effect relationship can be established.
7. A statistical enquiry passes through the following four stages :

Collection of data



Classification and tabulation of the data



Analysis of the data



Interpretation of the data

LIMITATIONS :

1. Statistics is applicable only to quantitative data and it is not suited to the study of qualitative phenomenon.
2. Statistics can be used only to analyze an aggregate of objects, and not individual objects.
3. Statistical decisions are true only on an average and also the average is to be taken for a large number of observations. They may not be true for a few cases.
4. Statistical decisions are to be made carefully by the experts. The use of statistical tools by untrained persons may lead to false conclusions. Misuse of statistics has, in fact, created some distrust on the subject.
5. Statistical data must be uniform, in the sense that they should be subject to a stable casual system. There should not be change in the group of factors responsible for variation in the data.

BIOSTATISTICS OR BIOMETRY :

It may be defined as the application of statistics as a science to Biology. The use of statistical methods is constantly increasing in biological sciences. The development of biological theories is closely associated with statistical methods. Therefore, a good understanding of biostatistics is essential for the students of biological sciences, as the methods of Biostatistics are indispensable tools for the design and analysis of data in the interpretation of experimental results for dependable conclusions.

DEFINITION OF FEW TERMS COMMONLY USED IN STATISTICS :

Population :

It refers to the complete set of observations or measurements about which inferences are to be made at a particular time. It simply means the totality of the set of objects under consideration. For example, all babies born in a particular year, all plants in a wheat field may constitute a population. Population may be finite or infinite. If a population consists of a fixed number of observations, the population is said to be finite. For example, the number of patients in a hospital, the number of wheat plants in a quadrant etc. On the other hand, a population which is unlimited in size is said to be an infinite population. Thus for an infinite population it is impossible to observe all the values. For example, the number of RBCs in the human body, the number of phytoplanktons in a pond etc.

Sample :

It is simply defined as a part of a population i.e. a sample is a selected number of individuals each of which is a member of the population. Thus, a sample represents the small collection of the population which has actually been observed. For example, all plants in a wheat field represent a population whereas individual observations on 10 high yielding plants from this population is a sample.

Data :

Numbers or measurements that are collected as a result of observations, known as data. For example, scores of a psychological or educational test.

Variables :

Any quality or quantity liable to show variation from one individual to the other in the same population is known as variable. An individual observation of any variable is known as variate. For example, plant height, the weights of pre school children etc.

Constant :

The value of which never changes is known as constant. For example, π (pi).

Quantitative variables :

Variables that can be measured in usual sense i.e. can be expressed in terms of numbers are said to be quantitative variables. For example, yield of crops, number of spikelets / spike, measurement on the height of adult male's etc.

Qualitative variables :

Variables that can not be measured in usual sense i.e. cannot be expressed in terms of numbers, but can be classified under different heads or categories are said to be qualitative variables. For examples, religion, sex, colour etc.

Continuous variable :

A variable is said to be continuous when it can assume any value within a specified interval of values assumed by the variable. For example, height, weight etc.

Discrete or discontinuous variables :

A discrete variable is a variable, which can take only some isolated values. For example, the number of plants in a given quadrant, the number of births in a hospital etc.

Random variable :

A variable whose values depends on chance and can not be predicted is called a random variable. Statistical theories deals with random variables only.

Primary and Secondary data :

The most important task of a statistician is to collect and assemble his data. He may prepare the data himself (primary data) or borrow them from other sources (secondary data).

Primary data :

These are collected for a specific purpose directly from the field of enquiry. Thus, these data are original in nature. Generally, trade associations collect data from their member concerns, government organisations collect data from its subordinate offices. They are considered as primary data. But individuals or any organisation can also collect primary data from the actual field of enquiry by appointing trained investigators.

Sources for collecting primary data :

An investigator may collect primary data by the following methods:

Direct personal observation :

In this method, the investigator may collect data by direct observation or measurement. An investigator may himself meet persons who can supply the requisite information. The method is time consuming and costly, but yields very accurate results. It is therefore suitable for such studies when the field of enquiry is small.

Indirect oral investigation :

In this method data are collected through indirect sources. Persons, who are likely to have information about the problem, are interrogated and on the basis of their answers, the data have to be compiled. Most of the commissions of enquiry or committees appointed by Government collect primary data by this method. The reliability of this data very much depends on the integrity of the persons selected.

Sending questionnaires by mail :

In this method the most important instrument is the questionnaire. This contains a set of questions relevant to the subject of enquiry and these questions are sent by mail to the selected persons with request to return them duly filled in. Though this method can cover large areas and is also comparatively cheaper, but the principal disadvantages of the method are – the low degree of reliability of the collected data, and a large number of non-responses.

Sending schedules through investigators :

This method is most widely used for the collection of primary data. In this method paid investigators are employed for the data collection. The investigators carry with them the printed schedules and meet the person concerned. The investigators fill up the schedules on the spot based on the answer received from the informants. The data are thus collected. This method is very popular and yields satisfactory results.

TWO PROCEDURES FOR COLLECTION OF DATA :

Complete enumeration or Census :

When information is collected in respect of every individual person or item of a given population, we say that the enquiry has been done on complete enumeration or census. This process is called census survey.

Sample survey :

In most cases of statistical enquiry, because of limitations of time and money, only a portion of the population is examined and the data collected therefrom. This process of partial enumeration is known as Sample survey. The results and findings are, however, made applicable to the whole field of enquiry. This is known as Sampling.

Secondary data :

These are numerical information, which have been previously collected as primary data by some agency for a specific purpose but now are compiled from that source for use in a different connection. In fact, data collected by some agency when used by another or collected for one purpose when used for another may be termed as secondary data. The same data is primary for its collecting authority but secondary to another agency who used them.

The chief sources of secondary data are :

- ♦ Official publications of State and Central Governments, Foreign Governments and International bodies like ILO, UNO, UNESCO, WHO etc.
- ♦ Publications and reports of various Chambers of commerce, Trade associations, Co-operative societies etc.
- ♦ Reports of commissions and committees of enquiry appointed from time to time for specific purposes of enquiry.
- ♦ Journals and Magazines published by private agencies.

Unpublished reports prepared by researchers, labours and trade unions.

Presentation of data :

There are three different ways in which statistical data may be presented.

Textual presentation :

In this method numerical data are presented in descriptive form. Example : Numerical data with regard to industrial diseases and deaths therefrom in Great Britain during 1935-39 and 1940-44 are as follows :

During 1935-39, there were in Great Britain 1775 cases of industrial diseases made up of 677 cases of lead poisoning, 111 of other poisoning, 144 of anthrax and 843 of gassing. The number of deaths reported was 20% of the cases for all the four disease taken together, that for lead poisoning was 135, for other poisoning 25, and that for anthrax was 30.

During 1940-44, the total number of cases reported was 2807. But lead poisoning case reported fell by 351 and anthrax cases by 35. Other poisoning cases increased by 784 between the two periods. The number of deaths reported decreased only by 2 for anthrax from the prewar to the postwar. In the later period, 52 deaths were reported for poisoning other than lead poisoning. The total number of deaths reported in 1940-44 including those from gassing was 64 greater than in 1935-39.

Disadvantages are :

1. It is a lengthy text.
2. There has been much repetition in words.
3. Comparison between the corresponding figures in the two periods is difficult.
4. It is difficult to grasp, from a lengthy text the important points if there be a number of them, making it all the more difficult to arrive any conclusion.

Tabular presentation :

Tabulation may be defined as the logical and systematic organisation of statistical data in rows and columns., designed to simplify the presentation and facilitate comparisons.

Tabular representation is also a form of presentation of quantitative data in a condensed form so that numerical figures and easy to understand.

The numerical descriptions of the previous example have been condensed in the form of table given below :

Example :

Tabel 1 : Deaths from industrial diseases in Great Britain

SL No.	Types of Diseases	1935 – 39			1940 – 44		
		No. of cases	No. of Deaths	% of Deaths	No. of cases	No. of Deaths	% of Deaths
1	2	3	4	5	6	7	8
1.	Lead poisoning	677	135	19.9	326	90	8
2.	Other poisoning	111	25	22.5	859	52	6.1
3.	Anthrax	144	30	20.8	109	28	25.7
4.	Gassing	843	165	19.6	1513	249	16.5
	Total	1775	355	20.0	2807	419	14.9

Advantages of tabulation are :

1. It enables the significance of data readily understood, and leaves a lasting impression than textual presentation.
2. It facilitates quick comparison of statistical data shown between rows and columns.
3. Errors and omissions can be readily detected when data are tabulated.
4. Repetitions of explanatory terms can be avoided, and the concise tabular form clearly reveals the characteristics of data.

RAW DATA :

Statistical Data may originally appear in a form where the collected data are not organised numerically. We call them Raw data.

Suppose that the number of grains per spike of a wheat variety 'Bansi' are as shown below :

Tabel 2 : Number of grains per spike in wheat

37	38	40	36	38	37	36	40	50	47
41	46	38	31	33	48	37	52	32	50
40	50	47	41	50	43	26	45	52	45
41	44	39	16	21	30	38	32	48	47
41	45	41	51	37	26	40	38	46	32

Tabel 3 : Array of data of table 2 (arranged in order ascending magnitudes)

16	21	26	26	30	31	32	32	32	33
36	36	37	37	37	37	38	38	38	38
38	39	40	40	40	40	41	41	41	41
41	43	44	45	45	45	46	46	47	47
47	48	48	50	50	50	50	51	52	52

The variable observed here are the number of grains per spike in wheat and the data obtained are known as observations. If n observations on a variable x are available, they are usually denoted by

$x_1, x_2, x_3, \dots, x_n$.

where x_1 denotes the first observation on x ,

x_2 denotes the second observation on x ,

x_3 denotes the third observation on x ,

|
|

x_n denotes the n -th observation on x .

In general x_i denotes the i th observation on x ($i = 1, 2, \dots, n$). In Table 2

$x_1 = 37, x_2 = 38, x_3 = 40, \dots, x_{30} = 32$. All the observation are not different, some of them are repeated.

Frequency :

Frequency of a value of the variable is the number of times it occurs in a given series of observation. In table 3, 16 occurs once, 21 occurs once, 26 occurs 2 times, 30 occurs once, 31 occurs once, 32 occurs 3 times etc. Hence, the frequencies of the values 16, 21, 26, 30, 31, 32 are 1, 1, 2, 1, 1, 3 respectively.

Tally sheet :

A tally sheet may be used to calculate the frequencies from the raw data. A tally mark (/) is put against the value when it occurs in the raw data. Having occurred 4 times, the fifth occurrence is represented by putting a cross tally mark (\) on the first four tally marks. This technique facilitates the counting of tally marks at the end.

Tabel 4 : Tally marks

No. of grains per spike (x)	Tally marks	Frequencies (f)
16	/	1
21	/	1
26	//	2
30	/	1
31	/	1
32	///	3
33	/	1
36	//	2
37	////	4
38	////\	5
39	/	1
40	////	4
41	////\	5
43	/	1
44	/	1
45	///	3
46	//	2
47	///	3
48	//	2
50	////	4
51	/	1
52	//	2
Total frequency		50

Such a representation of the data is known as the frequency distribution.

Frequency distribution is a statistical table which shows the values of the variable arranged in order of magnitude, either individually or in groups, and also corresponding frequencies side by side. There are two types of frequency distributions :

(1) Simple frequency distribution :

It shows the values of the variable individually (see Table 4).

(2) Grouped frequency distribution :

It shows the values of the variable in groups or intervals (see Table 5).

Tabel 5 : Grouped frequency distribution

No. of grains per spike	Frequencies (f)
16 – 20	1
21 – 25	1
26 – 30	3
31 – 35	5
36 – 40	16
41 – 45	10
46 – 50	11
51 – 55	3
Total	50

In constructing a frequency distribution, it is desirable to consider the following basic rules :

1. Range :

Range of a given data is the difference between the highest and lowest value of the variable. In Table 3, the highest value is 52 and the lowest is 16, hence the range = $52 - 16 = 36$.

2. Number of classes :

As a general rule, the number of classes should lie between 10 – 15, never more than 30 and not less than 6. However, the number of classes in a frequency distribution are not fixed.

3. Class intervals :

The class intervals depend on the range of the data and the numbers of classes. The class interval would be equal to the difference between the highest and the lowest value of the variable divided by the number of classes. The following formula may be used to estimate the class interval :

$$i = \frac{L - S}{C},$$

where i = class interval

L = largest value

S = smallest value

C = No. of classes.

The number of classes C can be decided with the help of the sturge rule. According to sturge, the class interval

$$C = 1 + 3.322 \log N$$

Where N = total number of observations

Log = logarithm.

Class limits :

The class limits are the lowest and the highest values which are included in the class. For example, in the class 41 – 45, the lowest value is 41 and the highest value is 45. It indicates that there can be no values in the class below the lower limit 41 and above the upper limit 45 of the class.

Class boundaries or true class limits :

The most extreme values which would ever be included in a class interval are called class boundaries or true class limits. If age is considered, the class interval 15 – 19 actually includes all ages between 14.5 and 19.5. Thus, class boundaries are, in fact, the real limits of a class interval. In this case the lower extreme point 14.5 is the lower class boundary and the upper extreme point is the upper class boundary. Class boundaries may be calculated from class limits by applying the following rule :

$$\text{Lower class boundary} = \text{lower class limit} - \frac{1}{2} d$$

$$\text{Upper class boundary} = \text{upper class limit} - \frac{1}{2} d$$

where d is the common difference between the upper class limit of any class interval and the lower class limit of the next class interval.

Mid – value :

The value exactly at the middle of a class interval is called the Mid – value.

$$\text{Mid – value} = (\text{lower class limit} + \text{upper class limit}) / 2$$

$$= (\text{lower class boundary} + \text{upper class boundary}) / 2.$$

There are two ways of classifying the data on the basis of class intervals :

Exclusive method :

In this method, the upper limit of a class is the lower limit of succeeding class. The following data are classified on this basis :

No. of grains per spike	Number of plants
16 but less than 21	1
21 but less than 26	1
26 but less than 31	3
31 but less than 36	5
36 but less than 41	16
41 but less than 46	10
46 but less than 51	11
51 but less than 56	3
Total	50

Inclusive method :

In this method the upper limit of one class is included in that class. The following data are classified on this basis:

No. of grains per spike	Number of plants
16 – 20	1
21 – 25	1
26 – 30	3
31 – 35	5
36 – 40	16
41 – 45	10
46 – 50	11
51 – 55	3
Total	50

Width of a class :

It is the difference between the lower and the upper class boundaries (not class limits). Width of a class = upper class boundary – lower class boundary.

Cumulative frequency and cumulative frequency distribution :

In statistical investigations, sometimes we are interested in the number of observations less than (or more than) a given value. In such cases our main objective is concerned with the accumulated frequency upto (or above) some value of the variable. This accumulated frequency is known as the cumulative frequency. Thus, cumulative frequency corresponding to a specified value of the variable may be defined as the number of observations less than (or more than) that specified value. The number of observations upto a given value is known as the cumulative frequency less than type, and the number of observations more than a given value is called the cumulative frequency more than type. When a frequency distribution is given, the cumulative totals of frequencies give the cumulative frequencies. When a grouped frequency distribution is given, the cumulative frequencies calculated there from must be shown against the class boundary points. A table showing the cumulative frequencies against values of the variable arranged in order of magnitude either in ascending (or in descending) order is known as the cumulative frequency distribution.

Example 1 : Construct (a) the cumulative frequency distribution, and (b) the grouped frequency distribution, from the following data :

<u>Number of grains</u>	<u>Number of plants</u>
less than 10	37
less than 20	118
less than 30	161
less than 40	185
less than 50	194
less than 60	200

Solution :

(a) The cumulative frequency distribution shows the values of the variable and the corresponding cumulative frequencies (less than).

Table 6: Cumulative frequency distribution

<u>Number of grains</u> (less than)	<u>Number of plants</u>
10	37
20	118
30	161
40	185
50	194
60	200

(b) The grouped frequency distribution shows the values of the variable in class intervals and the corresponding class frequencies:

Table 7: Grouped frequency distribution

<u>Number of grains</u> (less than)	<u>Number of plants</u>
0 - 10	37
10 - 20	81
20 - 30	43
30 - 40	24
40 - 50	9
50 - 60	6

Total number of plants = 200

Exmple 2 :

Construct the cumulative frequency distribution less than and more than type of the number of grains per spike from the Table 7.

Solution :

Cumulative frequency distribution.

<u>Number of grains</u> (less than)	<u>Number of plants</u>	<u>Number of grains</u> (more than)	<u>Number of plants</u>
10	37	0	200
20	$37 + 81 = 118$	10	$200 - 37 = 163$
30	$118 + 43 = 161$	20	$200 - 118 = 82$
40	$161 + 24 = 185$	30	$200 - 161 = 39$
50	$185 + 9 = 194$	40	$200 - 185 = 15$
60	$194 + 6 = 200$	50	$200 - 194 = 6$

Histogram :

It is the common form of diagrammatic representation of a group frequency distribution. It consists of a set of adjoining rectangles drawn on a horizontal line, with areas proportional to the class frequencies. The width of rectangles, extends over the class boundaries (not class limits) shown on the horizontal line. The histogram can be constructed in two ways depending upon the class-intervals :

- (1) For distribution that have equal class-intervals
- (2) For distributions that have unequal class-intervals

When the class-intervals are equal, the height of the rectangles will be proportional to the class frequency.

Example 3 :

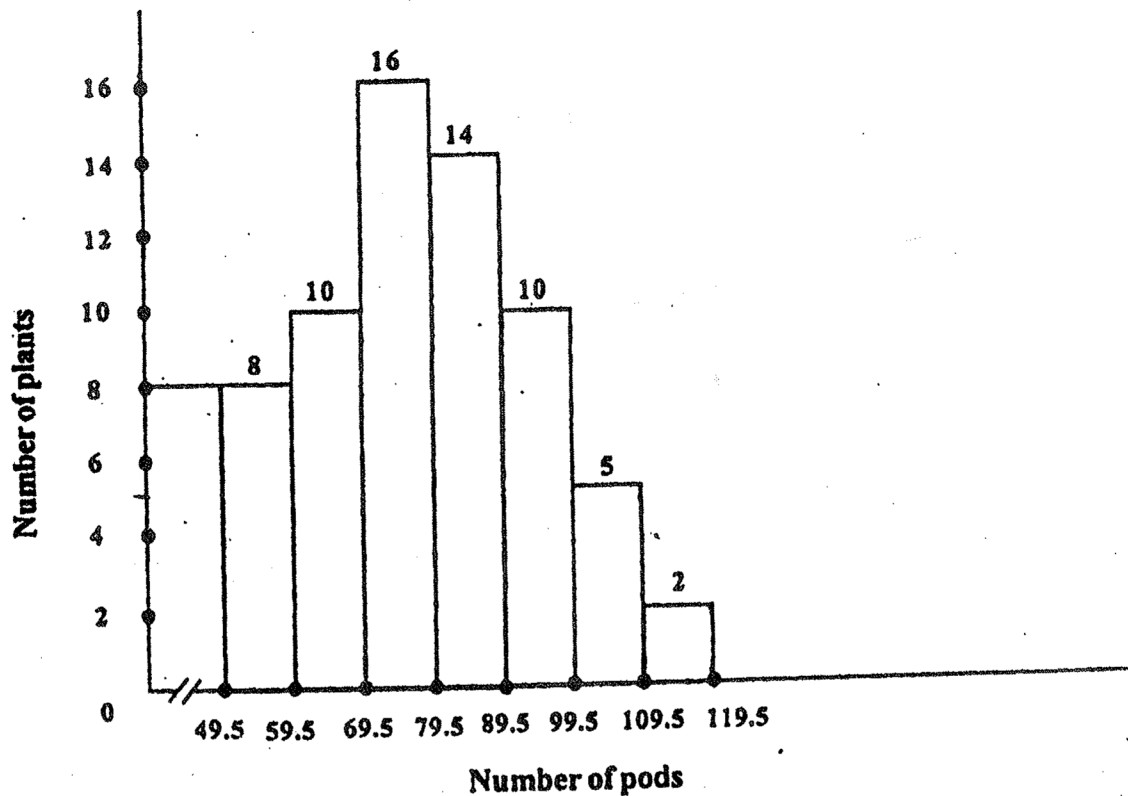
Draw a histogram for the following frequency distribution :

<u>Number of pods</u>	<u>Number of plants</u>
50 - 59	8
60 - 69	10
70 - 79	16
80 - 89	14
90 - 99	10
100 - 109	5
110 - 119	2

Solution:

Here the class intervals are defined by class limits and so we have to find the class boundaries for drawing the histogram. All the classes have the same width and therefore when histogram is drawn, heights of the rectangles may be represented by the class frequencies.

Number of pods (class limits)	Class boundaries	Number of plants (frequency)
50 - 59	49.5 - 59.5	8
60 - 69	59.5 - 69.5	10
70 - 79	69.5 - 79.5	16
80 - 89	79.5 - 89.5	14
90 - 99	89.5 - 99.5	10
100 - 109	99.5 - 109.5	5
110 - 119	109.5 - 119.5	2



Histogram representing the distribution of the number of pods per plants as given in Example 3.

Example 4 : Draw a histogram for the following frequency distribution :

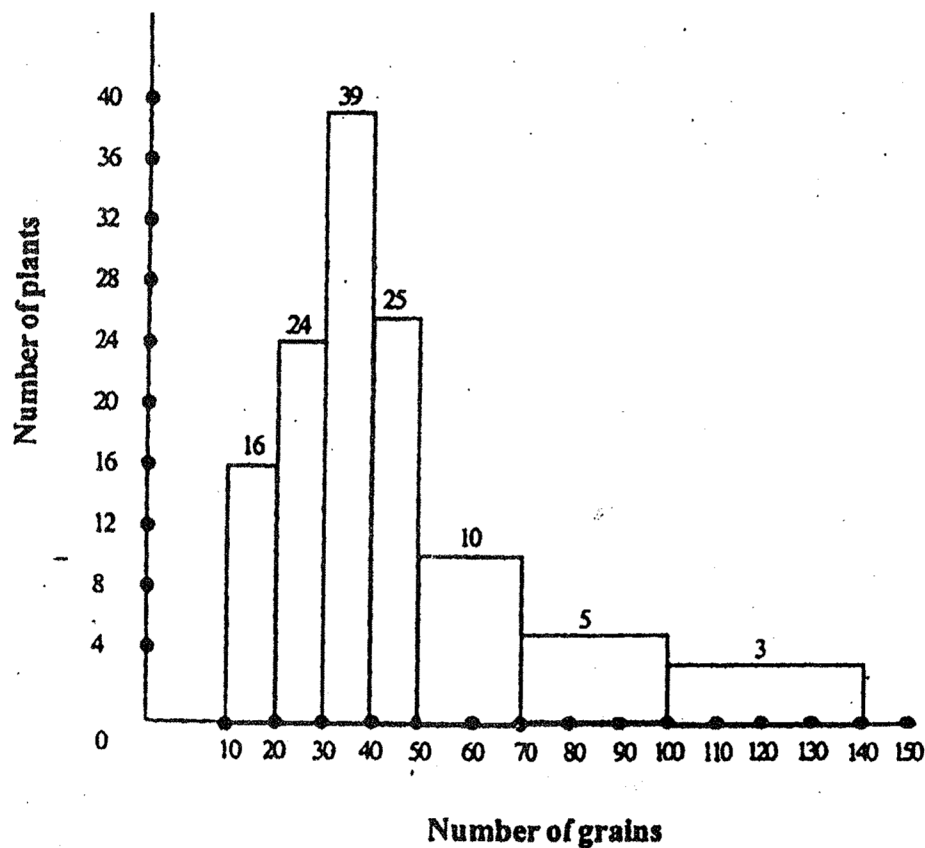
<u>Number of grains</u>	<u>Number of plants</u>
(less than)	16
10 - 20	24
20 - 30	39
30 - 40	25
40 - 50	20
50 - 70	20
70 - 110	12
110 - 150	

Solution :

Since the class intervals are unequal, frequency density has to be calculated. Taking the lowest class intervals 10 - 20, 20 - 30, 30 - 40, 40 - 50 etc. as 1 unit, the class interval 50 - 70 becomes 2 units, 70 - 110 and 110 - 150 becomes 4 units etc.

Thus, histogram can be drawn on the basis of the following data :

Number of grains	Units	Number of plants	Frequency Density
10 – 20	1	16	$16 \div 1 = 16$
20 – 30	1	24	$24 \div 1 = 24$
30 – 40	1	39	$39 \div 1 = 39$
40 – 50	1	25	$25 \div 1 = 25$
50 – 70	2	20	$20 \div 2 = 10$
70 – 110	4	20	$20 \div 4 = 5$
110 – 150	4	12	$12 \div 4 = 3$



Histogram representing the distribution of the number of grains per spike as given in Example 4.

Frequency polygon :

It is a graphical representation alternative to the histogram and may be looked upon as derived from histogram by joining the mid-points of the tops of consecutive rectangles. To construct a frequency polygon, we mark the frequencies along the vertical and the values of the variables along the horizontal as in the case of histogram. A dot is placed above the mid-point of each class and the height of a given dot corresponds to the frequency of the relevant class interval. Connecting these dots by a straight line, the frequency polygon is prepared.

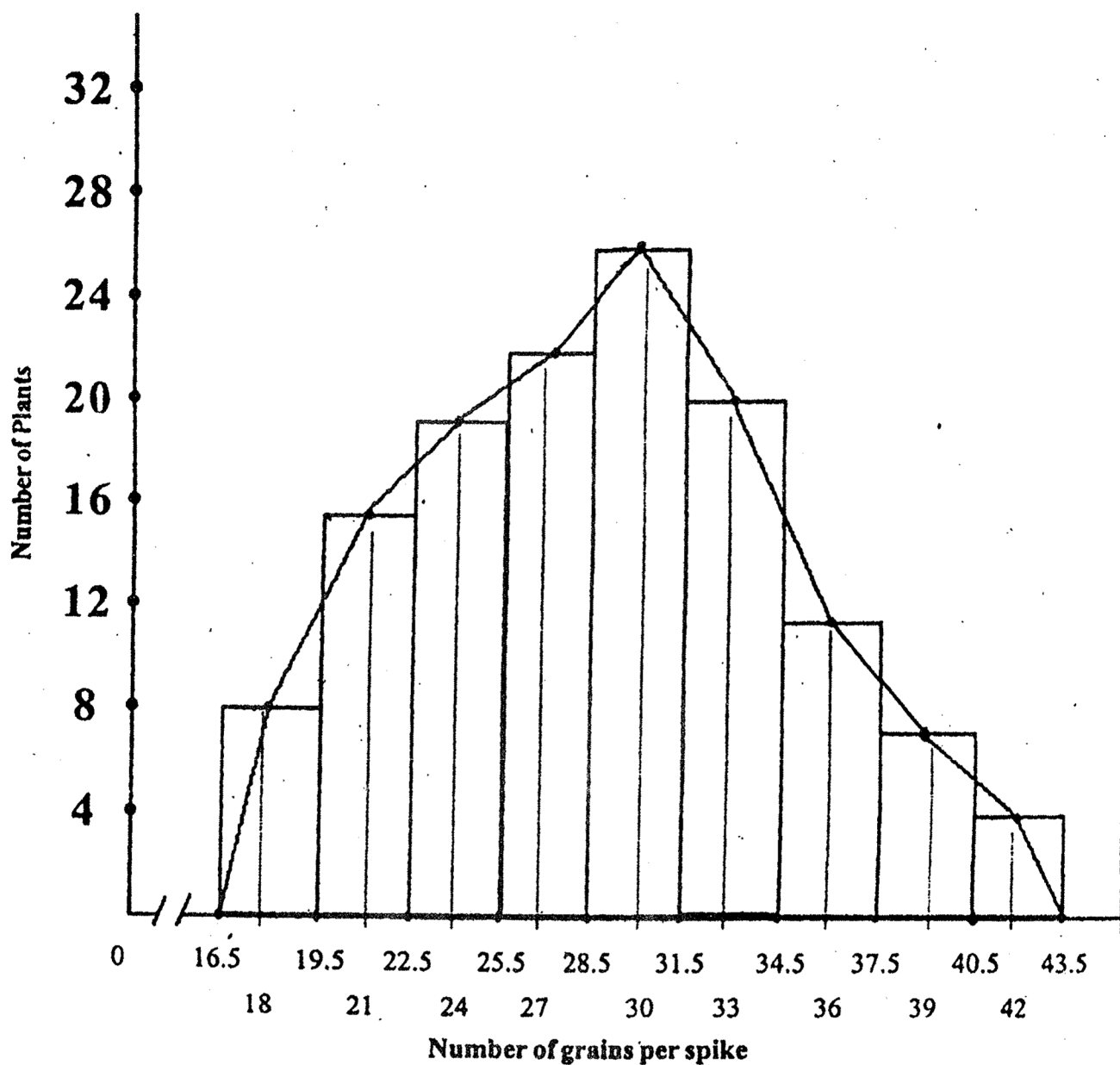
Example 5 :

Construct a frequency polygon and histogram for the following data :

<u>Number of grains</u>	<u>Number of plants</u>
17 - 19	8
20 - 22	15
23 - 25	18
26 - 28	21
29 - 31	26
32 - 34	19
35 - 37	12
38 - 40	7
41 - 43	4

Solution :

<u>Number of grains</u>	<u>Number of plants</u>
16.5 - 19.5	8
19.5 - 22.5	15
22.5 - 25.5	18
25.5 - 28.5	21
28.5 - 31.5	26
31.5 - 34.5	19
34.5 - 37.5	12
37.5 - 40.5	7
40.5 - 43.5	4



Frequency polygon and histogram representing the distribution of number of grains per spike as given in Example 5.

Sample Questions :

1. (a) Define variable and attribute. What is difference between a discrete variable and a continuous variable?
- (b) Define class limit and class boundary.
- (c) What do you understand by primary and secondary data?
2. Write short notes on the following :
 - (a) Different methods of data collection.
 - (b) Histogram
 - (c) Frequency polygon.

Content

<u>Sl. No.</u>	<u>Module No.</u>	<u>Paper</u>	<u>Page No.</u>
1.	Module No.- 19(a)	VII	1 – 51
2.	Module No.- 19(b)	VII	52 – 96

M.Sc. Course in Botany

PART-I

Paper-VII

Module No. - 19(a) BIostatistics (Introduction to Fuzzy Sets)

Contents

1. Introduction
2. Objectives
3. Keywords
4. Measures of Central Tendency
 - 4.1 Definition
 - 4.2 Types of Measures of Central Tendency
 - 4.3 Arithmetic Mean
 - 4.4 Short-cut method for Calculating Mean
 - 4.5 Step Deviation Method
 - 4.6 Corrected Mean
 - 4.7 Combined Mean
 - 4.8 Merits, Demerits and uses of Arithmetic Mean
 - 4.9 Median
 - 4.10 Calculation of Median
 - 4.11 Merits and Demerits of Median
 - 4.12 Mode

- 4.13 Computation of Mode
- 4.14 Merits and Demerits of Mode
- 4.15 Other Measures of Central Tendency
- 4.16 Exercises
- 5. Measures of Dispersion
 - 5.1 Range
 - 5.2 Quartile Deviation
 - 5.3 Mean Deviation
 - 5.4 Standard Deviation
 - 5.5 Merits and Demerits of Standard Deviation
 - 5.6 Computation of Standard Deviation
 - 5.7 Relative Measures of Dispersion
 - 5.8 Exercises
- 6. Moments, Skewness and Kurtosis
 - 6.1 Moments
 - 6.2 Skewness
 - 6.3 Kurtosis
 - 6.4 Exercises
- 7. Unit Summary
- 8. References/Suggested Further Readings

1. Introduction

This module contains three sections, in which first section deals with measures of central tendency. In measures of Central Tendency, we have discussed the definitions, merits and demerits, various examples and the exercises about mean, median and mode. The second section deals with measure of dispersion. In measure of dispersion, we have discussed the various absolute and relative

measures including the definitions, merits, demerits and examples. The last section describes the various moments, skewness and kurtosis with their definitions, properties and examples.

2. Objectives : To study this module, the reader will learn the following topics

- 1) Measures of Central Tendency
- 2) Measures of dispersion
- 3) Moments, skewness and kurtosis.

3. Keywords : Mean, Median, Mode, Range, Quartile Deviation, Mean Deviation, Standard Deviation, Moments, Skewness and Kurtosis.

1. Measures of Central Tendency :

The word 'average' or 'measures of Central Tendency' denotes a 'representative' or 'typical value' of a whole set of observations. It is a single figure which describes the entire series of observations with their varying sizes. Since a typical value usually occupies a Central position, so that some observations are larger and some others are smaller than it, averages are also known as measures of Central Tendency.

1.1 Definition : An average is a method of condensing a mass of data to a single figure, which is typical and fully representative of the entire data. It represents the whole group.

The functions of an average are :

- (i) To present the salient features of a mass of complex data.
- (ii) To facilitate comparison.
- (iii) To know about the population from a sample.
- (iv) To help in decision making.

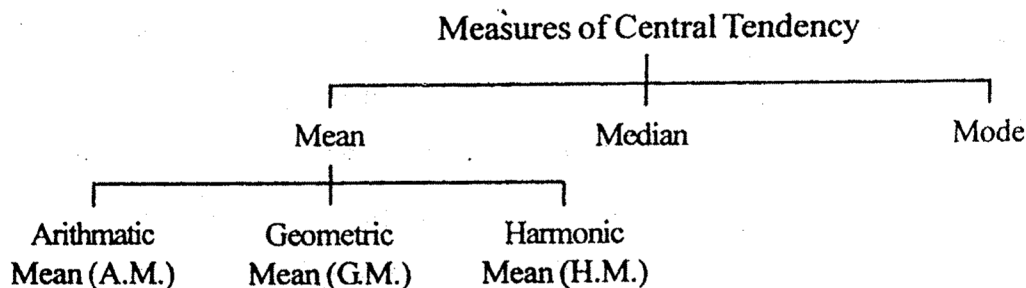
However, some desirable properties of a good measure of Central Tendency are as follows:

- (i) It should be rigidly defined.
- (ii) It should be easily comprehensible and easy to calculate.
- (iii) It should be capable of further mathematical treatment.
- (iv) It should be based on all the observations.
- (v) It should not be affected by fluctuations of random sampling.

- (vi) It should not be unduly affected by extreme observations.
- (vii) It should be easy to understand.
- (viii) It should have sampling stability.

1.2 Types of Measures of Central Tendency :

There are three measures of Central Tendency – Mean, Median and Mode. Again, Mean is of three types – Arithmetic Mean (A.M.). Geometric Mean (G.M.) and Harmonic Mean (H.M.). The words ‘mean’ and ‘average’ only refer to Arithmetic Mean.



4.3 Arithmetic Mean :

We have the following techniques to calculate the arithmetic mean of a given data.

Mean of Raw Data : We know that in an ungrouped raw data, we are given individual items. We also know that the average of n numbers is obtained by finding their sum (by adding and then dividing it by n), let x_1, x_2, \dots, x_n be n numbers, then their average is given by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Example 1. Find the arithmetic mean of the marks obtained by 10 students of a class in Mathematics in a certain examination. The marks obtained are :

25, 30, 21, 55, 47, 10, 15, 17, 45, 35.

Solution : Let \bar{x} be the average mark.

\therefore Sum of all the observations = $25 + 30 + 21 + 55 + 47 + 10 + 15 + 17 + 45 + 35 = 300$

Number of students = 10

$$\therefore \text{Arithmetic mean} = \frac{300}{10} = 30.$$

Mean of Grouped Data : Let x_1, x_2, \dots, x_n be the variates and let f_1, f_2, \dots, f_n be their corresponding frequencies, then their mean \bar{x} is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

where $N = \sum_{i=1}^n f_i = f_1 + f_2 + \dots + f_n$ total frequency.

Example 2 : Find the Arithmetic mean from the frequency table.

Marks	52	58	60	65	68	70	75
No. of Students	7	5	4	6	3	3	2

Solution : Let x be the marks and f be the frequency so that we have the following table :

x	f	xf
52	7	364
58	5	290
60	4	240
65	6	390
68	3	204
70	3	210
75	2	150

Here $N = \sum f = 30$, and $\sum fx = 1848$

$$\text{Mean } \bar{x} = \frac{1848}{30} = 61.6.$$

Example 3 : Calculate the Arithmetic Mean for the following data :

Class interval	Frequency	Class interval	Frequency
10 – 20	2	60 – 70	10
20 – 30	7	70 – 80	3
30 – 40	17	80 – 90	2
40 – 50	29	90 – 100	1
50 – 60	29		

Solution : While calculating the arithmetic mean for such a tabular data, it is assumed that the all the observations in any particular class interval have the same value. This value is the middle value or mid point of the class interval, i.e., we replace the classes by the mid values and proceed as follows:

Class interval	Mid values x	Frequency f	fx
10 – 20	15	2	30
20 – 30	25	7	175
30 – 40	35	17	595
40 – 50	45	29	1305
50 – 60	55	29	1595
60 – 70	65	10	650
70 – 80	75	3	225
80 – 90	85	2	170
90 – 100	95	1	95
TOTAL		100	4840

Here $N = \sum f = 100$, $\sum fx = 4840$

$$\text{Arithmetic Mean } \bar{x} = \frac{\sum fx}{\sum f} = \frac{4840}{100} = 48.4$$

4.4 Short-cut Method for Calculating Mean :

For Mean of ungrouped data, short-cut method is applied when the frequencies and the values of

the variables are quite large and it becomes very difficult to compute the arithmetic mean. In a frequency table of such a type the provisional mean is taken as that values of x (mid value of the class interval) which comes near the middle value of the frequency distribution. This number is called the Provisional Mean or Assumed Mean. Also find the deviations of the variates from this provisional mean. Then the arithmetic mean is given by the formula :

(i) In the case of ungrouped data

$$\bar{x} = a + \frac{\sum d}{n}, \quad \text{where } a = \text{assumed mean,}$$

n = number of items,

$d = x - a$ = deviations of any variate from a .

Working Rule for Short Cut method for ungrouped data :

Step I : Denote the variable for the discrete series by x or X .

Step II : Take any item of series, preferably the middle one, and denote it by a . This number a is called the assumed mean or provisional mean.

Step III : Take the difference $x - a$ and denote it by d or dx or $d' = x - a$, where d' is the deviation of any variate from ' a '.

Step IV. Find the sum of $\sum d$.

Step V. Use the following formula to calculate the arithmetic mean :

$$\bar{x} = a + \frac{\sum d}{N}$$

(i) In the case of grouped data

$$\bar{x} = a + \frac{\sum fd}{N}, \quad \text{where}$$

fd = product of the frequency and the corresponding deviation.

$N = \sum f$ = the sum of all the frequencies.

Working Rule for Short Cut Method for Grouped data

Step I. In the case of discrete series, denote the variable by x or X and the corresponding frequency by f . (But in the case of continuous series x is the mid value of the interval and f , the frequency corresponding to that interval).

Step II. Take any item x series, preferably the middle one and denote it by ' a '. This number ' a ' is called the **assumed mean** or **provisional mean**.

Step III. Take the difference $x - a$ and denote it by d or dx or $d = x - a$ = deviation of any variate x from a , the assumed mean.

Step IV. Multiply the respective f and d and denote the product under the column fd .

Step V. Find $\sum fd$.

Step VI. Use the following formula to calculate the arithmetic mean.

$$\bar{x} = a + \frac{\sum fd}{\sum f}$$

Example 4. Find, by short-cut-method, the mean height of the following 8 students whose height in centimetre are

59, 65, 71, 67, 61, 63, 69, 73

Solution. Let us take 65 as assumed mean, i.e., $a = 65$. Let us prepare the following table. (i)

x	$d = x - 65$
59	- 6
65	0
71	+ 6
67	+ 2
61	- 4
63	- 2
69	+ 4
73	+ 8

Total deviation = $\sum fd = 8$

Here $a = 65$, $n = 8$, $\sum fd = 8$

$$\therefore \bar{x} = a + \frac{\sum fd}{n} = 65 + \frac{8}{8} = 66 \text{ cm.}$$

Example 5. Ten coins were tossed together and the number of tails resulting from them were observed. The operation was performed 1050 times and the frequencies thus obtained for different number of tails (x) are shown in the following table. Calculate the arithmetic mean by the shortcut method.

x	0	1	2	3	4	5	6	7	8	9	10
f	2	8	43	133	207	260	213	120	54	9	1

Solution : Let 5 be the assumed mean, i.e., $a = 5$. Let us prepare the following table in order to calculate the arithmetic mean :

x	f	$d = x - 5$	fd
0	2	-5	-10
1	8	-4	-32
2	43	-3	-129
3	133	-2	-266
4	207	-1	-207
5	260	0	0
6	213	1	+213
7	120	2	+240
8	54	3	+162
9	9	4	+36
10	1	5	+5
$\sum f = 1050$			$\sum fd = 12$

$$\therefore \text{Arithmetic Mean} = \bar{x} = a + \frac{\sum fd}{\sum f} = 5 + \frac{12}{1050} = 5 + 0.0114 = 5.0114.$$

4.5 Step Deviation Method. When the class intervals in a grouped data are equal, then the calculations can be simplified further by taking out the common factor from the deviations. This common factor is equal to the width of the class-interval. In such cases, the deviation of variates x from the assumed mean ' a ' (i.e. $d = x - a$) are divided by the common factor. The arithmetic mean \bar{x} is then obtained by the following method :

$$\bar{x} = a + \frac{\sum fd}{N} \times i,$$

where a = assumed mean or Provisional Mean,

$$d = \frac{x - a}{i} = \text{the deviation of any variate from } a,$$

i = the width of the class interval,

N = the number of observations.

Example 6. Calculate, by step deviation method, the arithmetic mean of the following marks obtained by students in English.

Solution. Let $a = 20$ and $i = 5$.

X	f	$dx = x - a$	$dx' = dx / i$	fdx'
5	20	-25	-5	-100
10	43	-20	-4	-172
15	75	-15	-3	-225
20	67	-10	-2	-134
25	72	-5	-1	-72
30	45	0	0	0
35	39	5	1	39
40	9	10	2	18
45	8	15	3	24
50	6	20	4	24
$\sum f = 384$				$\sum fdx' = -598$

$$\begin{aligned}\text{Now } \bar{x} &= a + \frac{\sum fdx'}{\sum f} \times i = 30 - \frac{598}{384} \times 5 = 30 - 1.56 \times 5 \\ &= 30 - 7.80 = 22.2\end{aligned}$$

Example 7. (On weighted average). The following table gives the number of students in different classes in a Government Senior Secondary School and their tuition fees. Find the average tuition fee per student.

Class	No. of students	Tuition fee (Rs.)
V	65	0.50
VI	80	0.75
VII	95	1.00
VIII	90	1.50
IX	70	2.00

Solution. Here the values of x are 65, 80, 95, 90, 70 and their corresponding weights w 's are 0.50, 0.75, 1.00, 1.50, 2.00 respectively.

$$\begin{aligned}\text{Weighted Arithmetic Mean} &= \frac{\sum wx}{\sum x} \\ &= \frac{65 \times 0.50 + 80 \times 0.75 + 95 \times 1.00 + 90 \times 1.50 + 70 \times 2.00}{65 + 80 + 95 + 90 + 70} \\ &= \frac{32.5 + 60.0 + 95.0 + 135 + 140}{400} = \frac{462.5}{400} = 1.156.\end{aligned}$$

Example 8. Calculate the average marks, by the step deviation method, from the following data

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	42	44	58	35	26	15

Solution :

Marks	Mid value	$d = \frac{x-35}{10}$	No. of students (f)	fd
0-10	5	-3	42	-126
10-20	15	-2	44	-88
20-30	25	-1	58	-58
30-40	35	0	35	0
40-50	45	1	26	26
50-60	55	2	15	30
N = 220				$\sum fd = -216$

Here $a = 35$, $N = 220$, $\sum fd = -216$ $i = 10$

$$\therefore \bar{x} = a + \frac{\sum fd}{N} \times i$$

$$\therefore \bar{x} = 35 + \left(\frac{-216}{220} \right) \times 10 = 35 - 9.8 = 25.2$$

Example 9. In a study on patients, the following data were obtained. Find the arithmetic mean.

Age (in years)	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Number of cases	1	0	1	10	17	38	9	3

Solution : The data is presented in the form of an inclusive series. We have to transform the inclusive series in an exclusive series. It can be transformed as follows:

We measure the distance between the lower limit of the second class interval and the upper limit of the first class-interval. This is equal to $20-19=1$. We subtract $\frac{1}{2}$ of this distance (i.e, 0.5) from the lower limit and add it to the upper limit. the new classes will be formed as follows:

$$10 - 0.5 = 9.5; 19 + 0.5 = 19.5$$

The new data will be as follows:

(Age)	f	Mid value	$d = \frac{x - 44.5}{10}$	fd
(\bar{x})		x	(f)	
9.5 – 19.5	1	14.5	-3	-3
19.5 – 29.5	0	24.5	-2	0
29.5 – 39.5	1	34.5	-1	-1
39.5 – 49.5	10	44.5	0	0
49.5 – 59.5	17	54.5	1	17
59.5 – 69.5	38	64.5	2	76
69.5 – 79.5	9	74.5	3	27
79.5 – 89.5	3	84.5	4	12
$N_t = 79$				$\sum fd = 128$

$$\text{Now } \bar{x} = a + \frac{\sum fd}{N} \times i.$$

$$\bar{x} = 44.5 + \frac{128}{79} \times 10 = 44.5 + 16.2 = 60.7.$$

4.6 Corrected Mean.

Example 10. Mean of 25 observations was found to be 78.4. But later on it was found that 96 was misread as 69. Find the correct mean.

Solution. We know that the mean is given by

$$\bar{x} = \frac{\sum x}{n} \text{ or } \sum x = n\bar{x}$$

Here $\bar{x} = 78.4, n = 25$.

$$\therefore \sum x = 25 \times 78.4 = 1960.$$

But this $\sum x$ is incorrect as 96 was misread as 69.

$$\therefore \text{correct } \sum x = 1960 - 69 + 96 = 1987.$$

$$\text{correct mean} = \frac{1987}{25} = 79.48.$$

4.7 Combined mean. If we are given the mean of two series and their size, then the combined mean for the resultant series can be obtained by the formula.

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

where \bar{X} = Combined mean of the two series.

\bar{x}_1 = Mean of the first series.

\bar{x}_2 = Mean of the second series.

n_1 = Size of the first series.

n_2 = Size of the second series.

Example 11. A firm of ready-made garments make both men's and women's shirts. Its profit average is 6% of sales. Its profits in men's shirts average 8% of sales; and women's shirts comprise 60% of output. What is the average profit per sales rupee in women's shirts.

Solution. Here $\bar{X} = 6, \bar{x}_1 = 8, n_1 = 40, n_2 = 60$. Assuming that the total output is 100, we are required to find out \bar{x}_2 . We know that

$$\bar{X} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{40 \cdot 8 + 60 \cdot \bar{x}_2}{40 + 60}$$

$$6 = \frac{320 + 60 \bar{x}_2}{100}$$

$$\Rightarrow \bar{x}_2 = \frac{600 - 320}{60} = \frac{280}{60} = \frac{10}{3} = 4.66.$$

Thus, the average profits in women's shirt is Rs. 4.66 per cent of sales, and Rs. 0.0466 per sale rupee.

4.8 Merits, Demerits and uses of Arithmetic Mean

Merits :

1. It can be easily calculated.
2. Its calculation is based on all the observations.

3. It is easy to understand.
4. It is rightly defined by the mathematical formula.
5. It is least affected by sampling fluctuations.
6. It is the best measure to compare two or more series (datas).
7. It is the average obtained by calculations and it does not depend upon and position.

Demerits :

1. It may not be represented in actual data so it is theoretical.
2. The extreme values have greater affect on mean.
3. It cannot be calculated if all the values are not known.
4. It cannot be determined for the qualitative data such as love, beauty, honesty, etc.
5. Mean may lead to fallacious conditions in the absence of original observations.

Uses of Arithmetic Mean :

1. A common man uses for calculating average marks obtained by a student.
2. It is extremely used in practical statistics.
3. Estimates are always obtained by mean.
4. Businessman uses it to find out the operation cost, profit per unit of article, output per man and per machine, average monthly income and expenditure, etc. etc.

45.9 Median

Median is defined as the middle-most or the central value of the variable in a set of observations, when the observations are arranged either in ascending or in descending order of their magnitudes. It divides the arranged series in two equal parts. Median is a position average, whereas, the arithmetic mean is the calculated average. When a series consists of an even number of terms, median is the arithmetic mean of the two central items. It is generally denoted by M.

4.10 Calculation of Median.

(a) **When the data is ungrouped.** Arrange the n values of the variable in ascending (or descending) order of magnitudes.

Case I. When n is odd. In the case $\frac{n+1}{2}$ th value is the median

i.e. $M = \frac{n+1}{2}$ term.

Case II. When n is even. In this case there are two middle terms $\frac{n}{2}th$ and $\left(\frac{n}{2}+1\right)th$. The median is the average of these two terms, i.e.,

$$M = \frac{\frac{n}{2} + \left(\frac{n}{2} + 1\right)}{2}$$

Example 12. The number of runs scored by 11 players of a cricket team of a school are

5, 19, 42, 11, 50, 30, 21, 0, 52, 36, 27

Find the median.

Solution. Let us arrange the values in ascending order :

0, 5, 11, 19, 21, 27, 30, 36, 42, 50, 52.

$$\begin{aligned} \therefore \text{Median} = M &= \left(\frac{n+1}{2}\right)th \text{ value} = \left(\frac{11+1}{2}\right)th \text{ value.} \\ &= 6^{th} \text{ value.} \end{aligned}$$

Now the 6^{th} value in the data (i) is 27.

\therefore Median = 27 runs.

Example 13. Find the median of the following items :

6, 10, 4, 3, 9, 11, 22, 18

Solution. Let us arrange the items in ascending order.

3, 4, 6, 9, 10, 11, 18, 22.

In this data the number of items is $n = 8$, which is even.

$$\begin{aligned} \therefore \text{Median} = M &= \text{average of } \left(\frac{n}{2}\right)th \text{ and } \left(\frac{n}{2}+1\right)th \text{ terms.} \\ &= \text{Average of } \left(\frac{8}{2}\right)th \text{ and } \left(\frac{8}{2}+1\right)th \text{ terms.} \\ &= \text{average of } 4th \text{ and } 5th \text{ terms.} \\ &= \frac{9+10}{2} = \frac{19}{2} = 9.5. \end{aligned}$$

(b) When the data is grouped

Case (1). When the series is discrete. In this case the values of the variable are arranged in ascending or descending order of magnitudes. A table is prepared showing the corresponding frequencies and cumulative frequencies. Then the median is calculated by the following formula:

$$M = \left(\frac{n+1}{2} \right) \text{th}$$

where $n = \sum f = \text{total frequencies.}$

Example 14. Calculate median for the following data :

No. of students	6	4	16	7	8	2
Marks	20	9	25	50	40	80

Solution. Arranging the marks in ascending order and preparing the following table :

Marks	Frequency	Cumulative Frequency
9	4	4
20	6	10
25	16	26
40	8	34
50	7	41
80	2	43
$n = \sum f = 43$		

Here $n = 43$.

$$\therefore \text{Median} = M = \left(\frac{n+1}{2} \right) \text{th value}$$

$$= \left(\frac{43+1}{2} \right) \text{th value} = 22\text{nd value.}$$

The above table shows that all items from 11 to 26 have their values 25. Since 22nd item lies in this interval, therefore, its value is 25.

Hence Median = 25 marks.

(c) When the series is continuous. In this case the data is given in the form of a frequency table with class-interval, etc. and the following formula is used to calculate the Median.

$$M = L + \frac{\frac{n}{2} - C}{f} \times i, \text{ where}$$

L = lower limit of the class in which the median lies

n = total number of frequencies, i.e. $n = \sum f$

f = frequency of the class in which the median lies.

C = Cumulative frequency of the class preceding the median class.

i = Width of the class interval of the class in which the median lies.

Example 15. The following table gives the marks obtained by 80 students in Economics. Find the median.

Marks	No. of students	Marks	No. of students
10 – 14	4	30 – 34	7
15 – 19	6	35 – 39	3
20 – 24	10	40 – 44	9
25 – 29	5	45 – 49	6

Solution. Let us prepare the following table showing the frequencies and cumulative frequencies:

Marks	Frequency	Cumulative Frequency
10 – 14	4	4
15 – 19	6	10
20 – 24	10	20
25 – 29	5	25
30 – 34	7	32
35 – 39	3	35
40 – 44	9	44
45 – 49	6	50

Here $n = 50$, $\therefore \frac{n}{2} = 25$.

Also $\frac{n}{2} = 25$.

$\therefore L$ = Lower limit of the median class = 24.5.

C = Cumulative frequency of the class (20–24) preceding the median class = 20.

f = Frequency of the median class = 5.

i = Class-interval of the median class = 5.

$$\begin{aligned}\therefore \text{Median} &= 24.5 + \frac{25-20}{5} \times 5 \\ &= 24.5 + 5 = 29.5.\end{aligned}$$

Example 16. The following table gives the weekly expenditure of 100 families. Find the median weekly expenditure.

Weekly Expenditure (in Rs.)	Number of Families
0 – 10	14
10 – 20	23
20 – 30	27
30 – 40	21
40 – 50	15

Solution. Let us prepare a table which gives the frequency and cumulative frequency:

Weekly Expenditure (in Rs.)	Number of families (frequency)	Cumulative frequency
0 – 10	14	14
10 – 20	23	37
20 – 30	27	64
30 – 40	21	85
40 – 50	15	100
	Here $n = \sum f = 100$	

$$\therefore \text{Median} = \left(\frac{n}{2}\right)\text{th value} = \left(\frac{100}{2}\right)\text{th value} = 50\text{th value.}$$

Median class = 20 – 30.

Here $\frac{n}{2} = 50$, $L = 20$, $f = 27$, $C = 37$, $i = 10$.

$$\therefore \text{Median} = L + \frac{\frac{n}{2} - C}{f} \times i = 20 + \frac{50 - 37}{27} \times 10 = 24.81.$$

4.11 Merits and Demerits

Merits

1. It is easily understood.
2. It is not affected by extreme values.
3. It can be located graphically.
4. It is the best measure for qualitative data such as beauty, intelligence, etc.
5. It can be easily located even if the class-intervals in the series are unequal.
6. It can be determined even by inspection in many cases.

Demerits

1. It is not subject to algebraic treatments.
2. It cannot represent the irregular distribution series.

Example 17. In a survey of 950 families in a village, the following distribution of numbers of children was obtained :

No. of children	0-2	2-4	4-6	6-8	8-10	10-12
No. of families	272	328	205	120	15	10

Find the mean and median of the above distribution.

Solution. Let us prepare the following table by taking 7 as assumed mean, i.e. $a = 7$.

Class	x_i	f_i	c.f.	$d'_i = \frac{x_i - 7}{2}$	$f_i d'_i$
0-2	1	272	272	-3	-819
2-4	3	328	600	-2	-656
4-6	5	205	805	-1	-205
6-8	7	120	925	0	0
8-10	9	15	940	1	15
10-12	11	10	950	2	20
950					-1642

$$\text{Now } A.M. = a + \frac{\sum_{i=1}^n f_i d'_i}{\sum_{i=1}^n f_i} \times i = 7 + \frac{-1642}{950} \times 2$$

$$= 7 - \frac{3284}{950} = 7 - 3.46 = 3.54$$

For Median. $\frac{N}{2} = \frac{950}{2} = 475$

\therefore Median class = 2 - 4; $L = 2$; $C = 600$; $i = 2$.

$$\therefore \text{Median} = L + \frac{\frac{N}{2} - C}{f} \times i = 2 + \frac{475 - 272}{328} \times 2$$

$$= 2 + \frac{203}{328} \times 2 = 2 + \frac{406}{328} = 2 + 1.24$$

$$= 3.24.$$

4.12 Mode

Mode is defined as that value in a series which occurs most frequently. In a frequency distribution mode is that variate which has the maximum frequency. In other words, mode represents that value which is most frequent or typical or predominant. For example in the series 6, 5, 3, 4, 3, 7, 8, 9, 5, 5, 4 we notice that 5 occurs most frequently, therefore, 5 is the mode. Mode is also known as the Norm.

Example 18. A Bata shop in Delhi had sold 100 pairs of shoes of Bata exclusive on a certain day with the following distribution :

Size of Shoe	4	5	6	7	8	9	10
No. of Pairs	10	15	20	35	16	3	1

Find the mode of the distribution.

Solution. Let us prepare the table showing the frequency.

Size of Shoe	4	5	6	7	8	9	10
Frequency	10	15	20	35	16	3	1

In the above table we notice that the size 7 has the maximum frequency, viz. 35. Therefore, 7 is the mode of the distribution.

4.13 Computation of Mode. (a) **Simple series.** In the case of **Simple series** the value which is repeated maximum number of times is the mode of the series.

Example 19. In Rajdhani Rubber Industry, Tilak Nagar, New Delhi, seven labourers are receiving the daily wages of Rs. 5, 6, 6, 8, 8, 8 and 10. Find the modal wage.

Solution. In the series 5, 6, 6, 8, 8, 8, 10; 8 occurs thrice and no other item occurs three times or more than three times and hence the modal wage is Rs. 8.

(b) **Discrete frequency distribution series.** In the case of discrete frequency distribution, mode is the value of the variable corresponding to the maximum frequency.

Example 20. A set of numbers consists of four 4's, five 5's, six 6's and nine 9's. What is the mode?

Size of item	4	5	6	9
Frequency	4	5	6	9

Since 9 has the maximum frequency viz. 9, therefore 9 is the mode.

Thus, we notice that in discrete series, mode is determined by inspection and therefore, an error of judgement is possible in these cases where the difference between the maximum frequency and the frequency preceding or succeeding it is very small and the items are heavily concentrated on either side. Under such circumstances the value of mode is determined by preparing a grouping table and analysis table. A grouping table has the following six columns:

Column I. It has original frequencies and the maximum frequency is marked by bold type.

Column II. In this column the frequencies of column I are combined *two by two*. Here also the maximum frequency is marked by bold type.

Column III. Here, we leave the first frequency of the column I and combine the others in *two by two*. Again the maximum frequency is marked by bold type.

Column IV. In this column the frequencies of the column I are combined (grouped) in *three by three* and again the maximum frequency is marked by bold type.

Column V. Here we leave the first frequency of the column I and group the others *three by three*. Again mark the maximum frequency by bold type.

Column VI. Now leave the first two frequencies of column I and combine the others in *three by three*. Mark the maximum frequency by bold type.

After preparing the grouping table, we prepare the analysis table. While preparing this table we put the column numbers on the left hand side and the various probable values of the mode on the right-hand side, i.e., values against which frequencies are maximum marked in the grouping table. The value which occurs maximum number of times is the mode.

The procedure of preparing a grouping table and analysis table shall be clear from the following table:

Example 21. Calculate the mode of the following frequency distribution :

Size (x)	4	5	6	7	8	9	10	11	12	13
Frequency (f)	2	5	8	9	12	14	14	15	11	13

Solution. This problem is solved by the method of grouping as it an irregular distribution in the sense that the difference between maximum frequency 15 and frequency preceding it, is very small. Let us prepare the grouping and analysis table.

Grouping Table

Size x	Frequency f	Grouping				
		of two	of two leaving the first	of three	of three leaving the first	of three leaving the first
	(I)	(II)	(III)	(IV)	(V)	(VI)
4	2	} 7	} 13	} 15	} 22	} 29
5	5					
6	8	} 17	} 21	} 35	} 40	} 43
7	9					
8	12	} 26	} 28	} 40	} 39	
9	14					
10	14	} 29	} 26			
11	15					
12	11	} 24				
12	13					

Let us now prepare the analysis table.

Analysis Table

Columns	Size of items having maximum frequency
I	11
II	10, 11
III	9, 10
IV	10, 11, 12
V	8, 9, 10
VI	9, 10, 11

Since the item 10 occurs maximum number of times, viz: 5 times, Hence mode is 10.

Continuous Frequency Distribution.

(i) **Modal class.** It is that class in grouped frequency distribution in which the mode lies. The modal class can be determined either by inspection or with the help of grouping table. After finding the modal class, we calculate the mode by the following formula:

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i,$$

where l = the lower limit of the modal class.

i = the width of the modal class.

f_1 = the frequency of the class preceding modal class.

f_m = the frequency of the modal class.

f_2 = the frequency of the class succeeding modal class.

Sometimes it so happened that the above formula fails to give the mode. In this case, the modal value lies in a class other than the one containing maximum frequency. In such cases we take the help of the following formula;

$$\text{Mode} = l + \frac{f_2}{f_1 + f_2} \times i$$

where l, f_1, f_2, i have usual meanings.

Asymmetrical Distribution. A distribution in which mean, median and mode coincide is called asymmetrical distribution. If the distribution is moderately asymmetrical then, mean, median and mode are connected by the formula.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

The procedure of finding the mode by the above methods shall be clear by the following examples.

This method is generally used when we have to locate mean, median, mode simultaneously.

Example 22. Find the mode for the following data :

Marks	No. of Students
1 – 5	7
6 – 10	10
11 – 15	16
16 – 20	32
21 – 25	24

Solution. From the above table it is clear that the maximum frequency is 32 and it lies in the class 16 – 20. Thus the modal class is 16 – 20.

Here $l = 16$, $f_m = 32$, $f_1 = 16$, $f_2 = 24$, $i = 5$.

$$\begin{aligned} \therefore \text{Mode} &= l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i = 16 + \frac{32 - 16}{64 - 16 - 24} \\ &= 16 + \frac{16}{24} \times 5 = 16 + \frac{10}{3} = 16 + 3.33 = 19.33. \end{aligned}$$

4.14 Merits and Demerits of Mode.

Merits

- (1) It can be easily understood.
- (2) It can be located in some cases by inspection.
- (3) It is capable of being ascertained graphically.
- (4) It is not affected by extreme values.
- (5) It represents the most frequent value and hence it is very often in practice.

- (6) - The arrangement of data is not necessary if the items are a few.

Demerits

- (1) There are different formulae for its calculations which ordinarily given different answers.
- (2) Mode is determinate. Some series have two or more than two modes.

4.15 Other Measures of Central Tendency

Midrange. The midrange is the value midway between the smallest and largest values in the sample, that is, the arithmetic mean of the largest and the smallest values. for example, in the set of radiologic counts 4, 5, 9, 1, 2, the midrange is $(9 + 1)/2$ or 5. it is clear that the midrange will be influenced by extreme values.

Geometric Mean. The geometric mean of a set of observations is the n th root of their product. the computation of the geometric mean requires that all observations be positive, that is, greater than zero. For example, the geometric mean of the radiologic counts 4 and 9 is $\sqrt{4 \cdot 9} = \sqrt{36}$ or 6. A general formula for computing the geometric mean of the set of observations x_1, x_2, \dots, x_n is $\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$. The geometric mean is sometimes denoted by \bar{x}_g . An interesting property is that the logarithm of the geometric mean is the arithmetic mean of the logarithms of the individual observations. This result is expressed by the formula

$$\log \bar{x}_g = \frac{\sum_{i=1}^n \log x_i}{n}$$

In addition to other applications, the geometric mean is used in microbiology for computing average dilution titers.

Harmonic Mean. The harmonic mean of a set of observations is the reciprocal of the arithmetic mean of the reciprocals of the observations. That is, if the observations are x_1, x_2, \dots, x_n then the harmonic mean is

$$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

The harmonic mean is often denoted by \bar{x}_h . It is an interesting example of the usefulness of the harmonic mean. The problem of determining the average velocity of a car that has travelled the first 10 miles of a trip at 30 miles per hour and the second 10 miles at 60 miles per hour. At first glance the average velocity would seem to be the simple average of 30 and 60, that is, 45 miles per hour. However, this kind of average is usually defined to be total distance divided by total time. Here the total distance is 20 miles, whereas the total time $\frac{1}{3}$ hour plus $\frac{1}{6}$ hour of $\frac{1}{2}$ hour, producing an average velocity of 40 miles per hour rather than 45 miles per hour.

4.16 EXERCISES

1. What are the measures of central tendency and give their relative merits and demerits.
2. Give a critical review of the different measures of Central tendency, with examples.
3. How do you calculate the mean of a grouped frequency distribution?
4. Define mean. What are its merits and demerits? Also give its uses.
5. Define median. What are its merits and demerits? Also give its uses.
6. Define mode. What are its merits and demerits? Also give its uses.
7. Compute the mean from the frequency table :

Marks	70	50	60	52	65	75	68
No. of Students	3	5	4	7	6	2	3

8. Find the mean height from the following frequency distribution

Heights in cms	150	160	158	155	164	168
No. of Students	10	14	8	15	7	16

9. Find the median of the following items :

5, 8, 16, 12, 11, 15, 10, 13, 6, 18, 20.

10. Find the median for the following data :

Class Interval	Frequency
0 – 10	28
10 – 20	35
20 – 30	24
30 – 40	41
40 – 50	18
50 – 60	11
60 – 70	8
70 – 80	7
80 – 90	3
90 – 100	1

Ex. 11 Calculate the mode for the following distribution :

x	10	20	30	40	50	60	70
y	17	22	31	39	27	15	13

Ex. 12 Find mean median and mode for the following frequency distribution :

Marks	No. of Students	Marks	No. of Students
0 – 10	2	0 – 60	79
0 – 20	6	0 – 70	94
0 – 30	21	0 – 80	98
0 – 40	39	0 – 90	100
0 – 50	61		

Ex. 13 Find the mean and the median for the following data, and comment on the shape of the distribution :

Weight in kg. :	36–40	41–45	46–50	51–55	56–60	61–65	66–70
No. of Persons	14	26	40	53	50	37	25

Ex. 14 Find the mean and mode for the following :

Year under :	10	20	30	40	50	60
No. of Persons :	15	32	51	78	97	109

Ex. 15 Find the median and the median class of the data given below :

Class boundaries	15–25	25–35	35–45	45–55	55–65	65–75
Frequency	4	11	19	14	0	2

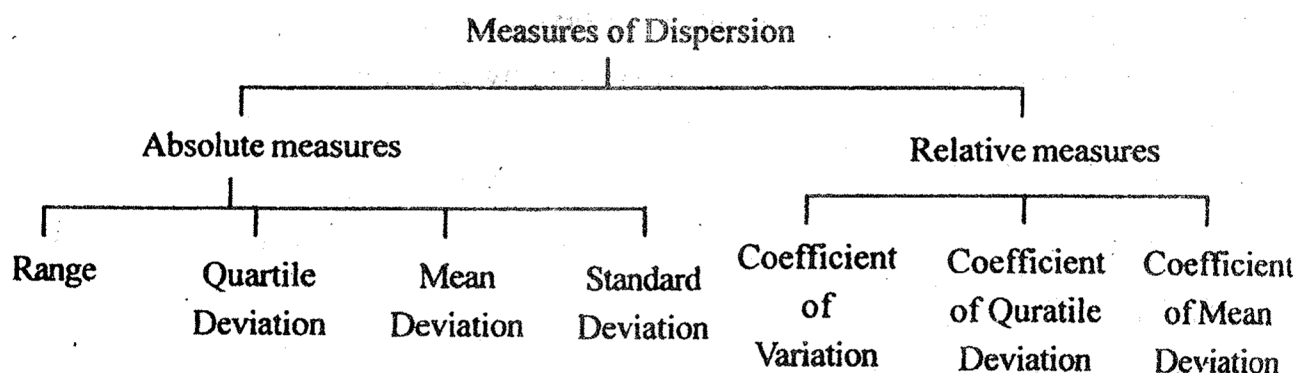
Answers :

(7) 60.27 (8) 159.5 (9) 12 (10) 30.24 (11) 40 (12) Mean = Median = Mode = 45 (13) Mean = 54.3 Kg. Median = 54.5 kg. (14) Mean = 29.95, Median = 35.00 (15) Median = 40.26, Median Class = 35–65.

5. Measures of Dispersion

The word dispersion or variability is used to denote the “degree of heterogeneity” in the data. It is an important characteristic indicating the extent to which observations vary among themselves. The dispersion of a given set of observations will be zero, only when all of them are equal. The wider discrepancy from one observation to another, the larger will be the dispersion.

A measure of dispersion is designed to state numerically the extent to which individual observations vary on the average. There are several measures to dispersion.



Now we discuss the different absolute measures one by one.

5.1 Range of a set of observations is the difference between the maximum and minimum values.

Range = Maximum value – Minimum value.

As for example, for the observations, (in Rs.) 6, 4, 1, 6, 5, 10, 3, the maximum and the minimum. Values are 10 and 1. Therefore,

$$\text{Range} = 10 - 1 = 9 \text{ Rs.}$$

Advantages :

- (i) It is easy to understand.
- (ii) Range is also simple to calculate.
- (iii) Its units are the same as the units of the variable being measured.

Limitations :

- (i) It does not depend on all observations, and is based on only the largest and the smallest among them. The values of intermediate observations, are not of all necessary for its calculation.
- (ii) It is highly affected by extreme values.
- (iii) It does not take into account the form of the distribution.
- (iv) The greatest disadvantage of Range is that it cannot be calculated from frequency distributions with open-end classes.

5.2 Quartile Deviation or Semi-interquartile Range :

Quartile Deviation is defined as half the difference between the upper and lower quartiles.

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

The difference $Q_3 - Q_1$ being the distance between the two quartiles, may be called interquartile range, and half of this is semi-interquartile Range.

Example 1. Calculate the quartile deviation from the following :

Class interval :	10-15	15-20	20-25	25-30	30-40	40-50	50-60	60-70
Frequency :	4	12	16	22	10	8	6	4

Solution. In order to compute Quartile deviation, we have to find Q_1 (1st quartile) and Q_3 (2nd quartile) i.e., values of the variable corresponding to cumulative frequencies $N/4$ and $3N/4$. Here total frequency $N = 82$. Therefore $N/4 = 20.5$, and $3N/4 = 61.5$.

Class boundary	Cumulative frequency (less than)
10	0
15	4
20	16
Q_1	$N/4 = 20.5$
25	32
30	54
Q_3	$3N/4 = 61.5$
40	64
50	72
60	78
70	$82 = N$

Applying simple interpolation

$$\frac{Q_1 - 20}{25 - 20} = \frac{20.5 - 16}{32 - 16}$$

$$\Rightarrow Q_1 - 20 = \frac{4.5 \times 5}{16} = 1.4$$

$$\Rightarrow Q_1 = 21.4$$

$$\text{Similarly, } \frac{Q_3 - 30}{40 - 30} = \frac{61.5 - 54}{64 - 54}$$

$$\Rightarrow Q_3 = 37.5$$

$$\therefore \text{Quartile deviation } \frac{Q_3 - Q_1}{2} = \frac{37.5 - 21.4}{2} = 8.0$$

5.3 Mean Deviation :

Mean Deviation (also called Mean Absolute Deviation) of a set of observations is the arithmetic mean of absolute deviations from mean or any other specified value. Given the observations x_1, x_2, \dots, x_n in order to find 'Mean Deviation about A', we first obtain the deviations $(x_1 - A), (x_2 - A), \dots, (x_n - A)$.

Some of these deviation may be positive and some negative. If we write $|x_i - A|$ to denote the positive value of $(x_i - A)$, whatever the actual sign, the sum of these 'absolute deviations' is

$$|x_1 - A| + |x_2 - A| + \dots + |x_n - A| = \sum |x_i - A|$$

and A.M. of the absolute deviations is

$$\text{Mean Deviation about } A = \frac{1}{n} \sum |x_i - A|$$

Mean Deviation is usually calculated about arithmetic mean (\bar{x}), and hence for simple series,

$$\text{Mean Deviation} = \frac{1}{n} \sum |(x_i - \bar{x})|$$

For frequency distribution,

$$\text{Mean Deviation} = \frac{1}{N} \sum f_i |x_i - \bar{x}|, \text{ where } N = \sum f_i$$

Example 2. Calculate the Mean Deviation of the following values about the median : 8, 15, 53, 49, 19, 62, 7, 15, 95, 77.

Solution. Since there are an even number of observations, so, 10, the median is the average of the two middle most observations, when arranged in order of magnitude 7, 8, 15, 15, (19, 49), 53, 62, 77, 95.

$$\text{Median} = \frac{19 + 49}{2} = 34.$$

Table for calculations of Mean Deviation

x	$ x - \text{median} $ i.e., difference from median
8	26
15	19
53	19
49	15
19	15
62	28
7	27
15	19
95	61
77	43
Total	272

Mean Deviation about median

$$= \frac{1}{n} \sum |x - \text{median}|$$

$$= \frac{1}{10} \times 272 = 27.2$$

Example 3. Find the mean deviation of the following series :

x	10	11	12	13	14	15
Frequency	3	12	18	12	3	48

Solution. Table for calculations for Mean Deviation

x	f	fx	$ x - \bar{x} $	$f x - \bar{x} $
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
Total	48	576	—	36

$$\text{Mean } \bar{x} = \frac{\sum fx}{N} = \frac{\sum fx}{\sum f} = \frac{576}{48} = 12$$

$$\text{So Mean Deviation about mean} = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{36}{48} = 0.75.$$

5.4 Standard Deviation (S.D.)

Standard Deviation of a set of observations is the square root of the arithmetic mean of squares of deviations from arithmetic mean. In short, S.D. may be defined as “Root-Mean-Square-Deviation from mean”. It is usually denoted by the Greek small letter σ (sigma).

If x_1, x_2, \dots, x_n be a set of observations and \bar{x} their A.M. then,

Deviations from : $(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$

Square-Deviations from mean : $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$

Mean-Square-Deviation from mean :

$$\frac{1}{n} \left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right] = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Root-Mean-Square-Deviation from mean, i.e.

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

The square of standard deviation, i.e., σ^2 , is known as variance.

$$\text{Variance} = (\text{S.D.})^2$$

$$\text{For simple series, } \sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

For frequency distribution,

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

S.D. is always considered as positive. Thus, S.D. is the positive square root of variance.

Important properties of S.D. :

- (a) S.D. is independent of the change of origin; i.e. if $y = x - c$ where c is a constant, then

$$\text{S.D. of } x = \text{S.D. of } y$$

In symbols, $\sigma_x = \sigma_y$.

This implies that the same S.D. will be obtained if each of the observations is increased or decreased by a constant.

- (b) If two variables x and z are so related that $z = ax + b$ for each $x = x_i$, where a and b are constants, then

$$\sigma_z = |a| \sigma_x$$

where $|a|$ denotes the positive value of a .

In particular, if $y = (x - c)/d$, where c and d are constants (d positive), then $\sigma_x = d \cdot \sigma_y$

This implies that S.D. does not depend on origin, but depends on scale of measurement. If each observation is multiplied or divided by a constant, S.D. will also be similarly affected.

(c) If a group of n_1 observations has mean \bar{x}_1 and S.D. σ_1 , and another group of n_2 observations has mean \bar{x}_2 and S.D. σ_2 then S.D. (σ) of the composite group of $n_1 + n_2$ ($= N$ say) observations can be obtained by the formula

$$N\sigma^2 = (n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1d_1^2 + n_2d_2^2) \dots\dots\dots(1)$$

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

Relation (1) may be extended to any number of groups :

$$N\sigma^2 = \sum n_i\sigma_i^2 + \sum n_id_i^2$$

where $d_i = \bar{x}_i - \bar{x}$, $N = \sum n_i$, and \bar{x} is the mean of composite group given by $N\bar{x} = \sum n_i\bar{x}_i$

(d) S.D. is the minimum root-mean-square-deviation, i.e.

$$\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \leq \sqrt{\frac{1}{n} \sum (x_i - A)^2}$$

whatever be the value of A .

5.5 Merits and Demerits of Standard Deviation :

Merits :

1. It is based on all the observations.
2. It is rigidly defined.
3. It lends itself to further algebraic treatment.
4. It is less affected by fluctuations of sampling as compared to other measures of dispersion.
5. It is extremely useful in correlation.
6. Like mean deviation, there is no artificiality in it.

Demerits

1. It is difficult to compute unlike other measures of dispersion.
2. It is not simple to understand.
3. It gives more weightage to extreme values.

Uses : The standard deviation is most useful in the case of the normal frequency distribution. The majority of measurements arising in biological and medical data form a distinct pattern.

5.6 Computation of Standard Deviation

The methods of calculating the standard deviation depend upon the nature of data and also on the number of observations for grouped data.

(a) Standard Deviations for grouped data

Direct method. In case of simple series, the standard deviation can be obtained by the formula.

$$\sigma = \sqrt{\frac{(x - \bar{x})^2}{n}} = \sqrt{\frac{\sum d^2}{n}}, \text{ where } d = x - \bar{x}$$

and x = value of the variable or observation.

\bar{x} = arithmetic mean.

n = total number of observations.

Example 4. Find the standard deviation of 16, 13, 17, 22.

Solution. Here $A.M. = \bar{x} = \frac{16+13+17+22}{4} = \frac{68}{4} = 17$.

Let us prepare the following table in order to calculate the standard deviation.

(x)	$d = x - \bar{x} = x - 17$	$(x - \bar{x})^2$
16	-1	1
13	-4	16
17	0	0
22	5	25
		$\sum d^2 = 42$

$$\text{Now } \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{42}{4}} = 3.2.$$

Short-cut Method. This method is applied to calculate the standard deviation, when the mean of the data comes out to be a fraction. In that case, it is very difficult and tedious to find the deviations of

all observations from the mean by the earlier method. The formula used is

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

where $d = x - A$, $A = \text{assumed mean}$,

$n = \text{total number of observations}$.

Example 5. Find the standard deviation of the following data :

48, 43, 65, 57, 31, 60, 37, 48, 59, 78.

Solution. Let us prepare the following table in order to calculate the value of S.D.:

Value (x)	$d = x - A, (A = 50)$	d^2
48	-2	4
43	-7	49
65	15	225
57	7	49
31	-19	361
60	10	100
37	-13	169
48	-2	4
59	9	81
78	28	784
$n = 10$	$\sum d = 26$	$\sum d^2 = 1826$

Here $\bar{x} = A + \frac{\sum d}{n} = 50 + \frac{26}{10} = 52.6$.

which is a fraction. Let us apply the short-cut formula in order to calculate S.D.

$$\begin{aligned} \therefore S.D. = \sigma &= \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{1826}{10} - \left(\frac{26}{10}\right)^2} \\ &= \sqrt{182.60 - 6.76} = \sqrt{175.84} = 13.26 \end{aligned}$$

(b) Standard deviation for grouped data or discrete series.

Direct Method. The standard deviation for the discrete series is given by formula

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

where \bar{x} is A.M., x is the size of the item, and f is the corresponding frequency in the case of discrete series. But when the mean has a fractional value, then the following formula is applied to calculate S.D.

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

where $d = x - A$, A = assumed mean, $n = \sum f$ = total frequency.

(c) Standard deviation in continuous series

Direct Method. The standard deviation in the case of continuous series is obtained by the following formula :

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

where x = mid-value, \bar{x} = A.M., f = frequency, n = total frequency.

Short Method. The formula for short method to find the standard deviation of continuous series is

$$\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times i,$$

where $d = \frac{x - A}{i}$, A = Assumed mean,

n = total frequency. i = class width.

Example 6. Find the standard deviation from the following data :

Size of the item :	10	11	12	13	14	15	16
Frequency :	2	7	11	15	10	4	1

Also find the coefficient of variation.

Solution. Let us prepare the following table :

Size of the items (x)	Frequency f	$d = x - A$ $A = 13$	fd	fd^2
10	2	-3	-6	18
11	7	-2	-14	28
12	11	-1	-11	11
13	15	0	0	0
14	10	1	10	10
15	4	2	8	16
16	1	3	3	9
Total	$n = \sum f = 50$		$\sum fd = -10$	$\sum fd^2 = 92$

$$\text{Now } A.M. = \bar{x} = A + \frac{\sum fd}{n} = 13 + \frac{(-10)}{50} = 12.8$$

Here $\bar{x} = 12.8$ is a fraction,

$$\begin{aligned} \therefore S.D. = \sigma &= \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} = \sqrt{\frac{92}{50} - \left(\frac{-10}{50}\right)^2} \\ &= \sqrt{1.84 - 0.04} = \sqrt{1.80} = 1.342. \end{aligned}$$

\therefore Now the coefficient of variation

Example 7. Find the standard deviation for the following distribution :

Marks :	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Student :	5	12	15	20	10	4	2

Solution. Let us prepare the following table in order to calculate the standard deviation :

Marks (Class Interval)	No. of Students (f)	Mid-value (x)	$d = \frac{x-45}{10}$	fd	fd ²
10 – 20	5	15	-3	-15	45
20 – 30	12	25	-2	-24	48
30 – 40	15	35	-1	-15	15
40 – 50	20	45	0	0	0
50 – 60	10	55	1	10	10
60 – 70	4	65	2	8	16
70 – 80	2	75	3	6	18
Total	$\sum f = n = 68$			$\sum fd = -30$	$\sum fd^2 = 152$

$$\therefore \sigma = i \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} = 10 \times \sqrt{\frac{152}{68} - \left(\frac{-30}{68}\right)^2} = 14.3 \text{ Approx.}$$

Example 8 : Calculate the mean, median and variance of the following data :

Height in cm.	95–105	105–115	115–125	125–135	135–145
No. of Children	19	23	36	70	25

Solution. Let us prepare the following data :

Class Interval	Mid-value (x)	No. of Children (f)	$d = \frac{x-a}{i}$ $= \frac{x-120}{10}$	fd	fd ²
95 – 105	100	19	-2	-38	76
105 – 115	110	23	-1	-23	23

115 – 125	12	36	0	0	0
125 – 135	130	70	1	70	70
135 – 145	140	52	2	104	208
$N = 200$				113	377

$$\text{Now Mean} = a + \frac{\sum fd}{N} \times i = 120 + \frac{113}{200} \times 10 = 120 + 5.65 = 125.65$$

Median. The median class is 125 – 135 as $N/2 = 100$ lies in it.

$$\begin{aligned} \therefore \text{Median} &= L + \frac{N/2 - C}{f} \times i = 125 + \frac{100 - 78}{70} \times 10 \\ &= 125 + 3.14 = 128.14. \end{aligned}$$

$$\text{Variance} : = \sigma^2 = \left[\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2 \right] \times i^2.$$

$$\therefore \sigma^2 = \left[\frac{377}{100} - \left(\frac{113}{100} \right)^2 \right] \times 100 = (3.77 - 1.2759) \times 100 = 249.31.$$

Example 9. In a study to test the effectiveness of a new variety of seeds, an experiment was performed with 50 experimental field and the following results of yield per hectare (in quintals) were obtained.

Yielding	No. of fields	Yield	No. of fields
31 – 35	2	51 – 55	16
36 – 40	3	56 – 60	5
41 – 45	8	61 – 65	2
46 – 50	12	66 – 70	2

Find the Mean Deviation from the Mean and Standard Deviation.

Solution. (a) Let us prepare the following table to calculate mean deviation :

Yield Per Hectare	No. of fields	Mid-value x	fx	$ d =$ $ x - 50 $	$f d $
31 – 35	2	33	66	17	34
36 – 40	3	38	114	12	36
41 – 45	8	43	344	7	56
46 – 50	12	48	576	2	24
51 – 55	16	53	848	3	48
56 – 60	5	58	290	8	40
61 – 65	2	63	126	13	26
66 – 70	2	68	136	18	36
$n = 50$		$\sum fdx = 2500$			$\sum f d = 300$

$$\text{Now A.M.} = \frac{\sum fx}{\sum f} = \frac{2500}{50} = 50.$$

$$\text{Mean Deviation} = \frac{\sum f|d|}{\sum f} = \frac{300}{50} = 6.$$

(b) **Calculation of Standard Deviation**

Yield Per Hectare	No. of fields	Mid-value x	$d = \frac{x - 53}{5}$	fd	fd^2
31 – 35	2	33	-4	-8	32
36 – 40	3	38	-3	-9	27
41 – 45	8	43	-2	-16	32
46 – 50	12	48	-1	-12	12
51 – 55	16	53	0	0	0

56 – 60	5	58	1	5	5
61 – 65	2	63	2	4	8
66 – 70	2	68	3	6	18
$n = 50$			$\sum fd = -30 \quad \sum fd^2 = 134$		

$$\text{Now } S.D. = i \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} = 5 \times \sqrt{\frac{134}{50} - \left(\frac{-30}{50}\right)^2}$$

$$= 5 \times \sqrt{\frac{134}{50} - \frac{900}{2500}} = 5 \times \sqrt{\frac{116}{50}} = 5 \times 1.5 = 7.5.$$

Example 10. In a study on patients, the following data was obtained. Find the standard deviation of the data :

Age (in yrs.)	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
Number of cases	1	0	1	17	17	38	9	3

Solution.

Age (in years)	No. of cases	Mid-value x	$d = \frac{x - 44.5}{10}$	fd	fd^2
10 – 19	1	14.5	– 3	– 3	9
20 – 29	0	24.5	– 2	0	0
30 – 39	1	34.5	– 1	– 1	1
40 – 49	10	44.5	0	0	0
50 – 59	17	54.5	1	17	17
60 – 69	38	64.5	2	76	152
70 – 79	9	74.5	3	27	81
80 – 89	3	84.5	4	12	48
$N = 79$			$\sum fd = 128 \quad \sum fd^2 = 308$		

$$\begin{aligned}
 \text{Standard Deviation} &= i \times \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \\
 &= 10 \times \sqrt{\frac{308}{79} - \left(\frac{128}{79}\right)^2} \\
 &= \frac{10}{79} \sqrt{24332 - 15384} = \frac{10}{79} \times 89.15 = 11.28.
 \end{aligned}$$

Relative measures of dispersion :

The four measures – Range, Quartile Deviation, Mean Deviation and Standard Deviation, are expressed in the same units as the original observations, and are called Absolute measures of variability. So, they can not be used for comparing the variability of two or more distributions given in different units. In order to meet such situations, the Relative measures of variability have been introduced which are independent of the units of measurement. There are 3 such measures –

- i) Coefficient of Variation = $100 \times \frac{\text{S.D.}}{\text{Mean}}$
- ii) Coefficient of Quartile Deviation = $100 \times \frac{\text{Quartile Deviation}}{\text{Median}}$
- iii) Coefficient of Mean Deviation = $100 \times \frac{\text{Mean Deviation}}{\text{Mean or Median}}$

Among these, coefficient of variation is the most important and is used in almost all cases.

The relative measures of variability may also be used to measure the precision of observations, although given in the same unit.

5.8 Exercises

1. Find the mean deviation about median from the following data :
46, 79, 26, 85, 39, 65, 99, 29, 56, 72.

2. Find the mean deviation for the following frequency distribution :

Variable :	3	5	7	9	11	13
Frequency :	2	7	10	9	5	1

3. Find the standard deviation for the distribution given below :

$x :$	1	2	3	4	5	6	7
Frequency :	10	20	30	35	14	10	2

4. Compute the arithmetic mean, standard deviation and mean deviation about the mean for the following data :

Sources :	4-5	6-7	8-9	10-11	12-13	14-15	Total
$f :$		4	10	20	15	8	3 60

Answers

1. 20.4
2. 2.02
3. 1.4
4. 9.23, 2.48, 2.03

6. Moments, Skewness & Kurtosis

6.1 **Moments :** Given n observations x_1, x_2, \dots, x_n and an arbitrary constant A ,

$\frac{1}{n} \sum (x - A)$ is called the 1st moment about A .

$\frac{1}{n} \sum (x - A)^2$ is called the 2nd moment about A .

$\frac{1}{n} \sum (x - A)^3$ is called the 3rd moment about A .

and so on. Let us denote these moments successively by m'_1, m'_2, m'_3 , etc.

$$\text{Then } m'_1 = \sum (x - A)/n = (\sum x - \sum A)/n = \frac{\sum x}{n} - A = \bar{x} - A.$$

i.e. the first moment about \bar{x} equals $(\bar{x} - \bar{x})$.

Basically moments are two kinds : One is moments about zero (Raw moments) and another moments about mean (central moments).

Moments about zero (Raw moments) :

$$\text{1st moment about zero} = \frac{1}{n} \sum x = \bar{x}$$

$$\text{2nd moment about zero} = \frac{1}{n} \sum x^2$$

$$\text{3rd moment about zero} = \frac{1}{n} \sum x^3 \text{ and so on.}$$

Note that the 1st moment about zero is the mean \bar{x} .

Moments about mean (Central moments) :

$$\text{1st moment about mean} = \frac{1}{n} \sum (x - \bar{x}) = 0$$

$$\text{2nd moment about zero} = \frac{1}{n} \sum (x - \bar{x})^2 = \sigma^2$$

$$\text{3rd moment about zero} = \frac{1}{n} \sum (x - \bar{x})^3 \text{ and so on.}$$

These are usually denoted by m_1, m_2, m_3 , etc.

Note that the 1st central moment about mean is zero and the 2nd central moment is the variance σ^2 , $m_1 = 0$ and $m_2 = \sigma^2$. The 3rd central moment m_3 is used to measure skewness and the 4th central moment m_4 is used to measure Kurtosis.

6.2 Skewness :

A frequency distribution is said to be 'symmetrical', if the frequencies are symmetrically distributed about mean, i.e., when values of the variable equi distant from mean have equal frequencies.

As for example for symmetrical distribution

$x :$	10	15	20	25	30
$f :$	3	7	16	7	3

In general, however, frequency distributions are not symmetrical, some are slightly asymmetrical and some others may be highly asymmetrical.

As for example for asymmetrical distribution

$x :$	5-8	9-12	13-16	17-20	21-24
$f :$	7	18	23	16	7

The word 'Skewness' is used to denote the 'extent of of asymmetry' in the data. When the frequency distribution is not symmetrical, it is said to be 'Skew'. The word 'skewness' literally denotes asymmetry, or 'lack of symmetry' and 'skey' denotes 'asymmetrical'. A symmetrical distribution has therefore zero skewness. Skewness may also be positive or negative.

Skewness is measured by the following formulae :

(1) Pearson's first measure –

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

(2) Pearson's second measure –

$$\text{Skewness} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

(3) Bowleys' measure –

$$\begin{aligned} \text{Skewness} &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \\ &= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \end{aligned}$$

where Q_1, Q_2, Q_3 denote the first, second and third quartiles of the distribution

(4) Moment measure : $\text{Skewness}(\gamma_1) = \sqrt{\beta_1} = \frac{m_3}{\sigma_3}$

Example. Calculate the coefficient of skewness based on quartiles from the following :

Class interval :	10-15	15-20	20-25	25-30	30-40	40-50	50-60	60-70	Total
Frequency :	4	12	16	22	10	8	6	4	82

Solution. In order to compute skewness, we have to find first Q_1 , Q_2 and Q_3 i.e. values of the variable corresponding to cumulative frequencies $\frac{N}{4}$, $\frac{N}{2}$, $\frac{3N}{4}$ respectively, where N = total frequency.

Here total frequency $N = 82$. Therefore $\frac{N}{4} = 20.5$, $\frac{N}{2} = 41$, and $\frac{3N}{4} = 61.5$.

Class boundary	Cumulative frequency (less than)
10	0
15	4
20	16
Q_1	$\frac{N}{4} = 20.5$
25	32
Q_2	$\frac{N}{2} = 41$
30	54
Q_3	$\frac{3N}{4} = 61.5$
40	64
50	72
60	78
70	$82 = N$

Cumulative Frequency Distribution.

Applying simple interpolation

$$\frac{Q_1 - 20}{25 - 20} = \frac{20.5 - 16}{32 - 16}$$

$$\Rightarrow Q_1 = 21.4.$$

$$\text{Similarly } \frac{Q_2 - 25}{30 - 25} = \frac{41 - 32}{54 - 32}$$

$$\Rightarrow Q_2 = 27.045$$

$$\text{and similarly } \frac{Q_3 - 30}{40 - 30} = \frac{61.5 - 54}{64 - 54}$$

$$\Rightarrow Q_3 = 37.5$$

$$\begin{aligned} \text{Now skewness} &= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \\ &= \frac{37.5 - 2 \times 27.045 + 21.4}{37.5 - 21.4} \\ &= \frac{4.809}{16.1} = 0.2987 \end{aligned}$$

$$\therefore \text{Skewness} = 0.2987.$$

6.3 Kurtosis : Kurtosis refers to the degree of “peakedness” of the frequency curve. Two distributions may have the same average, dispersion and skewness; yet, in one there may be high concentration of values near the mode; showing a sharper peak in the frequency curve than in the other. this characteristic of the frequency distribution is known as “kurtosis”. The only measure of kurtosis is based on moments,

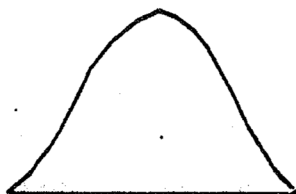
$$\text{viz. Kurtosis } (\gamma_2) = \frac{m_4}{\sigma^4} - 3 = \beta_2 - 3.$$

where m_4 and σ denote the fourth central moment and S.D. respectively, and $\beta_2 = \frac{m_4}{\sigma^4}$

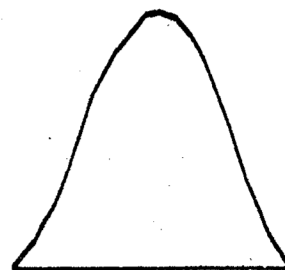
The following figures are shown different types of kurtosis.



(a) Platy kurtic



(b) Mesokurtic



(c) Leptokurtic

$\beta_2 < 3$, for platykurtic distribution

$\beta_2 = 3$, for mesokurtic distribution

$\beta_2 > 3$, for leptokurtic distribution

A distribution is said to be “platykurtic”, when γ_2 is negative; it is said to be “mesokurtic”, when $\gamma_2 = 0$, and “leptokurtic” when γ_2 is positive.

The frequency curve for a platykurtic distribution is relatively flat-topped, and for a leptokurtic distribution it has a relatively high peak. A mesokurtic distribution is of moderate peakedness.

6.4 EXERCISE

Ex. 1 Find the first four moments and the value of β_1 and β_2 from the following frequency distribution:

x :	21–24	25–28	29–32	33–36	37–40	41–44
f :	40	90	190	110	50	20

Also, find the measures of skewness and Kurtosis.

Ex. 1 Calculate the measure of skewness based on quartiles and median from the following data:

Variable:	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency:	358	2417	976	129	02	18	10

Ex. 3 Calculate the coefficient of skewness based on quartiles :

Class limits:	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89
Frequency:	5	9	14	20	25	15	8	4

Answers

1. Mean = 31.3, $m_1 = 23.04$, $m_2 = 26.11$, $m_3 = 1496.68$

$\beta_1 = .056$, $\beta_2 = 2.82$, Skewness(γ_1) = 0.24, Kurtosis(γ_2) = -0.18, -0.18

2. 0.13 (quartiles, 22.63, 26.73, 32.07)

3. -0.103 (quartiles 37.36, 50.30, 60.83)

7. **Unit Summary :** In this module, we have discussed the various types of measures of Central Tendency, the various types of measures of dispersion and various moments, skewness and kurtosis.

We have also discussed the merits and demerits of various types of measures of Central Tendency and the various types of measures of dispersion. We have solved lot of problems on the discussion topic in this module also.

8. Reference/ Suggested Further Readings :

1. N.G. Das, Statistical Methods, Vol. I & II, Das & Das Publishers, Calcutta.
2. P.N. Arora & P.K. Malhan, Bio Statistics, Himalaya Publishing House, Delhi, 1996.
3. I.A. Khan & A. Khanim, Fundamentals of Bio Statistics, Ukaaz Publications, Andhra Pradesh, 1994.
4. S.C. Gupta & Kappor, Fundamentals of Mathematical Statistics, Meerut, 2002.

---- 0 ----

**M.Sc. Course
in
Botany**

PART-I

Paper-VII

Module No. - 19(b)

BIOSTATISTICS

(INTRODUCTION TO FUZZY SETS)

Structure

1. Introduction
2. Objectives
3. Keywords
4. Probability
 - 4.1 Some Important terms and concepts
 - 4.2 Definition of Probability
 - 4.3 Theorems of Probability
 - 4.4 Theorem on Compound Probability
 - 4.5 Some Important Theoretical Distribution
 - 4.6 The Bionomial Distribution
 - 4.7 The Poisson Distribution
 - 4.8 The Normal Distribution
 - 4.9 Exercises
5. Correlation
 - 5.1 Correlation and Coefficient Correlation

- 5.2 Properties of Correlation Coefficient
- 5.3 Regression
- 5.4 Properties of Regression Coefficient
- 5.5 Exercises
- 6. Chi-Square Test
 - 6.1 Degrees of Freedom
 - 6.2 Chi-Square Distribution
 - 6.3 Properties of χ^2 -distribution
 - 6.4 Uses of χ^2 Test
 - 6.5 Conditions for using the chi-square test
 - 6.6 χ^2 -Test
 - 6.7 χ^2 -test for goodness of fit
 - 6.8 Exercises
- 7. Unit Summary
- 8. References/Suggested Further Readings

1. **Introduction :** This module contains three sections, in which first section deals with probability. In probability, we have discussed some important distributions, such as Binomial distribution, Poisson distribution and Normal distribution. In second section deals with correlation and regression analysis. We have also discussed the properties of correlation and regression coefficient. The last section describes the chi-square test and its application on biostatistics.

2. **Objectives :** To study this module, the reader will learn the following topics :

- (a) Probability and some important distributions.
- (b) Correlation and regression analysis.
- (c) Chi-square test and χ^2 -test for goodness of fit.

3. **Keywords :** Probability, Binomial, Poisson and Normal distribution, Correlation, Regression, Chi-square Test.

4. Probability :

Probability theory was originated from gambling theory. A large number of problems exist even today which are based on the game of chance, such as coin tossing, die throwing and playing cards. The utility of probability in business and economics is most emphatically revealed in the field of predictions for future. We have to anticipate, consequences of the each of these plans and finally we compare the results. Uncertainty plays an important role in business and probability is a concept which measures the degree of uncertainty and that of certainty also as a corollary. The probability when defined in the simplest way is *chance of occurrence of a certain event when expressed quantitatively*. The probability is defined in two different ways :

- (i) Mathematical (or a priori) definition
- (ii) Statistical (or empirical) definition

Before we study the probability theory in detail, it is appropriate to give the definitions of certain terms, which are essential for the study of *Probability theory*.

4.1 Some Important Terms and Concepts

1. **Random Experiment or Trial.** An experiment is characterized by the property that its observations under a given set of circumstances do not always lead to the same observed outcome but rather to different outcomes which follow a sort of statistical regularity. It is also called a *Trial*. For example tossing a coin, or throwing a die.

2. **Sample Space.** A set of all possible outcomes from an experiment is called a **Sample Space**.

Let us toss a coin. The result is either head or tail. Let 1 denote head and 0 denote tail. Mark the point 0, 1 on a straight line. Thus we get two different points 0 and 1 on a straight line. These points are called sample points or event points. For a given experiment there are different possible outcomes and hence different sample points. The collection of all such sample points is called a **Sample Space**. Toss two coins simultaneously, then the possible results are four pairs (0, 0), (1, 0), (0, 1), (1, 1), and they can be represented as points in coordinate plane. These points constitute a sample space of four sample points. Similarly, by tossing three coins simultaneously we get eight points which can be denoted by the triplets (0, 0, 0), (0, 1, 0), (0, 0, 1), (1, 0, 0), (1, 1, 0), (0, 1, 1), (1, 0, 1), (1, 1, 1). These

8 points form a sample space of 3 dimension. In general, in tossing n coins simultaneously we can have an n -dimensional sample space consisting of 2^n sample points.

3. **Discrete Sample Space.** A sample space whose elements are finite or infinite but countable is called a **discrete sample space**. For example, if we toss a coin as many times as we require for turning up one head then the sequence of points $S_1 = (1), S_2 = (0, 1), S_3 = (1, 0, 1), S_4 = (0, 0, 0, 1)$ etc., is a discrete sample space.

4. **Continuous Sample Space.** A sample space whose elements are infinite and uncountable or assume all the values on a real line R or on an interval of R is called **continuous sample space**. In this case the sample points build up a continuum, and the sample space is said to be continuous. For example, all the points on a line or all points on a plane is a sample space.

5. **Event.** A sub-collection of a number of sample points under the definite rule or law is called an **event**. For example, let us take a die. Let its faces 1, 2, 3, 4, 5, 6 be represented by $E_1, E_2, E_3, E_4, E_5, E_6$ respectively. Then all the E 's are sample points. Let E be the event of getting an even number on the die. Obviously $E = (E_2, E_4, E_6)$ which is a subset of the set $\{E_1, E_2, E_3, E_4, E_5, E_6\}$.

6. **Null Event.** An event having no sample point is called a null event and is denoted by ϕ .

7. **Simple Event.** An event consisting of only one sample point of a sample space is called a sample event.

For example, let a die be rolled once, and A be the event that face number 5 is turned up, then A is a simple event.

8. **Compound Events.** When an event is decomposable into a number of simple event, then it is called a compound event.

For example, the sum of the two numbers shown by the upper faces of the two dice is seven in the simultaneous throw of the two unbiased dice, is a compound event as it can be decomposable.

9. **Exhaustive Cases or Event.** It is the total number of all the possible outcomes of an experiment. For example, when we throw a die then any one of the six faces (1, 2, 3, 4, 5, 6) may turn up and, therefore, there are six possible outcomes. Hence, there are six exhaustive cases or events in throwing a die.

10. **Mutually Exclusive Events.** If in an experiment the occurrence of an event precludes or prevents or rules out the happening of all other events in the same experiment then these events are said to be **mutually exclusive events**. For example, in tossing a coin, the events **head** and **tail** are mutually exclusive, because if the outcome is head, then the possibility of getting a tail in the same trial is ruled out.

11. **Equally likely.** Events are said to be **equally likely** if there is no reason to expect any one in preference to other. For example, in throwing a die, all the six faces (1, 2, 3, 4, 5, 6) are equally likely to occur.

12. **Collectively Exhaustive Events.** The total number of events in a population exhausts the population. So they are known as **collectively exhaustive events**.

13. **Equally Probable Events.** If in an experiment all possible outcomes have an equal chance of occurrence, then such events are said to be **equally probable**. For example, in throwing a coin, the events **head** and **tail** have equal chances of occurrence, therefore, they are equally probable events.

14. **Favourable Cases.** The cases which ensure the occurrences of an event are said to be favourable to the events.

15. **Independent and Dependent Event.** When the experiments are conducted in such a way that the occurrences of an event in one trial does not have any effect on the occurrence of this or other events at a subsequent experiment, then the events are said to be independent. In other words, two or more events are said to be **independent** if the happening of any one does not depend on the happening of the other. Events which are not independent are called **dependent events**.

Illustration. If we draw a card in a pack of well shuffled cards and again draw a card from the rest of pack of cards (containing 51 cards), then the second draw is **dependent** on the first. But if on the other hand, we draw a second card from the pack by replacing the first card drawn, the second draw is known as **independent** of the first.

4.2 Definition of Probability

Classical definition of Probability. If an experiment has n mutually exclusive, equally likely and exhaustive cases, out of which m are favourable to the happening of the event A , then the probability

of the happening of A is denoted by $P(A)$ and is defined as

$$P(A) = \frac{m}{n} = \frac{\text{No. of cases favourable to } A}{\text{Total (Exhaustive) number of cases}}$$

Example 1. What is the probability of getting an even number in a single throw with a die?

Solution. The possible cases in the throw of a die are six, viz., 1, 2, 3, 4, 5, 6.

Favourable cases are those which are marked with 2, 4, 6 and these are three in number.

$$\therefore \text{Probability of getting an even number} = \frac{3}{6} = \frac{1}{2}.$$

Notes 1. Probability of an event which is certain to occur is 1 and the probability of an impossible event is zero.

2. The possibility of occurrence of any event lies between 0 and 1, both inclusive.

Example 2. What is the probability of getting tail in a throw of a coin?

Solution. When we toss a coin, there are two possible outcomes viz., Head or Tail. In this case the number of possible cases $n = 2$.

No. of favourable cases $= m = 1$.

(\because The outcome of tail is a favourable event)

$$\therefore \text{Probability of getting a tail} = \frac{m}{n} = \frac{1}{2}.$$

Example 3. A bag contains 6 white balls, 9 black balls. What is the probability of drawing a black ball?

Solution. The total number of equally likely and exhaustive cases

$$= n = 6 + 9 = 15.$$

Number of favourable cases $= m = 9$.

(\because Number of black balls $= 9$)

$$\text{Probability of drawing a black ball} = \frac{9}{15} = \frac{3}{5}.$$

Example 4. What is the probability that if a card is drawn at random from an ordinary pack of cards, it is (i) a red card, (ii) a club, (iii) one of the court cards (Jack of Queen or King).

Solution. No. of exhaustive cases = 52.

(i) There are 26 red cards and 26 black cards in an ordinary pack.

∴ Favourable cases = $n = 26$ (number of red cards).

∴ Probability of getting a red card = $\frac{26}{52} = \frac{1}{2}$.

(ii) Number of clubs in a pack = 13.

∴ Favourable cases = 13.

∴ Probability of getting a club = $\frac{13}{52} = \frac{1}{4}$.

(iii) There are $4 \times 3 = 12$ court cards in a pack of cards.

∴ Number of favourable cases = $m = 12$.

Number of exhaustive cases = $n = 52$.

∴ Probability of getting a court card = $\frac{12}{52} = \frac{3}{13}$.

Example 5. What is the probability that a leap year, selected at random, will have 53 Sundays?

Solution. There are 366 days in a leap year and it has 52 complete weeks and 2 days over. These two extra days can occur in following possible ways.

- (i) Sunday and Monday;
- (ii) Monday and Tuesday;
- (iii) Tuesday and Wednesday;
- (iv) Wednesday and Thursday;
- (v) Thursday and Friday;
- (vi) Friday and Saturday;
- (vii) Saturday and Sunday.

∴ No. of exhaustive cases = 7.

No. of favourable cases = 2.

{ ∴ There are two cases which have Sunday [(i) and (vii)] }

$$\therefore \text{Probability} = \frac{2}{7}.$$

Example 6. A card is drawn from an ordinary pack of playing cards and a person bets that it is a spade or an ace. What are the odds against his winning this bet?

Solution: In a pack of 52 cards, 1 card can be drawn in 52 ways. Since there are 13 spades and 3 aces (one ace is also present in spade), therefore, the favourable cases = $m = 13 + 3 = 16$.

No. of exhaustive cases = $n = 52$.

$$\text{Probability of getting a spade or an ace} = \frac{16}{52} = \frac{4}{13} = \frac{4}{9+4}.$$

\therefore Odds against winning the bet are 9 to 4.

Statistical or Empirical definition. Von Mises has given the following statistical or empirical definition of probability.

"If the experiment be repeated a large number of times, under essentially identical conditions, the limiting value of the ratio of the number of times the event A happens to the total number of trials of the experiment as the number of trials increases indefinitely is called the probability of happening of the event A ."

Symbolically. Let $P(A)$ denote the probability of the occurrence of A . Let m be the number of times in which an event A occurs in a series on n trials, then

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n},$$

Provided the limit is finite and unique.

4.3 Theorems of Probability

There are two important theorems probability, namely

1. The Addition theorem, or the theorem on Total Probability.
2. The Multiplication theorem or theorem on Compound Probability.

Additional theorem of the theorem on Total Probability

Statement. If the events are mutually exclusive, then the Probability of happening of any one of them is equal to the sum of the probabilities of the happening of the separate events, i.e., in

other words if $E_1, E_2, E_3, \dots, E_n$ be n events and $P(E_1), P(E_2), \dots, P(E_n)$, be their respective probabilities, then

$$P(E_1 + E_2 + E_3 + \dots + E_n) = P(E_1) + P(E_2) + P(E_3).$$

Example 7. A die is rolled. What is the probability that a number 1 or 6 may appear on the upper face?

Solution. The probability of appearing the number 1 on the upper face $= \frac{1}{6}$.

\therefore The probability 1 or 6 may appear on the face $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Example 8. If the probability of the horse A winning the race is $\frac{1}{5}$ and the probability of the horse B winning the same race is $\frac{1}{6}$, what is the probability that one of the horses will win the race?

Solution. Probability of the winning of the horse A $= \frac{1}{5}$. Probability of the winning of the horse B $= \frac{1}{6}$.

$$\therefore P(A + B) = P(A) + P(B) = \frac{1}{5} + \frac{1}{6} = \frac{11}{30}.$$

Multiplicative theorem or Theorem on Compound Probability

Before we proceed further in stating and proving this theorem, we require the following definitions;

Simple and compound events. A single event is called a **simple event**, whereas when two or more then two simple events occur in connections with each other, then their simultaneous occurrences is called a **compound event**. If A and B are two simple events, the simultaneous occurrence A and B is called a compound event and is denoted by AB.

Conditional Probability. The probability of the happening of an event A, when it is known

that B has already happened, is called the **conditional probability** of A and is denoted by $P(A/B)$;

i.e., $P(A/B) \Rightarrow$ conditional probability of A given that B has already occurred.

Similarly, $P(B/A) \Rightarrow$ conditional probability of B given that A has already happened.

Mutually independent Events. An event A is said to be independent of the event B if $P(A/B) = P(A)$, i.e., the probability of the happening of A is independent of the happening of B .

4.4 Theorem on Compound Probability

Theorem on Compound Probability. *The probability of the simultaneous occurrence of the two events A and B is equal to the probability of one of the events multiplied by the conditional probability of other, given the occurrence of the first, i.e.,*

$$P(AB) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$$

Example 9. *A card is drawn from a pack of 52 cards and then a second is drawn. What is the probability that both the cards drawn are queen?*

Solution. First draw. Probability of getting a queen $= \frac{4}{52} = \frac{1}{13}$.

Second draw. After drawing the first queen, we are left with 51 cards having 3 queens.

Probability of getting a queen in second draw $= \frac{3}{51} = \frac{1}{17}$.

Probability that both the cards are queen $= \frac{1}{13} \times \frac{1}{17} = \frac{1}{221}$.

Example 10. *A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both the balls drawn are black.*

Solution. Let $P(A)$, $P(B/A)$ denote the probability of drawing a black ball in the first and second attempt respectively.

\therefore Probability of drawing a black ball in the first attempt is $P(A)$,

$$\text{where } P(A) = \frac{\text{Favourable cases}}{\text{Total exhaustive cases}} = \frac{3}{5+3} = \frac{3}{8}$$

Probability of drawing the second black ball given the first ball drawn is black is

$$P(B/A) = \frac{\text{Favourable cases}}{\text{Total exhaustive cases}} = \frac{2}{5+2} = \frac{2}{7}$$

Probability that both the balls drawn are black is

$$P(AB) = P(A)P(B/A) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

Additional theorem for compatible events

Theorem. The probability of the occurrence of at least one of the events A and B is given by

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

Happening of at least one Event

Let A and B two independent events and p_1 and p_2 be the probabilities of their happening, then the chance that both A and B happen is $p_1 \times p_2$. Also the probabilities of not happening of A and B are $1 - p_1$, $1 - p_2$ respectively. The probability that both do not happen is $(1 - p_1) \times (1 - p_2)$.

Now the chance that at least one of them happens :

$$p_1 p_2 + p_1(1 - p_2) + p_2(1 - p_1) = 1 - (1 - p_1)(1 - p_2)$$

Thus the above result can be generalised for n events E_1, E_2, \dots, E_n with p_1, p_2, \dots, p_n respective probabilities. Then the probability that at least one of them happens is

$$1 - (1 - p_1)(1 - p_2) \dots (1 - p_n)$$

Example 11. 4 coins are tossed. Find the probability that at least one head turns up.

Solution. Probability of getting a head in a toss of a coin = $1/2$.

Probability of getting a tail in each case = $1/2$.

Probability of getting a tail in all the four cases = $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$.

\therefore Probability of getting at least one head = $1 - \frac{1}{16} = \frac{15}{16}$.

Example 12. In a throw of 3 dice, find the probability that at least one die shows up 1.

Solution. p = Probability of getting 1 in a throw of a die $= \frac{1}{6}$.

$$\therefore q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}.$$

Probability of getting at least one die shows up

$$= 1 - q^3 = 1 - \left(\frac{5}{6}\right)^3 = \frac{91}{216}.$$

Example 13. A problem in Mathematics is given to three students Dayanand, Ramesh, Naresh whose chances of solving it are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ respectively. What is the probability that the problem will be solved?

Solution. The probabilities of Dayanand, Ramesh and Naresh solving the problem are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ respectively.

\therefore The probabilities of Dayanand, Ramesh, Naresh not solving the problem are $1 - \frac{1}{2} = \frac{1}{2}; 1 - \frac{1}{3} = \frac{2}{3}; 1 - \frac{1}{4} = \frac{3}{4}$ respectively.

\therefore The probability that the problem is not solved by any of them is $= \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{1}{4}$.

The probability that the problem will be solved by at least one of them $1 - \frac{1}{4} = \frac{3}{4}$.

Example 14. A card is drawn at random from a well shuffled pack of 52 cards. Find the probability of getting a two of heart of diamond.

Solution. Since there is only one two of hearts and only one two of diamonds, therefore, the probability of getting a two of hearts $= \frac{1}{52}$ and the probability of getting a two of diamonds $= \frac{1}{52}$.

Since these are mutually exclusive cases, therefore, the probability of getting a two of heart or a two of diamonds $= \frac{1}{52} + \frac{1}{52} = \frac{2}{52} = \frac{1}{26}$.

Example 15. A man and his wife appear for an interview for two posts. The probability of the husband's selection is $\frac{1}{7}$ and that of the wife's selection is $\frac{1}{5}$. What is the probability that only one of them will be selected?

Solution. The probability that husband is not selected $= 1 - \frac{1}{7} = \frac{6}{7}$

The probability that wife is not selected $= 1 - \frac{1}{5} = \frac{4}{5}$.

Probability that only husband is selected $= \frac{1}{7} \times \frac{4}{5} = \frac{4}{35}$.

Probability that only wife is selected $= \frac{1}{5} \times \frac{6}{7} = \frac{6}{35}$.

Probability that only one of them is selected $= \frac{4}{35} + \frac{6}{35} = \frac{10}{35} = \frac{2}{7}$.

Example 16. A speaks truth in 75% and B in 80% of the cases. In what percentage of cases are they likely to contradict each other narrating the same incident?

Solution. Let $P(A)$, $P(B)$ be the probability of A and B speaking the truths, then

$$P(A) = \frac{75}{100} = \frac{3}{4}, P(B) = \frac{80}{100} = \frac{4}{5}$$

$$P(\bar{A}) = P(A \text{ tells a lie}) = 1 - P(A) = 1 - \frac{3}{4} = \frac{1}{4}$$

$$P(\bar{B}) = P(B \text{ tells a lie}) = 1 - P(B) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$\text{Now } P(A \text{ and } B \text{ will contradict}) = P(A)P(\bar{B}) + P(B)P(\bar{A})$$

$$= \frac{3}{4} \times \frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = \frac{7}{20} = 35\%$$

Example 17. Kamal and Monica appeared for an interview for two vacancies. The probability of Kamal's selection is $\frac{1}{3}$ and that of Monica's selection is $\frac{4}{5}$. Find the probability

that only one of them will be selected.

Solution. Let p_1 and p_2 be the probabilities of Kamal and Monica selection.

$$\text{Then } p_1 = \frac{1}{3} \text{ and } q_1 = \frac{2}{3};$$

$$p_2 = \frac{1}{5} \text{ and } q_2 = \frac{4}{5}.$$

Now the required probability is given by $p_1q_2 + q_1p_2$.

$$= \frac{1}{3} \times \frac{4}{5} + \frac{2}{3} \times \frac{1}{5}$$

$$= \frac{4}{15} + \frac{2}{15} = \frac{6}{15} = \frac{2}{5}.$$

4.5 Some Important Theoretical Distributions

The distributions which are based on actual data of experiment are called observed frequency distribution. It is sometimes possible, by assuming a certain hypothesis, to derive under mathematically the frequency distribution of a certain population. Such distributions are known as theoretical distributions. In other words, a theoretical distribution is the frequency distribution of certain events in which frequencies are obtained by mathematical computations. Some of the important theoretical distributions are :

1. Binomial distribution.
2. Poisson distribution.
3. Normal distribution.

There are two types of theoretical distribution

- (i) Discrete distribution.
- (ii) Continuous distribution.

Binomial and Poisson distributions are known as discrete distribution. Normal distributions is an example of a continuous distribution.

4.6 Binomial distribution :

Binomial distribution is a probability which is obtained when the probability p of the happening

of an event in same in all the trials, and there are only two events in each trial.

For example, the probability of getting a head, when a coin is tossed a number of times, must remain same in each trial, i.e., $\frac{1}{2}$.

Let an experiment consisting of n trials be performed and let the occurrence of an event in any trial be called a success and its non occurrence of failure. Let p be the probability of success and q be the probability of the failure in a single trial, where $q = 1 - p$, so that $p + q = 1$.

Let us assume that the trials are independent and the probability of success is same in each trial. Let us claim that we have n trials, then the probability of happening of an event r times and failing $(n - r)$ times in any specified order is $p^r q^{n-r}$. But the total number of ways in which the event can happen r times exactly in n trials is n_c . These n_c ways are equally likely, mutually exclusive and exhaustive.

Therefore, the probability r successes and $(n - r)$ failures in n trials in any order, whatsoever is $n_c p^r q^{n-r}$. It can also be expressed in the form

$$p(r) = n_c p^r q^{n-r}, r = 0, 1, 2, \dots, n.$$

where $p(r)$ is the probability distribution of the number of successes. Giving different values to r , i.e., putting $r = 0, 1, 2, \dots, n$. We get the corresponding probabilities $n_c q^n, n_c p q^{n-1}, n_c p^2 q^{n-2}, \dots, n_c p^n$, which are the different terms in the Binomial expansion of $(q+p)^n$.

As a result of it, the distribution $p(r) = n_c p^r q^{n-r}$ is called Binomial Probability distribution. The two independent constants, viz., n and p in the distribution are called the parameters of distribution.

Again if the experiment (each consisting n trials) be repeated N times, the frequent function of the Binomial distribution is given by -

$$f(r) = N p(r) = N n_c p^r q^{n-r}, r = 0, 1, 2, \dots, n.$$

where $p + q = 1$, which is also called the Binomial frequency distribution.

Mean and Variance of Binomial distribution :

The probability distribution of Binomial distribution for r success in n trials is given by

$$p(r) = n_c p^r q^{n-r}$$

Its mean $\mu = np$ and variance $\sigma^2 = npq$.

Ex. 1 A die is thrown three times. Getting a '3' or a '6' is considered a success. Find the probability of at least two successes.

Solution. Let $p(x)$ be the probability of getting the number x in a toss of a dice. Then

$$P(3) = \frac{1}{6}, P(6) = \frac{1}{6}.$$

$$\therefore P(3 \text{ or } 6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = p.$$

$$q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}.$$

Now $p(r) = n_c p^r q^{n-r}$. Here $n = 3$.

$$\therefore P(2) = 3_c \left(\frac{2}{3}\right)^{3-2} \left(\frac{1}{3}\right)^2 = \frac{2}{9}.$$

$$P(3) = 3_c \left(\frac{2}{3}\right)^{3-3} \left(\frac{1}{3}\right)^3 = \frac{1}{27}.$$

$$P(\text{at least 2 successes}) = P(2) + P(3) = \frac{2}{9} + \frac{1}{27} = \frac{7}{27}.$$

Ex. 2 There are 64 beds in a garden and 3 seeds of a particular type of flower are down in each bed. The probability of a flower being white is $\frac{1}{4}$. Find the number of beds with 3, 2, 1 and 0 white flowers.

Solution. The probability p of a white flower = $\frac{1}{4}$.

$$\therefore q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}.$$

Here $n = 3$, $N = 64$,

$$\therefore f(r) = {}^3C_r \left(\frac{1}{4}\right)^r \left(\frac{3}{4}\right)^{3-r}$$

Number of beds with zero white flower $N_f(0)$

$$= 64 {}^3C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{3-0} = 27$$

$$\text{Beds with 1 white flower} = 64 f(1) = 64 {}^3C_1 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^{3-1} = 27$$

$$\text{Beds with 2 white flowers} = 64 f(2) = 64 {}^3C_2 \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^{3-2} = 9$$

$$\text{Beds with 3 white flowers} = 64 f(3) = 64 {}^3C_3 \left(\frac{1}{4}\right)^3 = 1$$

Ex. 3 If the probability of a defective bolt is 0.1, find the mean and standard deviation for the distribution of defective bolt in a total of 500.

Solution. Let $(q + p)^n$, $q + p = 1$, be the Binomial distribution.

Here we are given that $p = 0.1$, $n = 500$,

$$\text{Mean} = np = 500 \times 0.1 = 50$$

$$\text{Now } p = 0.1, q = 1 - p = 1 - 0.1 = 0.9$$

$$\text{Variance} = npq = 500 \times 0.9 \times 0.1 = 45$$

$$\text{Standard deviation} = \sqrt{45} = 6.7$$

4.7 The Poisson Distribution :

Poisson distribution is the distribution of a random variable taking values 0, 1, 2, ... and is useful in a wide variety of applications dealing with counts. However, not all variables arising from counting applications follow this distribution.

The probability distribution of a random variable x is said to have a Poisson distribution if it takes only non-negative values and if its distribution is given by

$$P(X = x) = \frac{e^{-\mu} \mu^x}{x!}, x = 0, 1, 2, \dots$$

where x is the parameter.

The quantity e in the formula above is a constant with value approximately equal to 2.71828.

The mean and variance of Poisson distribution are equal and is equal to μ .

Ex. 4. Suppose a book of 585 pages contains 43 typographical errors. If these errors are randomly distributed throughout the book. What is the probability that 10 pages, selected at random, will be free from errors ?

$$[Use e^{-0.735} = 0.4795]$$

Solution. Here $n = 10$, $p = \frac{43}{585} = 0.0735$, mean $= np = 0.735$.

Thus Poisson distribution is given by

$$P(x) = \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-0.735} (0.735)^x}{x!}$$

$$\begin{aligned} \text{Probability of zero error} &= P(0) = \frac{e^{-0.735} (0.735)^0}{0!} \\ &= 0.4795. \end{aligned}$$

Ex. 5. The mortality rate for a certain disease is 7 in 1000. What is the probability for just 2 deaths on account of this disease in a group of 400? (Given $e^{-2.8} = 0.06$).

Solution. Here $p = \frac{7}{1000}$ and $n = 400$

$$\therefore \mu = np = 400 \times \frac{7}{1000} = 2.8$$

Now the required probability is given by

$$P(2) = \frac{\mu^2}{2!} e^{-\mu} = e^{-2.8} \frac{(2.8)^2}{2!} = 0.235.$$

4.8 Normal Distribution : Normal distribution is a continuous distribution. Normal distribution is also a limiting form of the Binomial distribution under the following conditions :

- (i) n , the number of trials is indefinitely large, i.e., in other words, $n \rightarrow \infty$.
- (ii) neither p nor q is very small.

The probability function of the normal distribution with mean at the origin is given by

$$P(x) = y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, -\infty < x \leq \infty,$$

where σ is the standard deviation.

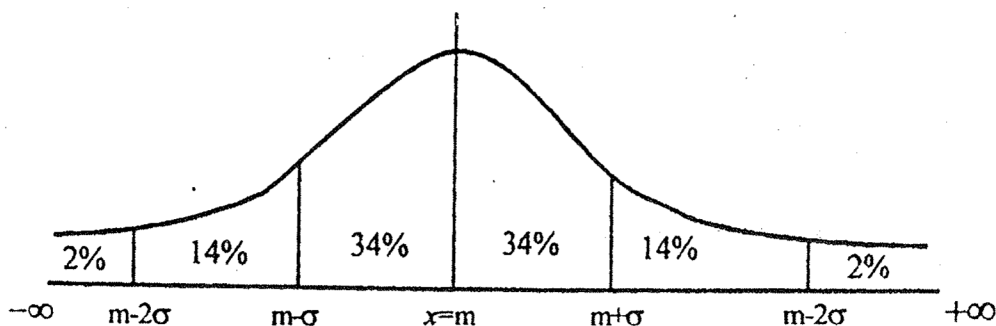
The probability function of a normal distribution with mean m is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, -\infty < x < \infty.$$

Properties of Normal Curve :

The normal probability curve with mean m and standard deviation σ has the following properties:

1. The equation of the curve is $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, -\infty < x < \infty$, and it is bell shaped.
2. The curve is symmetrical about the line $x = m$ and x ranges from $-\infty$ to ∞ .
3. Mean, Median and Mode are coincide at $x = m$.
4. It can be shown that it has Arithmetic mean $= m$ and variance $= \sigma^2$.
5. The total area under the normal curve is equal to unity and the if is shown in the following figures.



4.9 EXERCISES

1. The probability of A, B, C solving a problem are $\frac{1}{3}, \frac{2}{7}, \frac{3}{8}$ respectively. If all the three try to solve the problem simultaneously, find the probability that exactly one of them will solve it.
2. A bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black balls. If one ball is drawn from each bag, find the probability that (i) both are white (ii) both are black (iii) One is white and one is black.
3. Four numbers are selected at random. Show that the probability that the last digit in the product of these four numbers will be 1, 3, 7 or 9 is $\frac{16}{625}$.
4. The probability of hitting a target by three men are $\frac{1}{2}, \frac{1}{3}$ and $\frac{1}{4}$ respectively. Find the probability that one and only one of them will hit the target when they fire simultaneously.

Answer

1. $\frac{25}{56}$
2. $\frac{13}{24}$
4. $\frac{11}{24}$

5. Correlation

Before we study the correlation analysis we introduce the concept of "covariance" between two quantitative variables x and y . Let the corresponding values of the two variables x and y on the given set of n units of observations be given by the ordered pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots (x_n, y_n)$. Thus the covariance between x and y is denoted by $\text{cov}(x, y)$. It is defined as

$$\text{Cov}(x, y) = \frac{(x_1 - \bar{x}_1)(y_1 - \bar{y}_1) + (x_2 - \bar{x}_2)(y_2 - \bar{y}_2) + \dots + (x_n - \bar{x}_n)(y_n - \bar{y}_n)}{n}$$

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

where \bar{x} and \bar{y} are the means of x and y respectively.

$$\text{i.e. } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}, \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

5.1 Correlation and Coefficient of Correlation :

Correlation : Correlation may be defined as a tendency towards interrelation variation and the coefficient of correlation is a measure of such a tendency, i.e., the degree to which the two variables are interrelated is measured by a coefficient which is called the coefficient of correlation. It gives the degree of correlation.

Definition : The relationship between two variables such that a change in one variable results in a positive or negative change in the other and also a greater change in one variable results in corresponding greater or smaller change in the other variable is known as "correlation". The coefficient of correlation between the two variables x, y is generally denoted by r or r_{xy} or $e(x, y)$ or e and is defined as

$$e = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

5.2 Properties of Correlation Coefficient :

1. It is a measure of the closeness of a fit in a relative sense.
2. Correlation coefficient lies between -1 and $+1$, i.e. $-1 \leq r \leq 1$.
3. The correlation is perfect and positive if $r = 1$ and it is perfect and negative if $r = -1$.
4. If $r = 0$, then there is no correlation between the two variables and thus the variables are said to be independent.

Ex. 1. Calculate the correlation coefficient from the following data.

$$\sum_{i=1}^{100} x_i = 280, \sum_{i=1}^{100} y_i = 60, \sum_{i=1}^{100} x_i^2 = 2384, \sum_{i=1}^{100} y_i^2 = 117, \sum_{i=1}^{100} x_i y_i = 438$$

Solution. Correlation coefficient $r = \frac{Cov(x, y)}{\sigma_x \sigma_y}$

Here $\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 = \frac{2384}{100} - \left(\frac{280}{100} \right)^2 = 16$

$$\sigma_y^2 = \frac{\sum y^2}{n} - \left(\frac{\sum y}{n} \right)^2 = \frac{117}{100} - \left(\frac{60}{100} \right)^2 = 0.81$$

$$Cov(x, y) = \frac{\sum xy}{n} - \left(\frac{\sum x}{n} \right) \left(\frac{\sum y}{n} \right) = \frac{438}{100} - \frac{280}{100} \times \frac{60}{100} = 2.7$$

Substituting these values,

$$r = \frac{2.7}{\sqrt{16 \times 0.81}} = 0.75.$$

Ex. 2. Find the coefficient of correlation from the following data.

$x:$	65	63	67	64	68	62	70	66
$y:$	68	66	68	65	69	66	68	65

Solution. Since the correlation coefficient is unaffected by change of origin (and also scale), let us change the origins of x and y to 65 and 67 respectively i.e., write $x = X - 65$, $y = Y - 67$.

Table for Calculations for Correlation Coefficient

x	y	$x = x - 65$	$y = y - 67$	x^2	y^2	xy
65	68	0	1	0	1	0
63	66	-2	-1	4	1	2
67	68	+2	1	4	1	2
64	65	-1	-2	1	4	2
68	69	3	2	9	4	6
62	66	-3	-1	9	1	3
70	68	5	1	25	1	5
66	65	1	-2	1	4	-2
525	535	5	-1	53	17	18

$$\text{Now } \sigma_x^2 = \frac{53}{8} - \left(\frac{5}{8}\right)^2 = \frac{399}{64}$$

$$\sigma_y^2 = \frac{17}{8} - \left(-\frac{1}{8}\right)^2 = \frac{135}{64}$$

$$\text{Cov}(x, y) = \frac{18}{8} - \left(\frac{5}{8}\right)\left(-\frac{1}{8}\right) = \frac{149}{64}$$

$$\begin{aligned} \therefore \text{Correlation coefficient } r &= \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \\ &= \frac{\frac{149}{64}}{\sqrt{\frac{399}{64} \times \frac{135}{64}}} = 0.64. \end{aligned}$$

5.3 Regression : We know that the correlation studies the relationship between two variables x and y . In this section we shall consider the related problem of prediction “estimation” of the value of variable from a known value of variable from a known value of variable to which it is related. This would be discussed by means of regression lines.

Regression shows a relationship between the average values of two variables. Thus regression is very helpful in estimating and predicating the average value of one variable for a given value of the other variable. The estimate or prediction may be made with the help of regression line which shows the average value of one variable x for a given value of the other variable y . The best average value of one variable associated with the given value of the other variable may also be estimated or predicated by means of an equation and the equation is known as a Regression equation.

The two lines of regression are

1. Regression equation of x on y is

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

[If estimates x for a given value of y].

2. Regression equation of y on x is

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

[It estimates y for a given value of x].

where x = value of x , y = value of y

\bar{x} = Mean of x series

\bar{y} = Mean of y series

σ_x = Standard deviation of x

σ_y = Standard deviation of y

r = Correlation coefficient between x & y .

It is important to note that the regression equation of x on y should be used for predictions or timing the value of x for a given value of y and the regression equation of y on x should be used for predicting or estimating the value of y for a given value of x .

Regression Coefficients :

The regression coefficient of y on x is $b_{yx} = r \frac{\sigma_y}{\sigma_x}$ and that of x on y is $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

5.4 Properties of Regression Coefficients :

1) The coefficients of correlation is the geometric mean of the coefficients of regression
i.e., $r = \sqrt{b_{xy} \times b_{yx}}$.

2) If one of the regression coefficients is greater than unity, then the other is less than unity.

3) Arithmetic mean of the regression coefficient is greater than the correlation coefficient.

4) Regression coefficient are independent of change of origin but not of scale.

Ex. 3. From the data given below estimate the most likely height of a father son's height is 65"

Father's : Mean height is 67" and a. S.D. of 3.5"

Son's : Mean height is 65" and a S.D. of 2.5"

The coefficient of correlation between the heights of fathers and sons is + 0.8.

Estimate the height of father when height of son is 70".

Solution. Let x and y be the variables corresponding to the heights of sons and fathers.

\therefore Then $\bar{x} = 65$, $\bar{y} = 67$, $\sigma_x = 2.5$, $\sigma_y = 3.5$, $r_{xy} = 0.8$.

$$\text{Now } b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{(0.8)(3.5)}{2.5} = 1.12.$$

\therefore The equation of line of regression of y on x is

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

$$\Rightarrow y - 67 = 1.12(x - 65)$$

$$\Rightarrow y = 1.12x - 5.8 \dots (1)$$

Height of a father whose son's height is 70" is = estimate of y for $x = 70$.

Putting $x = 70$ in (1), we get

$$y = 1.12 \times 70 - 5.8 = 72.6''.$$

Ex. 4. Given the following results of the height and weight of 1,000 students :

$\bar{y} = 68$ inches, $\bar{x} = 150$ lbs, $r = 0.60$, $\sigma_y = 2.50$ inches,

$\sigma_x = 10$ lbs. Amit weights 100 lbs. Sumit is 5 feet tall. Estimate the height of Amit from his weight and the weight of Sumit from his height.

Solution. Here Height = y , weight = x , $\bar{y} = 68$ inches

$\bar{x} = 150$ lbs, $\sigma_x = 2.50$ inches, $\sigma_y = 10$ lbs, $r = 0.60$

(1) the regression equation of y on x is

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

$$\Rightarrow y - 68 = 0.60 \times \frac{2.50}{10} (x - 150)$$

$$\Rightarrow y = 0.15x + 63.$$

When Amit's weight $x = 100$ lbs, his height

$$y = 0.15 \times 100 + 63 = 15 + 63 = 78 \text{ inches.}$$

∴ Required height of Amit = 68 inches

(2) The regression equation of x on y is

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

$$\Rightarrow x - 150 = 0.60 \times \frac{10}{2.50} (y - 68)$$

$$\Rightarrow x = 2.4y - 13.2.$$

When Sumit's height $y = 5$ feet = 60 inches.

Then His weight $x = 2.4 \times 60 - 13.2 = 130.8$ lbs.

∴ Required weight of Sumit = 130.8 lbs.

Ex. 5. Find the regression of x on y from the following data :

$$\sum x = 24, \sum y = 44, \sum xy = 306, \sum x^2 = 164, \sum y^2 = 574, n = 4.$$

Find the value of x , when $y = 6$.

Solution. Here $\bar{x} = \frac{\sum x}{n} = \frac{24}{4} = 6.$

$$\bar{y} = \frac{\sum y}{n} = \frac{44}{4} = 11.$$

$$\begin{aligned} \text{Regression coefficient } (b_{xy}) &= r \cdot \frac{\sigma_x}{\sigma_y} \\ &= \frac{n \sum xy - \sum x \sum y}{x \sum y^2 - (\sum y)^2} \\ &= \frac{4 \times 306 - 24 \times 44}{4 \times 574 - (44)^2} \\ &= 0.47 \end{aligned}$$

The regression equation of x on y is

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

$$\Rightarrow x - 6 = 0.47(y - 11)$$

$$\Rightarrow x - 6 = 0.47y - 5.17$$

$$\Rightarrow x = 0.47y + 0.83.$$

$$\text{when } y = 6, x = 0.47 \times 6 + 0.83 = 3.65$$

\therefore Required value of $x = 3.65$.

5.5 Exercises

Ex.1. Find the lines of regression x on y and y on x for the following data :

$x :$	3	5	6	6	9
$y :$	2	3	4	6	5

Ex. 2. For 10 observations on price (x) and supply (y), the following data were obtained.

$$\sum x = 130, \sum y = 220, \sum x^2 = 2288, \sum y^2 = 5506, \sum xy = 3467$$

Obtain the line of regression of y on x and estimate the supply when the price is 16 units

Ex. 3. The following data give the correlation coefficients, means and standard deviation of rainfall and yield of paddy in a certain tract :

	Yield per acre (in lbs)	Annual rainfall
Mean	973.5	18.3
S.D.	38.4	2.0

Coefficient of correlation = 0.58

Estimate the most likely yield of paddy when the annual rainfall is 22", other factors being assumed to remain the same.

Ex. 4. Obtain the correlation coefficient from the following results.

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

Answers

1. $y = 0.6x + 0.4, x = 1.2y + 1.20$

2. $y = 8.8 + 1.05x, 25.04.$

3. 1014.7
4. -0.92
6. **Chi-square test**

The chi-square test, written as ψ^2 -test, is a useful measure of comparing experimentally obtained results with those expected theoretically and based on the hypothesis. It is used as a **test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared.** In general **Chi-square test** is applied to those problems in which we study whether the frequency with which a given event has occurred, is significantly different from the one as expected theoretically. The measure of **Chi-square** enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained between observed frequencies and theoretical frequencies is due to error of sampling or due to chance.

The chi-square is computed on the basis of frequencies in a sample and thus the value of chi-square so obtained is a **statistic**. Chi-square is **not a parameter** as its value is not derived from the observations in a population. Hence **Chi-square test** is a **Non-Parametric test**. Chi-test is not concerned with any population distribution and its observations.

The ψ^2 -test was first used in testing statistical hypothesis by Karl Pearson in the year 1900. It is defined as

$$\psi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where O_i = observed frequency of *i*th event
 E_i = Expected frequency of *i*th event.

We require the following steps to calculate ψ^2 .

- Step 1.** Calculate all the expected frequencies i.e., E_i for all values of $i = 1, 2, \dots, n$.
- Step 2.** Take the difference between each observed frequency O_i and the corresponding expected frequency E for each value of i , i.e., and $(O_i - E_i)$.
- Step 3.** Square the difference for each value of i , i.e., calculate $(O_i - E_i)^2$ for all values of $i = 1, 2, 3, \dots, n$.
- Step 4.** Divide each square difference by the corresponding expected

frequency i.e., Calculate $\frac{(O_i - E_i)^2}{E_i}$ for all values of $i = 1, 2, 3, \dots, n$.

Step 5. Add all these quotients obtained in Step 4, then

$$\psi^2 = \sum_{i=1}^n \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

is the required value of **Chi-square**.

It should be noted that

- (a) The value of ψ^2 is always positive as each pair is squared up.
- (b) ψ^2 will be zero if each pair is zero and it may assume any value extending to infinity, when the difference between the observed frequency and expected frequency in each pair are unequal. Thus ψ^2 lies between 0 and ∞ .
- (c) The significance test on ψ^2 is always based on **One Tailed Test** of the right hand side of standard normal curve as ψ^2 is **always non-negative**.
- (d) As ψ^2 is a statistic and not a parameter, so it does not involve any assumption about the form of original distribution from which the observation come.

6.1 Degrees of Freedom

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the **degrees of freedom** and is denoted by ν .

Case I. If the data is given in the form of a series of variables in a row or column, then the **degree of freedom = (number of items in the series) - 1** i.e. $= n-1$, where n is the number of variables in the series in a row or column.

Case II. When the number of frequencies are put in cells in a contingency table, the degrees of freedom will be the product of **(number of rows less one)** and the **(number of columns less one)** i.e., $\nu = (R-1)(C-1)$, where R is the number of rows and C is the number of columns.

6.2 Chi-square distribution

ψ^2 - distribution is a continuous distribution whose probability density function is given by

$$P(\psi^2) = y_0 (\psi^2)^{\frac{1}{2}(v-2)} e^{-\frac{1}{2}\psi^2}$$

where y_0 = constant depending on the degrees of freedom.

v = degrees of freedom = $n - 1$.

6.3 Properties of ψ^2 -distribution

1. Chi-square curve is always positively skewed.
2. The mean of distribution is the number of degrees of freedom.
3. The standard deviation of ψ^2 distribution = $\sqrt{2v}$, where v is the degrees of freedom.
4. Chi-square values increase with the increase with the increase in degrees of freedom.
5. The value of ψ^2 lies between zero and infinity i.e., $0 \leq \psi^2 < \infty$.
6. The sum of two ψ^2 distribution is again a ψ^2 distribution i.e. if ψ_1^2 and ψ_2^2 are two independent and they have a ψ^2 -distribution with n_1 and n_2 degrees of freedom respectively, then $(\psi_1^2 + \psi_2^2)$ is also a ψ^2 -distribution with $(n_1 + n_2)$ degrees of freedom.
7. For different degrees of freedom, the shape of the curve will be different. See figure 1.
8. Like t-distribution its shape depends on the degree of freedom but it is not a symmetrical distribution.

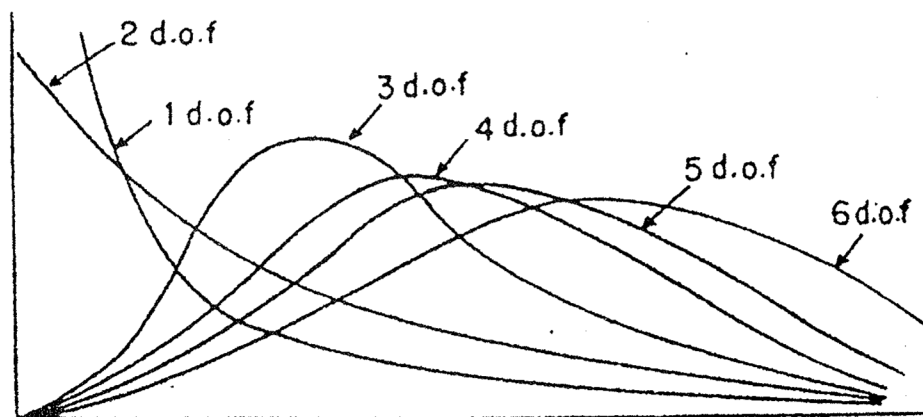


Fig. 1

6.4 Uses of χ^2 test

The χ^2 test is a very powerful test for testing the hypothesis of a number of statistical problems.

The important uses of χ^2 test are :

1. **Test of Goodness of Fit.** If the two curves, viz. (i) Observed frequency curve and (ii) the expected frequency curve, are drawn. then the Chi-square statistic may be used to determine whether the two curves so drawn are fitted good or not. Thus the term **goodness of fit** is used to test the concordance of the fitness of these two curves. Under this test there is only one variable, i.e., the degrees of freedom is $\nu = n - 1$.
2. **Test of Independence of Attributes.** The Chi-square test is used to see that the principles of classification of attributes are independent. In this test the attributes are classified into a two way table or a contingency table as the case may be. The observed frequency in each cell (square) is known as **Cell frequency**. The total frequency in each row or column of the two way contingency table is known as **Marginal frequency**. The degrees of freedom is $\nu = (R - 1) (C - 1)$,
where R = number of rows, C = number of columns in the two way contingency table. This test discloses whether there is any association or relationship between two or more attributes.
3. **Test of Homogeneity or a Test for a Specified Standard deviation.**
The Chi-square test may be used to test the homogeneity of the attributes in respect of a particular characteristic or it may also be used to test the population variance. In the case of a specified standard deviation, the test statistic is given by $\chi^2 = (n - 1) s^2 / \sigma_0^2$, where s^2 = sample variance and σ_0^2 is the hypothesized value of population variance.

6.5 Conditions for using the Chi-square test

1. Each of the observations making up the sample for this test should be independent of each other.
2. The expected frequency of any items or cell should not be less than 5. If it is less than 5,

then frequencies taking from the adjacent items or cells be pooled together in order to make it 5 or more than 5.

3. The total number of observations used in this test must be large i.e. $n \geq 50$.
4. This test is used only for drawing inferences by testing hypothesis. It cannot be used for estimation of parameter or any other value.
5. It is wholly dependent on the degrees of freedom.
6. The frequencies used in χ^2 -test should be absolute and not relative in terms.
7. The observation collected for χ^2 -test should be on random basis of sampling.

6.6 χ^2 -Test

The Chi-square test is widely used to test the independence of attributes. It is applied to test the association between the attributes when the sample data is presented in the form of a contingency table with any number of rows or columns.

WORKING RULE

Step 1. Set up the Null Hypothesis H_0 : No Association exists between the attributes.

Alternative Hypothesis H_1 : An association exists between the attributes.

Step 2. Calculate the expected frequency E corresponding to each cell by the formula

$$E_{ij} = \frac{R_i \times C_j}{n}$$

where R_i = Sum total of the row in which E_{ij} is lying

C_j = Sum total of the column in which E_{ij} is lying

n = Total sample size.

Step 3 Calculate χ^2 -statistic by the formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

The characteristics of this distribution are completely defined by the number of degrees of freedom ν which is given by $\nu = (R - 1)(C - 1)$,

where R = number of rows and

C = number of columns in the contingency table

Step 4 Find from the table the value of ψ^2 for a given value of the level of significance α and for the degrees of freedom ν , **Calculated in Step 2.**

If no value for α is mentioned, then take $\alpha = 0.05$.

Step 5. Compare the computed value of ψ^2 , with the tabled value of ψ_α^2 found in step 4.

(a) If calculated value of $\psi^2 <$ tabulated value of then ψ_α^2 then accept null hypothesis H_0 .

(b) If calculated value of $\psi^2 >$ tabulated value of ψ_α^2 , then rejected null hypothesis H_0 and accept the alternative hypothesis H_1 .

6.7 ψ^2 -Test for goodness of fit

ψ^2 -test is a measure of probabilities of association between the attributes. It gives us an idea about the divergence between the observed and expected frequencies. Thus the test is also described as the test of "**Goodness of Fit**". If the curves of these two distributions, when superimposed do not coincide or appear to diverge much we say that the fit is poor. On the other hand if they don't diverge much, then the fit is less poor.

This concept is illustrated by the following examples.

Example 1. In a sample survey of public opinion, answers to the questions

(i) Do you drink?

(ii) Are you in favour of local option on sale of liquor?

are tabulated below:

Question (1)			
	Yes	No	Total
Yes	56	31	87
No	18	6	24
Total	74	37	111

Can you infer whether or not the local option on the sale of liquor is dependent on individual drink.

Solution. Null hypothesis H_0 : The option on the sale of liquor is independent or not associated with individual drinking.

The theoretical frequencies are tabulated as below

Question (1)

	Yes	No
Yes	$\frac{87 \times 74}{111} = 58$	$\frac{37 \times 87}{111} = 29$
No	$\frac{74 \times 24}{111} = 16$	$\frac{24 \times 37}{111} = 8$
Total	74	37

Computation of test statistic

Applying the formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}, \text{ we get}$$

$$\begin{aligned} \chi^2 &= \frac{(56 - 58)^2}{58} + \frac{(18 - 16)^2}{16} + \frac{(31 - 29)^2}{29} + \frac{(6 - 8)^2}{8} \\ &= \frac{4}{58} + \frac{4}{16} + \frac{4}{29} + \frac{4}{8} = \frac{111}{116} = 0.957. \end{aligned}$$

3. **Degree of freedom** $\nu = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$.

4. **Decision.** The tabulated value of χ^2 at $\alpha = 0.05$ and one degree of freedom is $\chi^2_{0.05} = 3.441$.

Since calculated $\chi^2 = 0.957 < \text{Tabulated value } \chi^2_{0.05} = 3.441 \Rightarrow$ the Null hypothesis H_0 is accepted \Rightarrow Sale of liquor is independent or not associated with the individual drinking.

Example 2. From the table given below, whether the colour of son's eyes is associated with that of father's eyes.

Eye colour in Sons

		Not light	Light
Eye Colour in fathers	Not light	230	148
	Light	151	471

Solution. The observed frequencies are given by following Table.

		Eye colour in Sons		
		Not light	Light	Total
Eye colour in fathers	Not light	230	148	378
	Light	151	471	622
Total		381	619	1000

The theoretical frequencies of Expected frequencies are given by the table

Colour in sons		
		Light
Eye colour in fathers	Not light	$\frac{381 \times 378}{1000} = 144$
	Light	$\frac{381 \times 622}{1000} = 237$
		$\frac{619 \times 378}{1000} = 234$
		$\frac{619 \times 622}{1000} = 385$

1. Null hypothesis H_0 : The colour of the son's eyes is not associated with the colour of father's eyes.

2. Calculation of test Statistic.

$$\begin{aligned}\psi^2 &= \frac{(230-144)^2}{144} + \frac{(151-237)^2}{237} + \frac{(148-234)^2}{234} + \frac{(471-385)^2}{385} \\ &= \frac{7396}{144} + \frac{7396}{237} + \frac{7396}{234} + \frac{7396}{385} \\ &= 51.36 + 31.21 + 31.61 + 19.21 = 133.39\end{aligned}$$

3. Degree of freedom. $\nu = (R-1)(C-1) = (2-1)(2-1) = 1$

4. Decision. The tabulated value of ψ^2 for $\alpha=0.05$ and 1 degree of freedom is $\psi_{0.05,1}^2 = 3.84$.

Now calculated $\psi^2 = 133.39 > \psi_{0.05,1}^2 = 3.84 \Rightarrow$ the Null hypothesis is rejected \Rightarrow there is an

association between the colours of eyes of son's and colours of eyes of father's.

Example 3. A survey of 320 families with 5 children each revealed the following distribution.

No. of boys :	5	4	5	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that the male and female births are equally probable?

Solution. Null hypothesis H_0 : Male and female births are equally probable.

Alternative hypothesis H_1 : Male and female births are not equally probable.

2. Calculation of test statistic. On the basis of null hypothesis the expected frequencies are

$$320 \left(\frac{1}{2} + \frac{1}{2} \right)^5 = 320 \left(\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32} \right)$$

$$= 10, 50, 100, 100, 50, 10.$$

Table for ψ^2

O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
14	10	4	16	1.60
56	50	6	36	1.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	10	100	2.00
12	10	2	4	0.40
				$\psi^2 = \sum \frac{(O - E)^2}{E} = 7.16$

3. Level of significance. Take $\alpha = 0.05$.

4. Critical value. The table value of ψ^2 at $\alpha = 0.05$ for $(6-1)=5$ degrees of freedom is $\psi_{0.05}^2 = 11.07$.

5. Decision. Since the calculated value of $\psi^2 = 7.16 < \text{table value } \psi_{0.05}^2$ for 5 d.f.=11.07, so the null hypothesis is accepted \Rightarrow the male and female births are equally probable.

Example 4. In 60 throws of a dice, face one turned up 6 times, face two or three 18 times, face four or five 24 times and face six, 12 times. Test at 10% significance level, if the dice is honest, it being given that $P(\chi^2 > 6.25) = 0.1$ for 3 degrees of freedom.

Solution.

1. Null hypothesis H_0 = The dice is honest.

Alternative Hypothesis H_1 : The dice is not honest.

2. Computation of test statistics. Under the assumption of null hypothesis that the dice is

honest, the expected frequency for each face is $60 \times \frac{1}{6} = 10$

[\because Probability of turning up any one of the numbers

1, 2, 3, 4, 5, 6, is $\frac{1}{6}$].

We prepare the following table.

Computation Table for

Face of the die	Observed frequency O	Expected frequency E	O-E	$\frac{(O-E)^2}{E}$
1	6	10	-4	1.6
2 } 3 }	18	10 } 20 10 }	-2	0.2
4 } 5 }	24	10 } 20 10 }	4	0.8
6	12	10 10	2	0.4
Total	60	60	$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.0$	

$$\therefore \chi^2 = 3.0.$$

3. Critical value. The table value of χ^2 at $\alpha = 0.1$ for $4 - 1 = 3$ degrees of freedom is 6.25.

4. **Decision.** The tabulated value of $\psi^2 = 3.0 < \psi_{0.1}^2$ for 3 d.f. = 6.25, therefore, the Null hypothesis that the die is honest is accepted.

Example 5. Calculate the expected frequencies for the following data presuming the two attributes viz., conditions of home and condition of child as independent

		Condition of home	
		Clean	Dirty
Condition of child	Clean	70	50
	Fairly clean	80	20
	Dirty	35	45

Use chi-square test at 5% level of significance to state whether the two attributes are independent.

(Table value of ψ^2 at 5% for d.o.f. is 5.991 and for 3 d.o.f. is 7.815 and 4 d.o.f. is 9.488).

Solution. The expected frequency E , corresponding to each cell in the table is given by

$$E = \frac{R \times C}{n}$$

where R = a row total, C = a column total and n = the sample size.

We form the table of expected frequencies, with the help of the above rule and write the observed frequencies in each cell within brackets.

Condition of home			Total
	Clean	Dirty	
Clean	(70) $\frac{185 \times 120}{300} = 74$	(50) $\frac{115 \times 120}{300} = 46$	120
Fairly clean	(80) $\frac{185 \times 100}{300} = 61.67$	(20) $\frac{115 \times 100}{300} = 38.33$	100
Dirty	(35) $\frac{185 \times 80}{300} = 49.33$	(45) $\frac{115 \times 80}{300} = 30.67$	80
Total	185	115	300

1. Null Hypothesis H_0 : There does not exist any association between the attributes.
2. Alternative Hypothesis H_1 : An association exists between the attributes.
3. Calculation of test statistics

$$\begin{aligned}\psi^2 &= \frac{\sum (O - E)^2}{E} = \frac{(70 - 74)^2}{74} + \frac{(50 - 64)^2}{64} + \frac{(80 - 61.67)^2}{61.67} \\ &\quad + \frac{(20 - 38.33)^2}{38.33} + \frac{(35 - 49.33)^2}{49.33} + \frac{(45 - 30.67)^2}{30.67} \\ &= 0.2162 + 0.3478 + 5.4482 + 8.7657 + 4.1627 + 6.6954 \\ &= 25.636\end{aligned}$$

Degree of freedom = $(R - 1)(C - 1) = 2 \times 1 = 2$.

Decision. The table value of ψ^2 at $\alpha = 0.05$ for 2 degrees of freedom is $\psi_{0.05, 2}^2 = 5.991$.

Also the calculated $\psi^2 = 25.636 > \psi_{0.05}^2$ for 2 d.f. = 5.991

\Rightarrow the Null hypothesis is rejected \Rightarrow Alternative Hypothesis H_1 is accepted from which we conclude that there exists an association between the attributes.

Example 6. The following figures show the distribution of digits in number chosen at random from a telephone directory :

Digits :	0	1	2	3	4	5	6	7	8	9
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

(The table value of ψ^2 for 9 degrees of freedom = 16.919).

Solution.

1. Null Hypothesis H_0 : The digits occur equly frequently in the directory.

Alternative Hypothesis H_1 : The digits do not occur equally frequently.

2. Calculation of Test Statistic. Here the total number of frequencies

$$n = 1026 + 1107 + 997 + 966 + 1075 + 933 + 1107 + 972 + 964 + 853 = 10,000.$$

Expected frequency for each of the digits 0, 1, 2, 3 ..., 9 is $E = \frac{10,000}{10} = 1000$. Thus the

value of ψ^2 is

$$\begin{aligned}\psi^2 &= \frac{(1026-1000)^2}{1000} + \frac{(1107-1000)^2}{1000} + \frac{(997-1000)^2}{1000} \\ &\quad + \frac{(966-1000)^2}{1000} + \frac{(1075-1000)^2}{1000} + \frac{(933-1000)^2}{1000} \\ &\quad + \frac{(1107-1000)^2}{1000} + \frac{(972-1000)^2}{1000} \\ &\quad + \frac{(964-1000)^2}{1000} + \frac{(853-1000)^2}{1000} \\ &= 0.676 + 11.449 + 0.009 + 1.156 + 5.625 + 4.489 + 11.449 \\ &\quad + 0.784 + 1.296 + 1.296 + 21.609 = 58.542\end{aligned}$$

3. **Critical value.** The tabulated value or critical value of ψ^2 at $\alpha = 0.05$ for $10-1=9$ degrees of freedom is $\psi_{0.05,9}^2 = 16.919$.

Since calculated value of $\psi^2 = 58.542$ is $>$ Tabulated $\psi^2 = 16.919$ the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted \Rightarrow the digits do not occur equally frequently.

Example 7. A dice is tossed 120 times with the following results :

Number turned up :	1	2	3	4	5	6	Total
Frequency :	30	25	18	10	22	15	120

Test the hypothesis that the dice is unbiased.

Solution. 1. Null Hypothesis H_0 : The dice is unbiased one.

Alternative Hypothesis H_1 : the dice is a biased one.

2. **Calculation of test statistic.** On the hypothesis that the dice is unbiased, the expected frequency is $120 \times \frac{1}{6} = 20$. We calculate ψ^2 from the following table.

O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
30	20	10	100	5.00
25	20	5	25	1.25
15	20	-2	4	0.20
10	20	-10	100	5.00
22	20	2	4	0.20
15	20	-5	25	1.25
				$\psi^2 = \sum \frac{(O - E)^2}{E} = 12.90$

3. **Critical value.** The control value of ψ^2 at $\alpha = 0.05$ and for $6 - 1 = 5$ degrees of freedom is $\psi_{0.05, 5}^2 = 11.070$.

4. **Decision.** Since the calculated value $\psi^2 = 12.90$ is $\psi_{0.05, 5}^2 = 11.07$, so null the hypothesis H_0 is rejected and the alternative Hypothesis H_1 is accepted \Rightarrow the dice is a biased one.

Example 8. From the adult male population of seven large cities random sample given 2×7 contingency table of married and unmarried men, as given below were taken. Can it be said that there is a significant variation among the cities in the tendency of men to marry?

City	A	B	C	D	E	F	G	Total
Married	133	164	155	106	153	123	146	980
Unmarried	36	57	40	37	55	33	36	294
Total	169	221	195	143	208	156	182	1274

$$[At (2 - 1)(7 - 1) d.f \text{ Take } \psi_{0.05, 6}^2 = 12.6]$$

Solution 1. Null Hypothesis H_0 : There is no significant variation among the cities in the tendency of men to marry.

Alternative Hypothesis H_1 : There is a significant variation among the cities in the tendency of men to marry.

2. Calculation of test statistic. On the basis of Null Hypothesis the expected frequencies are:

$$\text{Expected number of married people in City A} = \frac{980}{1274} \times 169 = 130$$

$$\text{Expected number of married people in City B} = \frac{980}{1274} \times 221 = 170$$

$$\text{Expected number of married people in City C} = \frac{980}{1274} \times 195 = 150$$

$$\text{Expected number of married people in City D} = \frac{980}{1274} \times 143 = 110$$

$$\text{Expected number of married people in City E} = \frac{980}{1274} \times 208 = 160$$

$$\text{Expected number of married people in City F} = \frac{980}{1274} \times 156 = 120$$

$$\text{Expected number of married people in City G} = \frac{980}{1274} \times 182 = 140$$

Similarly, the expected frequencies for unmarried are 39, 51, 45, 33, 48, 36 and 42.

Table for expected frequencies

Married	130	170	150	110	160	120	140
Unmarried	39	51	45	33	48	36	42
Total	169	221	195	143	208	156	182

Table for calculation of χ^2

O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
133	130	3	9	0.069
164	170	-6	36	0.212

155	150	5	25	0.167
106	110	-4	16	0.145
153	160	-7	49	0.306
123	120	3	9	0.075
146	140	6	36	0.257
36	39	-3	9	0.231
57	51	6	36	0.706
40	45	-5	25	0.556
37	33	4	16	0.485
55	48	7	49	1.021
33	36	-3	9	0.250
36	42	-6	36	0.857
$\therefore \psi^2 = \sum \frac{(O-E)^2}{E} = 5.337$				

3. **Level of significance.** Take $\alpha = 0.05$.

4. **Critical value.** The critical value or table value of ψ^2 at $\alpha = 0.05$ for $(2-1)(7-1) = 6$ degrees of freedom is $\psi_{0.05,6}^2 = 12.6$.

5. **Decision.** Since the calculated value of $\psi^2 = 5.337 < \text{critical value } \psi_{0.05}^2$ for 6 d.f. = 12.6, so the Null hypothesis is accepted \Rightarrow that there is no significant variation among the cities in the tendency of men to marry.

6.8 EXERCISES

1. A sample of 300 students of Under-Graduate and 300 students of Post-Graduate classes of a University were asked to give their opinion toward the autonomous colleges. 190 of the Under-Graduate and 210 of the Post-Graduate students favoured the autonomous status.

Present the above data in the form of frequency table and test of 5% levels, the opinion of Under-Graduate and Post-Graduate students on autonomous status of colleges are independent (Table

association between the colours of eyes of son's and colours of eyes of father's.

Example 3. A survey of 320 families with 5 children each revealed the following distribution.

No. of boys :	5	4	5	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that the male and female births are equally probable?

Solution. Null hypothesis H_0 : Male and female births are equally probable.

Alternative hypothesis H_1 : Male and female births are not equally probable.

2. Calculation of test statistic. On the basis of null hypothesis the expected frequencies are

$$320 \left(\frac{1}{2} + \frac{1}{2} \right)^5 = 320 \left(\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32} \right)$$

$$= 10, 50, 100, 100, 50, 10.$$

Table for ψ^2

O	E	O - E	(O - E) ²	$\frac{(O - E)^2}{E}$
14	10	4	16	1.60
56	50	6	36	1.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	10	100	2.00
12	10	2	4	0.40
				$\psi^2 = \sum \frac{(O - E)^2}{E} = 7.16$

3. Level of significance. Take $\alpha = 0.05$.

4. Critical value. The table value of ψ^2 at $\alpha = 0.05$ for $(6-1)=5$ degrees of freedom is $\psi_{0.05}^2 = 11.07$.

5. Decision. Since the calculated value of $\psi^2 = 7.16 < \text{table value } \psi_{0.05}^2$ for 5 d.f.=11.07, so the null hypothesis is accepted \Rightarrow the male and female births are equally probable.

Example 4. In 60 throws of a dice, face one turned up 6 times, face two or three 18 times, face four or five 24 times and face six, 12 times. Test at 10% significance level, if the dice is honest, it being given that $P(\chi^2 > 6.25) = 0.1$ for 3 degrees of freedom.

Solution.

1. Null hypothesis H_0 = The dice is honest.

Alternative Hypothesis H_1 : The dice is not honest.

2. Computation of test statistics. Under the assumption of null hypothesis that the dice is

honest, the expected frequency for each face is $60 \times \frac{1}{6} = 10$

[\because Probability of turning up any one of the numbers

1, 2, 3, 4, 5, 6, is $\frac{1}{6}$].

We prepare the following table.

Computation Table for

Face of the die	Observed frequency O	Expected frequency E	O-E	$\frac{(O-E)^2}{E}$
1	6	10	-4	1.6
2 } 3 }	18	10 } 20 10 }	-2	0.2
4 } 5 }	24	10 } 20 10 }	4	0.8
6	12	10 10	2	0.4
Total	60	60	$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.0$	

$$\therefore \chi^2 = 3.0.$$

3. Critical value. The table value of χ^2 at $\alpha = 0.1$ for $4 - 1 = 3$ degrees of freedom is 6.25.

4. **Decision.** The tabulated value of $\psi^2 = 3.0 < \psi_{0.1}^2$ for 3 d.f. = 6.25, therefore, the Null hypothesis that the die is honest is accepted.

Example 5. Calculate the expected frequencies for the following data presuming the two attributes viz., conditions of home and condition of child as independent

		Condition of home	
		Clean	Dirty
Condition of child	Clean	70	50
	Fairly clean	80	20
	Dirty	35	45

Use chi-square test at 5% level of significance to state whether the two attributes are independent.

(Table value of ψ^2 at 5% for d.o.f. is 5.991 and for 3 d.o.f. is 7.815 and 4 d.o.f. is 9.488).

Solution. The expected frequency E , corresponding to each cell in the table is given by

$$E = \frac{R \times C}{n}$$

where R = a row total, C = a column total and n = the sample size.

We form the table of expected frequencies, with the help of the above rule and write the observed frequencies in each cell within brackets.

Condition of home			Total
	Clean	Dirty	
Clean	(70) $\frac{185 \times 120}{300} = 74$	(50) $\frac{115 \times 120}{300} = 46$	120
Fairly clean	(80) $\frac{185 \times 100}{300} = 61.67$	(20) $\frac{115 \times 100}{300} = 38.33$	100
Dirty	(35) $\frac{185 \times 80}{300} = 49.33$	(45) $\frac{115 \times 80}{300} = 30.67$	80
Total	185	115	300

1. Null Hypothesis H_0 : There does not exist any association between the attributes.
2. Alternative Hypothesis H_1 : An association exists between the attributes.
3. Calculation of test statistics

$$\begin{aligned}\psi^2 &= \frac{\sum(O-E)^2}{E} = \frac{(70-74)^2}{74} + \frac{(50-64)^2}{64} + \frac{(80-61.67)^2}{61.67} \\ &\quad + \frac{(20-38.33)^2}{38.33} + \frac{(35-49.33)^2}{49.33} + \frac{(45-30.67)^2}{30.67} \\ &= 0.2162 + 0.3478 + 5.4482 + 8.7657 + 4.1627 + 6.6954 \\ &= 25.636\end{aligned}$$

Degree of freedom = $(R-1)(C-1) = 2 \times 1 = 2$.

Decision. The table value of ψ^2 at $\alpha = 0.05$ for 2 degrees of freedom is $\psi_{0.05,2}^2 = 5.991$.

Also the calculated $\psi^2 = 25.636 > \psi_{0.05}^2$ for 2 d.f. = 5.991

\Rightarrow the Null hypothesis is rejected \Rightarrow Alternative Hypothesis H_1 is accepted from which we conclude that there exists an association between the attributes.

Example 6. The following figures show the distribution of digits in number chosen at random from a telephone directory :

Digits :	0	1	2	3	4	5	6	7	8	9
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853

Test at 5% level whether the digits may be taken to occur equally frequently in the directory.

(The table value of ψ^2 for 9 degrees of freedom = 16.919).

Solution.

1. Null Hypothesis H_0 : The digits occur equilly frequently in the directory.

Alternative Hypothesis H_1 : The digits do not occur equally frequently.

2. Calculation of Test Statistic. Here the total number of frequencies

$$n = 1026 + 1107 + 997 + 966 + 1075 + 933 + 1107 + 972 + 964 + 853 = 10,000.$$

Expected frequency for each of the digits 0, 1, 2, 3 ..., 9 is $E = \frac{10,000}{10} = 1000$. Thus the value of ψ^2 is

$$\begin{aligned}\psi^2 &= \frac{(1026-1000)^2}{1000} + \frac{(1107-1000)^2}{1000} + \frac{(997-1000)^2}{1000} \\ &\quad + \frac{(966-1000)^2}{1000} + \frac{(1075-1000)^2}{1000} + \frac{(933-1000)^2}{1000} \\ &\quad + \frac{(1107-1000)^2}{1000} + \frac{(972-1000)^2}{1000} \\ &\quad + \frac{(964-1000)^2}{1000} + \frac{(853-1000)^2}{1000} \\ &= 0.676 + 11.449 + 0.009 + 1.156 + 5.625 + 4.489 + 11.449 \\ &\quad + 0.784 + 1.296 + 1.296 + 21.609 = 58.542\end{aligned}$$

3. **Critical value.** The tabulated value or critical value of ψ^2 at $\alpha = 0.05$ for $10-1=9$ degrees of freedom is $\psi_{0.05,9}^2 = 16.919$.

Since calculated value of $\psi^2 = 58.542$ is $>$ Tabulated $\psi^2 = 16.919$ the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted \Rightarrow the digits do not occur equally frequently.

Example 7. A dice is tossed 120 times with the following results :

Number turned up :	1	2	3	4	5	6	Total
Frequency :	30	25	18	10	22	15	120

Test the hypothesis that the dice is unbiased.

Solution. 1. Null Hypothesis H_0 : The dice is unbiased one.

Alternative Hypothesis H_1 : the dice is a biased one.

2. **Calculation of test statistic.** On the hypothesis that the dice is unbiased, the expected frequency is $120 \times \frac{1}{6} = 20$. We calculate ψ^2 from the following table.

O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
30	20	10	100	5.00
25	20	5	25	1.25
15	20	-2	4	0.20
10	20	-10	100	5.00
22	20	2	4	0.20
15	20	-5	25	1.25
$\psi^2 = \sum \frac{(O - E)^2}{E} = 12.90$				

3. **Critical value.** The control value of ψ^2 at $\alpha = 0.05$ and for $6 - 1 = 5$ degrees of freedom is $\psi_{0.05, 5}^2 = 11.070$.

4. **Decision.** Since the calculated value $\psi^2 = 12.90$ is $\psi_{0.05, 5}^2 = 11.07$, so null the hypothesis H_0 is rejected and the alternative Hypothesis H_1 is accepted \Rightarrow the dice is a biased one.

Example 8. From the adult male population of seven large cities random sample given 2×7 contingency table of married and unmarried men, as given below were taken. Can it be said that there is a significant variation among the cities in the tendency of men to marry?

City	A	B	C	D	E	F	G	Total
Married	133	164	155	106	153	123	146	980
Unmarried	36	57	40	37	55	33	36	294
Total	169	221	195	143	208	156	182	1274

[At $(2 - 1)(7 - 1)$ d.f Take $\psi_{0.05, 6}^2 = 12.6$]

Solution 1. Null Hypothesis H_0 : There is no significant variation among the cities in the tendency of men to marry.

Alternative Hypothesis H_1 : There is a significant variation among the cities in the tendency of men to marry.

2. Calculation of test statistic. On the basis of Null Hypothesis the expected frequencies are:

$$\text{Expected number of married people in City A} = \frac{980}{1274} \times 169 = 130$$

$$\text{Expected number of married people in City B} = \frac{980}{1274} \times 221 = 170$$

$$\text{Expected number of married people in City C} = \frac{980}{1274} \times 195 = 150$$

$$\text{Expected number of married people in City D} = \frac{980}{1274} \times 143 = 110$$

$$\text{Expected number of married people in City E} = \frac{980}{1274} \times 208 = 160$$

$$\text{Expected number of married people in City F} = \frac{980}{1274} \times 156 = 120$$

$$\text{Expected number of married people in City G} = \frac{980}{1274} \times 182 = 140$$

Similarly, the expected frequencies for unmarried are 39, 51, 45, 33, 48, 36 and 42.

Table for expected frequencies

Married	130	170	150	110	160	120	140
Unmarried	39	51	45	33	48	36	42
Total	169	221	195	143	208	156	182

Table for calculation of χ^2

O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
133	130	3	9	0.069
164	170	-6	36	0.212

155	150	5	25	0.167
106	110	-4	16	0.145
153	160	-7	49	0.306
123	120	3	9	0.075
146	140	6	36	0.257
36	39	-3	9	0.231
57	51	6	36	0.706
40	45	-5	25	0.556
37	33	4	16	0.485
55	48	7	49	1.021
33	36	-3	9	0.250
36	42	-6	36	0.857
$\therefore \psi^2 = \sum \frac{(O-E)^2}{E} = 5.337$				

3. **Level of significance.** Take $\alpha = 0.05$.

4. **Critical value.** The critical value or table value of ψ^2 at $\alpha = 0.05$ for $(2-1)(7-1) = 6$ degrees of freedom is $\psi^2_{0.05,6} = 12.6$.

5. **Decision.** Since the calculated value of $\psi^2 = 5.337 < \text{critical value } \psi^2_{0.05}$ for 6 d.f. = 12.6, so the Null hypothesis is accepted \Rightarrow that there is no significant variation among the cities in the tendency of men to marry.

6.8 EXERCISES

1. A sample of 300 students of Under-Graduate and 300 students of Post-Graduate classes of a University were asked to give their opinion toward the autonomous colleges. 190 of the Under-Graduate and 210 of the Post-Graduate students favoured the autonomous status.

Present the above data in the form of frequency table and test of 5% levels, the opinion of Under-Graduate and Post-Graduate students on autonomous status of colleges are independent (Table

"Learner's Feed-back"

After going through the Modules / Units please answer the following questionnaire.
Cut the portion and send the same to the Directorate.

To
The Director
Directorate of Distance Education,
Vidyasagar University,
Midnapore - 721 102

1. The modules are : (give ✓ in appropriate box)

☐ easily understandable; ☐ very hand; ☐ partially understandable.

2. Write the number of the Modules/Units which are very difficult to understand :

.....
.....
.....

3. Write the number of the Modules/Units which according to you should be re-written :

.....
.....
.....

4. Which portion / page is not understandable to (mention the page no. and portion)

.....
.....
.....
.....

5. Write a short comment about the study material as a learner.

.....
.....
.....
.....
.....

Date :

.....
(Full Signature of the Learner)

Enrolment No.

Phone / Mobile No.

value of ψ^2 at 5% level for 1 d.f. is 3.84) [Hint. The contingency table of observed frequencies in which expected frequencies are shown in braces is

	Favour	Not in favour	Total
Under Graduate :	190 (200)	110 (100)	300
Post Graduate :	210 (200)	90 (100)	300
	400	200	600

Null Hypothesis H_0 : Opinion on autonomous status and level of Graduation are independent.

$$\psi^2 = \frac{100}{200} + \frac{100}{200} + \frac{100}{100} + \frac{100}{100} = 3.$$

$$\text{d.f.} = (2 - 1)(2 - 1) = 1.$$

Table value of ψ^2 at $\alpha = 0.05$ and for 1 d.f. is $\psi_{0.05}^2 = 3.84$.

Now calculated $\psi^2 = 3 < \psi_{0.05}^2 = 3.84$. Null hypothesis is accepted

Opinion on autonomous status and level of Graduation are independent.

2. A certain drug is claimed to be effective in curing colds. In an experiment of 164 people with cold, half of them were given the drug and half of them given sugar pills. The patient's reactions to the treatment are recorded in the following table. Test the hypothesis that the drug is not better than sugar pills for sugar colds

	Helped	Harmed	No effect
Drug :	52	10	20
Sugar Pills :	44	12	26

3. In a survey of 200 boys, of which 75 were intelligent, 40 had skilled fathers, while 85 of the unintelligent boys had unskilled fathers. Do these support the hypothesis that skilled fathers have intelligent boys. Use ψ^2 test. Values of ψ^2 for 1-degree of freedom at 5% level is 3.84.

4. Four dice were thrown 112 times and the number of times 1, 3 or 5 were as under :

Number of dice showing 1, 3 or 5	0	1	2	3	4
Frequency	10	25	40	30	7

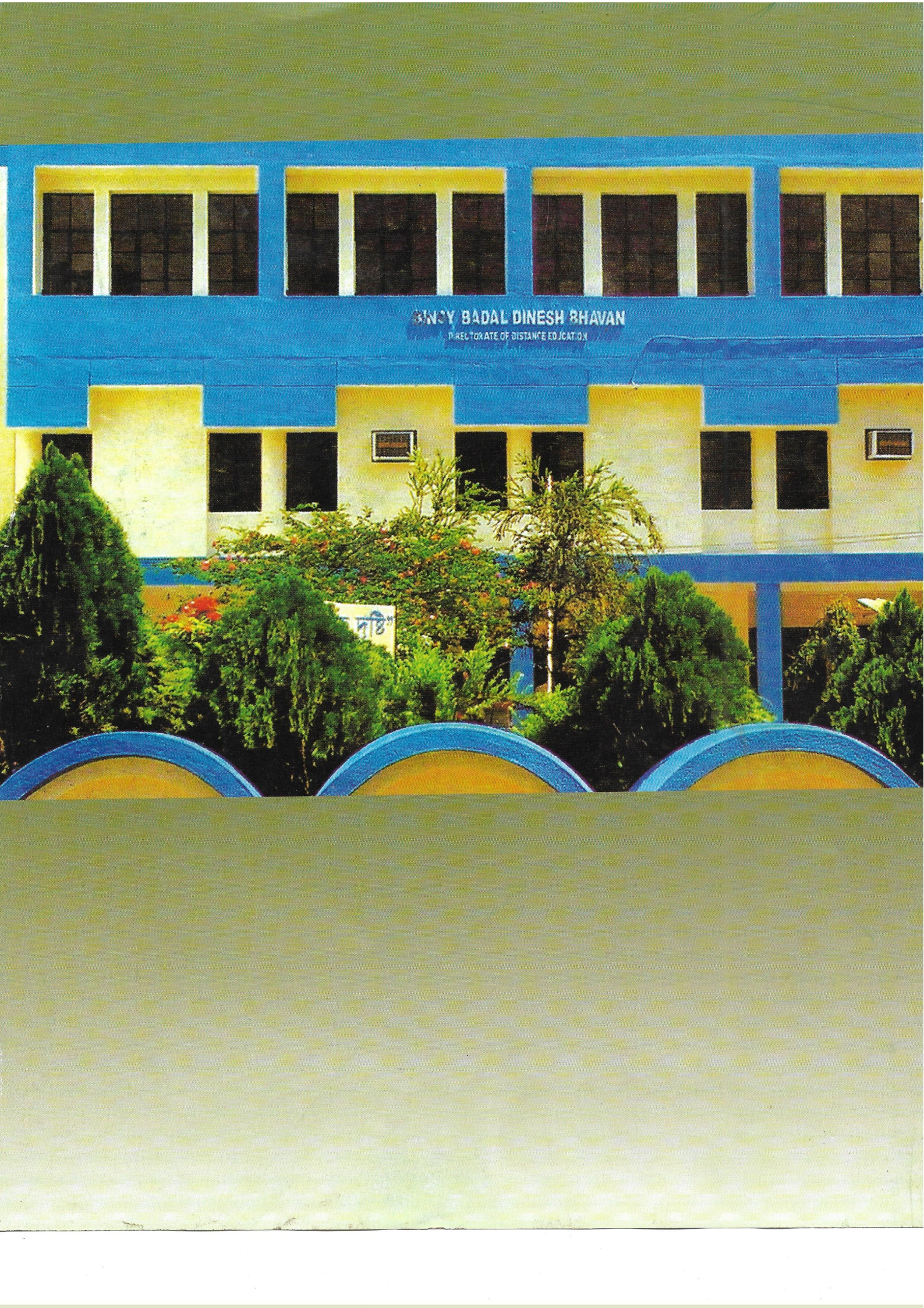
Answers

- Opinion on autonomous status and level of graduations are independent.
- Computed $\chi^2 = 1.622$, $\chi^2_{0.05} = 5.991$. Drug is not better than sugar points.
- $\chi^2 = 0.88$, $\chi^2_{0.05}$ 1 d.f. = 3.84; skilled fathers have intelligent boys.
- All the four dice are fair.
- Unit Summary :** In this module, we have discussed probability, some important distributions, correlation and regression analysis, and chi-square test. Also we have discussed the properties of correlation and regression coefficient. Finally a lot of problems have solved of our discussion on Biostatistics.

8. Reference / Suggested Further Readings

- N.G. Das, Statistical Methods, Vol. I & II.
Das & Das Publishers, Calcutta.
- P.N. Arora & P.K. Malhan, Biostatistics, Himalaya Publishing House, Delhi, 1996.
- I.A. Khan & A. Khanum, Fundamentals of Biostatistics, Ukaaz Publications, Andhra Pradesh, 1994.
- S.C. Gupta & Kappor, Fundamentals of Mathematical Statistics, Meerut, 2002.

--- 0 ---



BINAY BADAL DINESH BHAVAN

DIRECTORATE OF DISTANCE EDUCATION

दृष्टि