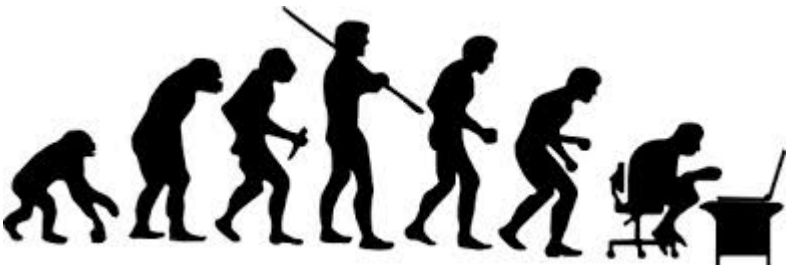
MZO-04

Vardhman Mahaveer Open University, Kota



Evolution, Bio-Statistics and Computer Applications in Zoology

MZO-04

Vardhman Mahaveer Open University, Kota

Evolution, Bio-Statistics and Computer Applications in Zoology

Course Development Committee

Chair Person

Prof. Vinay Kumar Pathak
Vice-Chancellor
Vardhman Mahaveer Open University, Kota

Coordinator and Members

Convener

SANDEEP HOODA
Department of Zoology
School of Science & Technology
Vardhman Mahaveer Open University, Kota

Members

- **Prof. L.R.Gurjar**
  Director (Academic)
  Vardhman Mahaveer Open
  University, Kota
- **Dr. Anuradha Dubey**
  Deputy Director
  School of Science & Technology
  Vardhman Mahaveer Open
  University, Kota
- **Dr. Arvind Pareek**
  Director (Regional Centre)
  Vardhman Mahaveer Open
  University, Kota
- **Prof. K.K. Sharma**
  MDSU,Ajmer
- **Prof. Maheep Bhatnagar**
  MLSU, Udaipur
- **Prof. S.C. Joshi**
  University of Rajasthan,Jaipur
- **Dr. Anuradha Singh**
  Department of Zoology
  Govt. College, Kota
- **Dr.M.M.Ranga**
  Department of Zoology
  Govt. College, Ajmer

Editing and Course Writing

Editor

SANDEEP HOODA
Assistant Professor
& Convener of Zoology
School of Science & Technology
Vardhman Mahaveer Open University, Kota

Writers:

Writer Name	Unit No.	Writer Name	Unit No.
Dr. Rajesh Yadav Dept. of Zoology, JECRC University Jaipur Mr. Pankaj Kumar DMRC, Jodhpur	1, 9, 11	**Sunil Choudhary** Arid Forest Research Institute, Jodhpur	2,3,4
Prof. K.K.Sharma Ex Head, Dept. of Zoology, MDS University, Ajmer	5,6	**Dr. Varsha Gupta** Dept. of Microbiology, JECRC University Jaipur	7,8
Dr. Abhishek Rajpurohit Department of Zoology Lachoo Memorial College of Science &Technology ,Jodhpur	**10,12,13**	**Mohammed Kasim** Dept. of Zoology,JNVU Jodhpur	14
Ms. Mugdha Agarwal Banasthali Vidhyapeeth	15		

Academic and Administrative Management

Prof. Vinay Kumar Pathak
Vice-Chancellor
Vardhman Mahaveer Open University, Kota

Prof. L.R. Gurjar
Director (Academic)
Vardhman Mahaveer Open University, Kota

Prof. Karan Singh
Director (MP&D)
Vardhman Mahaveer Open University, Kota

Dr. Subodh Kumar
Additional Director (MP&D)
Vardhman Mahaveer Open University, Kota

ISBN :

Printed and Published on behalf of the Registrar, V.M. Open University, Kota.
Printed by :

Vardhman Mahaveer Open University, Kota

Index

Vardhman Mahaveer Open University, Kota

Preface

The present book entitled "**Evolution, Bio-Statistics and Computer Applications in Zoology"** has been designed so as to cover the unit-wise syllabus of MZO-04 course for M.Sc. Zoology (Previous) students of Vardhman Mahaveer Open University, Kota. The basic principles and theory have been explained in simple, concise and lucid manner. Adequate examples, diagrammes, photographs and self-learning exercises have also been included to enable the students to grasp the subject easily. The unit writers have consulted various standard books and internet on the subject and they are thankful to the authors of these reference books.

Unit - 1

Evolutionary Mechanism

Structure of the Unit

1.0 Objectives

After going through this unit you will be able to understand:

- Gene pool and gene frequency
- Hardy-Weinberg Law of genetic equilibrium
- What are the different types of destabilizing forces?
- What is Molecular Evolution?
- What are the causes of molecular Evolution?
- How the evolution of gene families occur?
- What is Molecular drive?

1.1 Introduction

Evolution has been described as "descent with modification" by Darwin. Evolution cannot be an individual phenomenon, because an individual lives and dies with a fixed genotype. Recently, evolution has been regarded as a change in the genetic composition of the population rather than the change at individual level. Thus, the population rather than the individual has been regarded as the functional unit of evolution. With the development of the concept of gene and alleles, the genetic basis of inherited variation was established. And the changes over time in relative frequencies of the genetic variations in sub-population are the basis for the evolution of the species.

The total genetic stock of the population is its *gene pool*. Individuals have a selection of alleles from that gene pool, possibly taken randomly. The *Hardy-Weinberg equilibrium* is a means of calculating expected genotype frequencies from allele/gene frequencies determined in the same population (and vice-versa) assuming random mating, equal reproductive success, no mutations and no effects of selection or migration affecting particular genotypes. If a population does not fit Hardy-Weinberg predictions then that is evidence of some real effect (e.g. Natural selection) operating to disturb it. Thus, Gene frequencies and the forces that change these frequencies such as mutation, migration selection and genetic drift are being studied widely these days.

1.2 Hardy-Weinberg Law of genetic equilibrium

The concept of population is generally used to denote a group of individuals which have common features. A *population* may be defined as an assemblage of living beings that represents a closely interacting system. Or, more elaborately, *population* may be defined as *a group of individuals; each of them could potentially mate with a member of the same group of the opposite sex and hence could take part in the reproduction of the population.* The potential to reproduce by mating with another member of the population is the main aspect of population. The particular individual may not take part in actual reproduction but the potential for it remains within.

Populations are dynamic, they may grow and expand or diminish and contract through changes in birth or death rates, by migration or by merging with other populations. The members of one population could mate with members of another population, but they usually do not, because they are reproductively isolated by some barriers. The resulting gain or loss of inherited variation over time within populations of the species determines the course of evolution.

1.2.1 Gene pool

In a Mendelian population, the set of genetic information carried by all members which can interbreed is known as the *gene pool*. Or, in other words, it is the sum total of alleles/genes present in a Mendelian population. The genes are temporarily embodied in the individuals and are passed on to the next generation in the reproductive gametes of the population. Therefore, gene pool going in the gametes is called *gametic pool*.

The study of gene pool of a population tells us about the kind of genes present in the population, the proportion of different kinds of genes and also the way in which these kinds are distributed among the individuals of the population.

The ratio of a gene in the gene pool of a population is called *gene frequency*. The gene pool of each population maintains its integrity as long as there is free random interbreeding between the individuals of this population. When there is interbreeding among populations of a species, the genes from one gene pool enter another and vice-versa. This exchange of genes among populations is called *gene flow*. This causes mixing and reshuffling of gene pool.

Gene pool becomes larger by the addition of genes to the gene pool. This is possible by immigration or inward migration of individuals into a population from other populations.

Gene pool becomes smaller by the loss or removal of genes from the gene pool. This occurs by emigration, natural selection and genetic drifts.

The gene pool is not static. In sexually reproducing and cross fertilizing organisms, gametogenesis, meiosis and fertilization bring about constant reshuffling of alleles of a gene pool.

1.2.2 Gene frequency

The relative frequency of an allele in a given sample of population is called *gene frequency or allele frequency*. It is also defined as the percentage of alleles of a given type in a population. In population genetics, the focus is on groups rather than individuals and on the measurement of gene and genotype frequency from generation to generation rather on the distribution of genotypes resulting from a single mating. Gene frequency provides a description of genetic variability transmitted from generation to generation in a population.

With the help of gene frequency of a given controlling known trait, the genetic structure of population can be studied. And this is possible once the mode of inheritance and the number of different alleles present in the population have been established. Sometimes, it is not possible to determine directly the gene frequency because; in many cases only the phenotypes and not genotype are observed. However, if alleles expressed in a co-dominant fashion are considered, phenotypes are equivalent to genotype. For example, MN blood groups in human.

Table 1.1 MN Blood groups in man

Blood Type	Genotype	Reaction with antibodies	
		Anti-M	Anti-N
M	L^ML^M	+	-
N	L^NL^N	-	+
MN	L^ML^N	+	+

Here, the locus L on the chromosome-2 bears two alleles, L^M and L^N. Each allele controls the production of respective antigen, i.e. antigen M and antigen N and the genotypes can be $L^M L^M$, $L^N L^N$, $L^M L^N$. A simple count of alleles provides the gene frequency. Genotype frequencies of the individuals of the population give us the gene frequencies.

For example, in a population of 100 white Americans, 36 were MM, 48 were MN and 16 were NN. The frequencies of the M and N alleles in a population can be determined by counting the number of individuals of each phenotype.

Table 1.2 Measuring gene frequencies of M and N alleles in a population

Genotype/Phenotype	MM	MN	NN	Total
No. of individuals	36	48	16	100
No. of M alleles	72	48	0	120
No. of N alleles	0	48	32	80
Total no. of alleles	72	96	32	200

Frequency of M in population =120/200=0.6 or 60%

Frequency of N in population = 80/200=0.4 or 40%

Thus, the gene frequency convey us all the information about the transmission of genes of the successive generation and about the gene pool of the population. Genetic variability which can be transmitted is the function of the allele frequencies of both co-dominant and dominant genes. It also provides us the information about the two independent phenotype frequencies for each co-dominant locus.

1.2.3 The Hardy-Weinberg Law

Recent concept about evolution regards evolution as a change in the genetic composition of population rather than change at individual level, this shows that the unit of evolution is population. A population of similar individuals living within a circumscribed area at a given time and capable of interbreeding and exchanging genes freely, is known as *genetic population* or *Mendelian population*. All the individuals of Mendelian population have somewhat similar genes arranged

in similar fashion, so that there is free gene flow among them. The sum total of genes of all the individuals of a population constitutes the *gene pool*.

In 1908, **G.H. Hardy,** a British Mathematician and **W. Weinberg,** a German Physician, proposed a mathematical expression for the distribution of gene and genotype frequencies in a population of diploid, sexually reproducing individuals. This law occupies the position of central importance in the modern theory of evolution. According to this law, "**The relative frequencies of various genes in a large and randomly mating panmictic (Panmictic means the mating pattern that leads to random union of gametes from the individuals) population tend to remain constant from generation to generation in the absence of mutation, selection and gene flow, that is, the population remains in a genetic equilibrium**".

According to Hardy-Weinberg Law, the distribution of genotypes produced by random mating in the next generation can be expressed as:

$p^2 + 2pq + q^2 = 1$

(Where p is frequency of dominant gene and q is frequency of recessive gene)

The law, therefore, indicates that if gene frequencies of a population do not change, i.e. if the population is in genetic equilibrium then the rate of evolution is zero.

The Hardy-Weinberg law (HWL) is one of the fundamental concepts in population genetics and forms the basis of the later. In this law, the following conditions must be considered:

1. The population must be infinitely large or at least large enough so that sampling error is negligible.
2. Mating with the population occurs at random,
3. There must not be any selective advantage for any genotype, so that all genotypes produced by random mating are equally viable and fertile, and
4. There should not be mutation, migration, selection and genetic drift.

Mathematical formulation of the Hardy-Weinberg's Law

To explain the law, let us presume that 'p' represents the frequency of a dominant allele (A) and 'q' the frequency of a recessive allele (a) in a population. As the sum of these frequencies represents 100 per cent of the population, so, $p + q = 1$.

On mating two heterozygous individuals for a given trait, the F_2 generation consists of 25% dominant homozygous, 50% heterozygous and 25% homozygous recessive. This is because the two types of gametes whether male or female are formed by the mating individuals in equal numbers. If we assume that in a population random mating occurs and all the organisms of the population produces the same quantity of gametes, then the proportion of the genotypes in the next generation with respect to gene 'A' and its allele 'a' will be as follows-

Table 1.3 Genotype frequencies of 'A' and 'a' alleles in a random mating population

Eggs / Sperms	A (p=0.5)	a (q=0.5)
A (p=0.5)	AA (P^2=0.25)	Aa (pq=0.25)
A (q=0.5)	Aa (pq=0.25)	Aa (q^2 =0.25)

Genotype ratio :- 0.25 AA : 0.50 Aa : 0.25 aa

or 25%AA: 50%Aa : 25%aa

Thus we find that these are the same proportions as in the previous generations. If the previous generation were made up of other proportions of genotypes such as 0.36AA, 0.16aa and 0.48Aa, then the succeeding generation will show the same unchanged genotype proportion (i.e. 0.36AA, 0.16aa and 0.48Aa) (Table 1.4).

Table 1.4 Genotype frequencies with random mating

Eggs / Sperms	A (p=0.6)	a (q=0.4)

A (p=0.6)	AA (P^2=0.36)	Aa (pq=0.24)
A (q=0.4)	Aa (pq=0.24)	Aa (q^2 =0.16)

Genotype ratio :- 0.36 AA : 0.48 Aa : 0.16 aa

or 36%AA: 48%Aa : 16%aa

Therefore, in a checker board, the distribution of genotypes produced by random mating in the next generation can be expressed as:

$P^2 + 2\ pq + q^2 = 1$

(And this is the mathematical representation of Hardy-Weinberg's law)

Where, 'p' represents the proportion or frequency of gene 'A' and 'q' of its allele 'a' in a large population. The above equation is actually an expression of the binomial theorem $(p+q)^2$, where p+q=1. By comparing the genotype proportions used in the foregoing examples with the Hardy-Weinberg's equation it can readily be recognized that p^2 represents the frequency of one of the homozygous genotypes (eg. AA), q^2 for that of the other (aa) and 2pq for that of the heterozygous one (Aa).

Thus, a population in which a gene frequency remains constant generation after generation is said to be in a genetic equilibrium for that gene. In the above case, as the frequencies of gene 'A' and 'a' remain constant, the condition listed for Hardy-Weinberg law is held true in the above population.

Salient features of Hardy-Weinberg's law: According to Hardy-Weinberg's law, the gene and genotype frequencies of each allele in a population remain at an equilibrium (static) generation after generation, if that population exhibits the following attributes:

1. Random Mating

The first attribute is **random mating**, which means that the probability that two genotypes will mate is the product of the frequencies (or probabilities) of the genotypes in the population. If the *MM* genotype makes up 90% of a population, then any individual has a 90% chance (probability =0.9) of mating with a person with an *MM* genotype. The probability of a *MM* by *MM* mating is (0.9) (0.9) or 0.81.

Deviations from random mating come about for two reasons: choice or circumstance. If members of a population choose individuals of a particular phenotype as mates more or less often than at random. The population is engaged in **assortative mating**. If individuals with similar phenotypes are mating more often than at random, **positive assortative mating** is in force; if matings occur between individuals with dissimilar phenotypes more often than at random, **negative assortative mating,** or **disassortative mating**, is at work.

Deviations from random mating also arise when mating individuals are either more closely related genetically or more distantly related than individuals chosen at random from the population. **Inbreeding** is the mating of related individuals, and **outbreeding** is the mating of genetically unrelated individuals. Inbreeding is a consequence of pedigree relatedness (e.g. cousins) and small population size.

One of the first counterintuitive observations of population genetics is that deviations from random mating alter genotypic frequencies but not allelic frequencies. Envision a population in which every individual is the parent of two children. On the average, each individual will pass on one copy of each of his or her alleles. Assortative mating and inbreeding will change the zygotic (genotypic) combinations from one generation to the next, but will not change which alleles are passed into the next generation. Thus genotypic, but not allelic, frequencies change under non-random mating.

2. Large Population size

Even when an extremely large number of gametes is produced in each generation, each successive generation is the result of a sampling of a relatively small portion of the gametes of the previous generation. A sample may not be an accurate representation of a population, especially if the sample is small. Thus, the second attribute of the Hardy-Weinberg equilibrium is that the population is infinitely

large. A large population produces a large sample of successful gametes. The larger is the sample size, the greater the probability that the allelic frequencies of the offspring will accurately represent the allelic frequencies in the parental population. When populations are small or when alleles are rare, changes in allelic frequencies take place due to chance alone. These changes are referred to as **random genetic drift**, or just **genetic drift**.

3. Absence of evolutionary forces

The gene frequency will remain static only in the absence of the following evolutionary forces like:

a. No Mutation or migration

Gene and genotypic frequencies may change through the loss or addition of alleles through mutation or migration (immigration or emigration) of individuals from or into a population. So, the third and fourth attributes of the Hardy-Weinberg equilibrium are that neither mutation nor migration causes such allelic loss or addition in the population.

b. No Natural selection

The final attribute necessary to the Hardy-Weinberg equilibrium is that no individual will have a reproductive advantage over another individual because of its genotype i.e. all genotypes in a population shall reproduce equally successfully. In other words, no natural selection is occurring. The absence of selection means every gamete is viable, every gametic union (zygote) also survives and gametes or zygotes are not segregated into classes of varying survival value. Under such conditions the genetic composition of population will remain unchanged.

In short, the following major assumptions/ attributes are necessary for the Hardy-Weinberg Equilibrium to hold:

1. In a population, the mating is a completely random phenomenon.
2. The equilibrium in the gene and genotype frequencies occur only in large sized populations. In small populations, gene frequencies may be unpredictable.
3. All the genotypes in a population reproduce equally successfully.
4. According to Hardy Weinberg's law, particular alleles will be neither differentially added to nor differentially subtracted from a population (i.e. natural selection or differential reproduction is not operating).

5. When population is in equilibrium there is no possibility for evolutionary change and hence, there is no evolution.
6. Equilibrium tends to conserve gain or modifications in the gene pool that have been introduced during past.
7. Equilibrium maintains recessive genes in the population.
8. Equilibrium maintains heterozygosity in the population.
9. The gene and genotype frequencies of each allele in a population remain at an equilibrium generation after generation.

From the above discussion, it becomes clear that populations exhibiting genetic equilibrium or following Hardy-Weinberg's law are static with zero evolutionary rates. That means such populations do not evolve and any population evolving is not static and does not exhibit genetic equilibrium.

From the foregoing description, it is apparent that if the conditions specified in the Hardy-Weinberg's law are disturbed either singly or in combination, then the genetic equilibrium of a population becomes altered and there may occur an evolutionary change. Thus, when natural selection (or differential reproduction) occurs then mating does not remain random, when genes also mutate; and/or when populations are small, then also evolution occurs.

Significance of Hardy-Weinberg's Law

The Hardy-Weinberg's law is important because it explains the situation in which there is genetic equilibrium and no evolution. Thus:

1. It provides a theoretical baseline for measuring evolutionary change.
2. Equilibrium maintains heterozygosity in the population.
3. The equilibrium tendency tends to conserve gains which have been made in the past and also avoid too rapid changes.
4. Equilibrium prevents evolutionary progress.

The Hardy-Weinberg's law has also proved to be a highly useful tool in studies of population genetics. It has been especially helpful in human genetics, where controlled test mating does not exist, for clarifying the mode of inheritance of certain traits.

1.3 Destabilizing forces Or Evolutionary Factors/Forces

Hardy-Weinberg's law provides a situation where the genes in the population have reached the equilibrium and the gene pool is constant. It means there will be no change and no evolution. But it has been observed in nature that over a long period of time this equilibrium is disturbed and changes occur on account of several forces. Those factors which operate on the gene pool of a population to change the genetic equilibrium of its genes actually bring about evolution. There are different destabilizing forces operating in a population which may alter or change gene frequency. These are:

1.3.1 Natural selection

1.3.2 Mutation

1.3.3 Isolation

1.3.4 Genetic drift

1.3.5 Migration

1.3.6 Meiotic drive

1.3.1 Natural selection or differential reproduction

The gene mutations and the variations caused due to the chromosomal mutations are often harmful and many a times deleterious in homozygous state. Therefore, the individuals possessing them are eliminated from the population. But all the mutations are not harmful and some of them and their combinations are found to be beneficial to the individual possessing them and they can utilize the environment more successfully. Natural selection favors such genes, which assure the highest level of adaptive efficiency between the population and its environment. When two or more gene combinations are present, selection favors increased reproduction of the gene combinations which are found to be most efficient under the environmental circumstances. The differential rate of reproduction is caused on account of greater survival of the zygotes and gametes possessing, beneficial combinations. These contribute proportionately greater percentage of one type to the gene pool of the next generation. Thus, natural selection is **differential reproduction** of genes. It changes the equilibrium of these genes in the gene pool of the population. The impact of total environment on the reproduction of gene combination is known as **natural selection**. Natural

selection moulds the genetic variations present in the population, but it cannot directly produce new genes or new gene combinations.

Thus, Selection is the main force which can alter gene frequencies within a population bringing about evolutionary changes. *Selection* can be defined as all those effects occurring during the life cycle that contribute to differential survival of an allele in the reproduction of the population. Selection may occur as differential survival of different alleles during gametogenesis or during embryogenesis and development. The relative strength of selection varies with the amount of advantage available. The possibility that a particular phenotype will survive and reproduce is a measure of its *fitness*. Thus, fitness refers to total reproductive potential or efficiency. It is expressed in relative terms by comparing a particular genotype / phenotype combination with one regarded as optimum. The difference between the fitness of a given genotype and the optimum one is called as the *selection coefficient or constant (s)*. For example, or a particular phenotype expressed by genotype 'aa' , only 99 out of 100 individuals the produce successfully , then =0.01

And if the 'aa' genotype is a homozygous lethal, then S=1.0

And the 'a' allele is transmitted only in the heterozygous state.

Selection acts on the genotype/phenotype combination and they are also passed on to the next generation. Such quantitative traits may cause change in physical characteristics such as height or weight. Selection for such traits can be classified as (i) directional, (ii) Stabilizing or (iii) disruptive.

i) **Directional selection**: In this, the trait is polygenic and the most extreme phenotype that the genotype can express will appear in the population only after prolonged selection. In nature, directional selection can occur when one of the phenotype extremes becomes selected for or against, usually as a result of changes in the environment. For example, domestic corn kernel for oil content.

Stabilizing selection: It tends to favor intermediate forms, with both phenotype extremes being selected against. Generally, stabilizing selection represents a situation where a population is adapted to its environment. For example, human birth weight and survival.

ii) **Disruptive selection**: It is selection against the intermediate and for both phenotypic extremes. It may be taken as the opposite of stabilizing selection. For example, shell pattern in limpets.

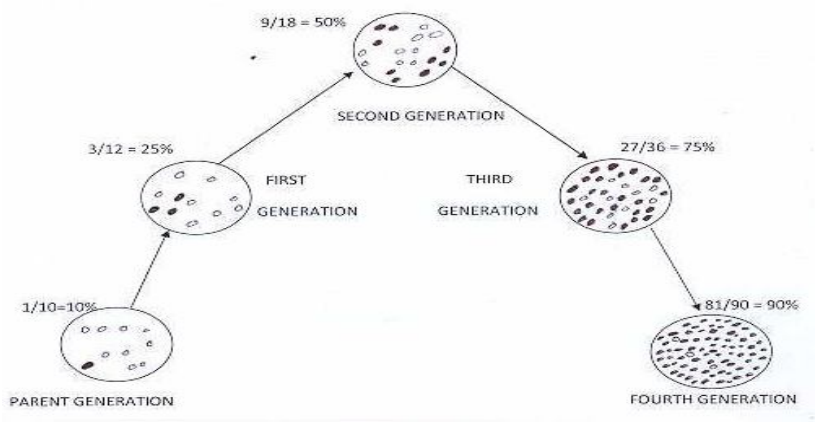


Figure 1.1: Process of Natural selection (Differential reproduction) (Black dot shows variation originating in parental generation, and white dots represents non-variant members

One of the most convincing examples of the power of selection to change the genetic constitution of a population given by Kettlewell is that of English peppered moth- *Biston betularia* for industrial melanism. These moths were grayish–white with black spotting before 1800 but in some populations today totally black forms occur. It has been observed that the difference in coloration between the two forms is regulated by a single locus with two alleles and that the dark form is due to a single dominate allele. The allele that determines the black (or melanic) form is dominant to the allele that determines the typical grayish coloration. Initially the dominant mutant gene was disadvantageous and was maintained at an extremely low frequency. As a result of industrialization, the mutant gene was favored by natural selection, but following the adaptation of laws to restrict environmental pollution in 1964, the frequency of non-melanic forms started to increase significantly.

The role of natural selection in a theoretical population of organisms can be illustrated as under- suppose a population was found to contain two alleles (a_1 & a_2) at locus 'a'. The gene frequency for a_1 is (p=0.9) and for a_2 is (q=0.1). The population at equilibrium will have constant gene frequency. But, if this population

is exposed to selection pressure on account of a change in the physical environment, and it is found that individuals homozygous for a_2a_2 could survive only 80 out of 100. It means that selection coefficient 'S' for a_2 is 0.2. The genotype a_1a_1 and a_1a_2 are not affected by selection, only the individuals with genotype a_2a_2 produces fewer offspring's in comparison to a_1a_1 and a_1a_2 . And if this continues for several generations a stage will come when gene a_2 will be eliminated from the gene pool of the population. This is how natural selection operates eliminating the non-suitable or harmful genes and encouraging the favorable ones. It has been noted that the higher the gene frequency of the deleterious mutations, the more rapidly its alleles are removed from the population.

1.3.2 Mutation

Mutations or changes in genes form the ultimate raw material of evolution. Mutations are changes in the chemical composition of genes. Mutations or changes in genes form the ultimate raw material of evolution. Within a population, new combinations in the progeny are produced due to reshuffling of gene pool in each generation. But the number of possible combinations is very large, so the members of the present population represent only a small fraction of all possible genotypes. Though there are a large number of new combinations, yet no new allele is produced. *Mutation* alone acts to create new alleles and is a force in increasing genetic variations. During mutation, the genetic variation would increase to a point frequency of 0.5 and then decreases as the mutant allele approached fixation at a frequency of 0.0 or 1.0. A mutation in actual would not replace the original allele. A mutation rate of 10^{-5} per locus per generation would mean that it would take a very large number of generations for the transition to fixation of the allele. As the mutant type becomes more common, *back mutation* will slow the increase in frequency of the mutant allele.

Mutation acts as a significant force in altering the gene frequencies. For this, the rate at which mutations are produced is measured. Generally mutations are recessive, so it is difficult to determine mutation rates directly in diploid individuals. Indirect methods such as probabilities or statistics can be employed. However, for some dominant mutations, a direct method can be applied. Following are the conditions for getting accurate mutation rate:

1. The allele must produce a well-distinct phenotypic trait clearly distinguished from similar traits produced by recessive allele.

2. The trait should never be produced by any mutagen such as chemicals.
3. The trait must be fully expressed or completely penetrant.

The Mathematical research indicates that the number of expected alleles in a population resulting from mutation alone is directly proportional to the population size. The smaller the population size, fewer is the number of alleles that can be maintained in the population by the force of mutation alone. Even if the rate of mutation was increased through exposure to higher levels of radioactivity or chemical mutagens, the impact of mutation on gene frequency would be extremely weak.

Role of Mutations in evolution

Mutations provide the bulk of hereditary variations which furnish the raw material on which natural selection operates. The mutations may be harmful, neutral or advantageous. The harmful or less useful mutations are gradually eliminated by natural selection due to differential reproduction in the population. The useful mutations are preserved and established in the population. These keep on accumulating generation after generation to bring about divergence in the naturally breeding populations.

When environment changes, the adaptedness of the inhabitant are disrupted and the harmony between environment and its inhabitants can only be established again by changing the genotype. The changing of genotype is brought about by mutations. Some mutants may prove useful in new environments and replace the old normal genes of the population forming the new adaptive forms.

Mutations are very important for the survival of the species. A living species that would suppress mutation process might gain a temporary advantage in an unchanged environment. But when the environment changes and some of the mutants found to be better fitted than the normal, a non-mutable species would be the loser and the mutated species with the accumulation of new mutations will be selected by the natural selection.

Mutation and Selection

The influence of a mutation affecting the frequency of a particular gene in a population is called *mutation pressure*. And the influence of selection altering the gene frequency in population is called *selection pressure*. Since, both mutation and selection pressure are operating together in natural population, and so both must be

recognized in defining population trends. In an unchanged environment, mutant individuals usually are at a disadvantage when compared with non-mutant forms. A few changes might be inconsequential or natural, but the chance of making an already well-adjusted organism better to meet the conditions of an environment by random change is remote indeed. The random origin of newness in higher organisms will not alone explain population changes, but it is a significant factor when combined with natural selection and reproductive isolation.

For example, a single bacterium *Escherichia coli* in an environment where penicillin is present, should undergo a spontaneous mutation and become resistant to penicillin; it might give rise to an entire population of penicillin resistant bacteria in a short period of time. Here, penicillin is not a mutagen but a selective agent. Animals as far up the evolutionary ladder as insects have been found to undergo chance mutations that happened to make them more fit for the particular environment in which they were located.

Pathogenic micro-organisms causing epidemic disorders of man and other animal probably respond in the same way. Mutations occur as frequently in higher, sexually reproducing individuals as in micro-organisms, but the numbers are very few to allow many mutations to become established even in long periods of geologic time. In higher organisms, mutation pressure alone is not a major factor in evolution. Under specific conditions, mutation and selection are forming a balance with regard to gene frequencies. In sexually reproducing organisms, recombinations of genes are already present in a population. The genotype of each individual incorporates genetic contributions from many pre-existing members of the species. With time passing, such mutations become widely distributed throughout the entire population.

1.3.3 Isolation

Isolation results in the genetic diversity between two populations so that each acquires new mutations and is acted upon by forces like genetic drift natural selection etc. That is, isolation assists in preserving mutations in the populations and thus in the splitting of the species into incipient races. Mutations and variations assisted with natural selection introduce difference in the population and isolation helps in the accumulation of variations leading to divergence among the population of a species which after attaining reproductive isolation become independent species.

Isolation has been recognized as one of the most important factors in the process of speciation or species formation. It helps in **allopatric speciation** (evolution of species occupying different areas) as well as **sympatric speciation** (evolution of species occurring in the same area).

A. Allopatric or parapatric Speciation

A species population usually has a discontinuous distribution. Even a more or less continuously distributed species having a wide range, does not form one large randomly mating population. Moreover, territoriality among animals tend to split the species population into a number of allopatric or parapatric breeding populations. Therefore, as a general rule, a species is composed of a number of allopatric breeding populations, each physically separated to some extent from others and pursuing its own independent evolutionary path. Even though, initially the genetic composition of these populations may be very similar, no two environments are likely to be biologically or physically identical. Selection plus the random aspects of mutations and in small populations the genetic drifts will bring about divergence in the hereditary characteristics. If these populations remain separated for a long time, and if the interacting forces of evolution, particularly selection operate to produce divergence, allopatric species are formed from allopatric populations due to the establishment of reproductive isolation. The allopatric species originate by the interplay of both geographical as well as reproductive isolation.

B. Sympatric Speciation

Sympatric species originate by the instantaneous development of reproductive isolation between segments of species population due to sudden change in their genotype. As a result the species population is splitted into two or more reproductively isolated populations. Once reproductive isolation is established, these populations follow their own evolutionary course and become sympatric species. These populations whether continue to occupy the same region or move out of the original habitat will remain distinct.

Speciation results in an increase in the existing number of species. As a first step, there is a formation of new demes and races due to migration and fragmentation. The original population tends to expand into the surrounding areas. Gene flow continues, in the early stages, between all sections of the population, but gradually,

localized subpopulations develop in the most suitable habitats. These subpopulations become isolated from each other which results in genetic divergence due to different mutation pressures, selection pressures, random genetic drift or the net effect of all three forces. After isolation, microevolution within the demes may produce widely different populations. Isolation is, therefore, the most important factor in the origin of new population and in its absence speciation is not possible.

1.3.4 Genetic drift:

It is the random fluctuation in gene frequencies, whose effect is negligible in large populations. For example, in a comparatively small population of 16 individuals consisting of four AA, eight Aa and four aa, the frequency of A(p) is 1/2 and that of a(q) is 1/2. And if there were eight males and eight females or eight pairs and each pair producing four offspring, then the result would be a new population, double the size of the first, with the gene frequencies remaining the same as the original, where p = q = 1/2. Instead of mating all 16 individuals, a specific pair might be taken to raise a new population. The pair selected could have the following different genotypes with the gene frequencies as shown in Table 1.5.

Table 1.5 showing possible mating in the F_2 population

Genotype	**Frequencies**	
	p	**Q**
AA X AA	1.0	0
AA X Aa	0.75	0.25
AA X aa	0.5	0.5
Aa X Aa	0.5	0.5
Aa X aa	0.25	0.75
aa X aa	0	1.0

Thus, the gene frequencies of p = q = 1/2 can be completely changed in one generation, if small non-random samples are selected to raise a new population. Then, *genetic drift* can be defined as the result of sampling error, plus the expression of a small population into a large population. The most convincing example of genetic drift can be seen with MN blood group series in man. It has been observed that in one population, the frequency of N gene (q) in the population was 0.75, while in another it was 0.24, almost completely the reverse. These differences could have arisen as a result of a sampling error in the original population.

Gene frequencies are subjected to functions about their mean, from generation to generation. If a population is large, the numerical fluctuations are small and have little or no effect. On the other hand, if the population is small, random fluctuations could lead to complete fixation of one allele or another. An allele, even though it might have high adaptive value, could be completely lost from a small breeding population by change alone, and an allele, with little or no adaptive value could become established or fixed by chance in a small population.

According to Hardy-Weinberg principle, in a large randomly mating population,, without selection and mutation, the gene frequency remains constant. But in a small population, the gene frequencies are found to fluctuate purely by chance and smaller the population, the greater the fluctuations. These random changes in gene frequency occurring by chance are called **Genetic drift**. These gene frequencies will continue to fluctuate until one allele is lost and the other is fixed. It means as a result of genetic drifts a new mutation arising in a small population is either lost or is fixed by chance irrespective of its utility, because in them the heterozygous gene pairs tend to become homozygous for one allele or the other only by chance rather than by selection. This may lead to the accumulation of certain disadvantageous characters and subsequent elimination of group possessing them.

Thus, genetic drift is an evolutionary force operating in small populations. It was described by **Sewall Wright** in 1931. Hence, it is also called **'Sewall Wright effect'**. It changes gene frequency in a small population purely by chance. It shows the following main characteristic features:

1. Genetic drift is an evolutionary force operating in small populations.
2. The gene frequency in small population changes purely by chance, not by selection.

3. Genetic drift may lead to the complete loss of an allele from the gene pool of a small population.
4. Genetic drift results in fixation of gene, that is, it may lead to the fixation or preserving certain genes and eliminating others.
5. A new mutation in a small population may be lost or fixed as a result of genetic drift.
6. Genetic drifts lead to preservation or loss of genes without preference, whether favorable, neutral or unfavorable. Hence, genetic drifts work against natural selection.

Genetic drifts favors homozygosity.

1.3.5 Migration

Another factor which may influence the gene frequencies is *differential migration*. In local units of a species, gene frequencies may be altered by an exchange of genes with other breeding units. This exchange is effective in changing gene frequencies if the breeding populations have been partially or completely separated for enough time to have developed markedly different frequencies for the same genes.

Difference in mutation rate and selective pressure can establish different gene frequencies in the sub-populations. Migration occurs when a large influx of organisms moves into another population and interbreeds with the latter. The phenomenon called *gene flow* takes place if one population contributes an allele to the other population.

Let us suppose a single pair of alleles as A(p) and A(q), then change in the frequency of A in one generation can be calculated as:

$P = m (pm - p)$

(Where P = change in frequency of A in one generation, m = migration rate, pm = frequency of A in immigrants, and p = frequency of A in existing population)

. Mathematically, if p = 0.4, and pm = 0.6, then 10% of the parents giving rise to the next generation are immigrants (m=0.1). Then, the change in the frequency of A in one generation will be

$P = m (pm-p) = 0.1x (0.6-0.4) = 0.1x (0.2) =0.02$

In the next generation, the frequency of A (p_1) will increase as below:

$p_1 = p + P = 0.4 + 0.02 = 0.42$

If m or p is larger, then a large change in the frequency of A will occur in a single generation. However, a balance can be obtained if p = pm, under similar conditions. From the above calculations, it becomes evident that change in gene frequency due to migration is proportional to the difference in frequency between the donor and recipient population. Since, *migration constant or coefficient* can have a wide range of values; the effect of migration can substantially change gene frequencies in populations.

1.3.6 Meiotic drive

Another factor which can bring about changes in the gene frequencies is the *meiotic drive* in which irregularity in the mechanics of the meiotic divisions takes place. Normally, an individual which is heterozygous (Aa), produces gametes A and a in equal proportions and these gametes have equal probabilities of fertilization and development. For many years, no unusual recordings were taken with respect to heredity. But now examples of systematic deviations from Mendelian ratios, which have a genetic basis, are on record.

For example, in *Drosophila*, the SD (*segregation distorter*) locus in chromosome II has two known alleles, of which one is wild type and other a distorter of the wild allele. In the presence of the homozygous wild-type allele, normal segregation of the chromosomes occurs. Heterozygous males having the mutant allele, under certain environmental conditions, shows a marked departure from the 1:1 ratio. The mutant allele apparently interacts with its wild type homologue, causing it to fragment and behave irregularly during spermatogenesis. Due to which only a few sperms contain the normal allele. During meiosis, duplication of chromosomes or its parts may occur producing irregularities in the gametes.

Similarly, in female *Drosophila*, in heterozygous conditions, different chromosomes may have preferential segregation. One member of the pair may be retained in the egg and other given to the non-functional polar body. If the two homologues are of unequal length, the shorter one is usually retained in the egg nucleus. Chromosomes that have undergone structural changes are often extruded.

Sex ratio has also been found to be altered by abnormal segregation of X and Y chromosomes. In *Drosophila pseudo obscura*, abnormal spermatogenesis occurs resulting in the failure of the X and Y chromosomes to pair. The Y chromosome

degenerates and the X undergoes an extra division. As a result, X chromosomes are segregated to all sperms. When these males, carrying only X in their sperms, are crossed, excessive numbers of females are produced in their progeny.

Meiotic drive acts as significant means of evolution. With it, even most favorable genes would not be perpetuated if the chromosomes in which they were carried were systematically excluded from gametes.

1.4 **Molecular Evolution**

Evolution is a process of change. Molecular evolution is evolution viewed at a molecular level. It can also be defined as a discipline of biology that utilizes molecular data (usually DNA or protein sequences) to address evolutionary questions. At the molecular level, this process involves the insertion, deletion, or substitution of nucleotides in the DNA. If the DNA encodes a polypeptide, these events may cause a change in the amino acid sequence. Over time, such changes can accumulate, leading to a molecule that bears little resemblance to its progenitor. Recent advances in molecular biology have made it possible to determine the nucleotide sequences of DNA and the amino acid sequences of polypeptides. By comparing related sequences, the molecular details of evolution can be analyzed.

It usually refers to incremental changes in the nucleotide sequence of DNA. It also refers to changes in the DNA of chromosomes which take place over the history of a species and distinguish the species from its ancestors. Therefore, it can be referred to as natural history of DNA. The evolution of molecular sequences relates more to general evolution. Molecular evolution takes place not only in the gene sequences coding for structural, enzymatic or other gene products, but also in the DNA with no known function (like 'Junk DNA'). Thus, the DNA of lineage might 'evolve molecularly', even though the phenotype of descendents remains constant.

Molecular evolution with the help of molecular data answers the evolutionary questions of whole genome sequences. The molecular data are usually DNA or protein sequences but may also include other types of data, such as the three-dimensional structure of some biochemical properties of a protein. DNA and proteins are referred to as *macromolecules* because they are much larger than molecules such as oxygen or ethanol. The field addresses diverse questions,

ranging from traditional questions of life sciences to new questions driven by recently emerged data, such as whole genome sequences.

The evolutionary questions considered in molecular evolution can be divided into two categories.

The first one concerns the process and mechanisms of evolution- in other words, how and why evolution of macromolecules has occurred. For example, we may want to know how and why hemoglobin, the oxygen carrier in animal blood, has evolved through time.

The second one concerns the evolutionary history of organisms, genes, or genomes. In particular, the molecular data have become a powerful tool to elucidate the differences and relatedness between species. For example, humans thought that they were very different from the apes, but DNA sequences data have shown that human and chimpanzee are actually closer to each other than either of them is to gorilla. When the study of molecular evolution deals with this category of questions, it is known as *evolutionary or molecular systematic or phylogenetics.* Therefore, molecular evolution is a multidisciplinary subject, requiring training in molecular biology, evolution, statistics, mathematics, and computer science.

Birth of Molecular Evolution: Although molecular evolution is a relatively young branch of science, attempts to use molecular techniques to study phylogenetics relationships among mammals were made before the dawn of 20^{th} century.

In the 1960s new technology revealed the amino acid sequences of proteins. By comparing the sequences of proteins such as hemoglobin and cytochrome C from different species, and using paleontological estimates of times to last common ancestors, biologists could estimate the rate of evolutionary changes in the protein sequences. Based on the analyses of protein sequence data of hemoglobin from different mammals, **Zuckerkandl and Pauling (1965)** proposed the molecular clock hypothesis, postulating that for any given protein the rate of molecular evolution is approximately constant in time and across evolutionary lineages. If macromolecules evolve at a constant rate, they can be used for estimating the dates of species divergence and for reconstructing phylogenetics relationships among organisms. However, the concept of rate constancy is diametrically opposite to the

observed 'erratic tempo of evolution' at the morphological and physiological levels. Therefore, the hypotheses also provoke a great controversy.

Causes of molecular evolution: There are three main causes of molecular evolution. They are: (a) Mutation: It is the ultimate source of genetic variation, though its role in evolution is minor.

(b) Random genetic drift: Its role is negligible in evolution.

(c) Natural selection: It plays a dominant role in evolution.

The Neutral theory of Molecular Evolution: This view has been well accepted for morphological evolution, but its validity for molecular evolution has been challenged. **Kimura (1968)** and **King and Jukes (1969)** proposed that most of the evolutionary change at the molecular level occurs as a consequence of random genetic drift of mutant alleles that are selectively neutral or nearly neutral. A mutation is said to be nearly neutral if its effect on selection co-efficient is smaller than ($1/2N_e$) of the effective size of the population. This proposal is known as the *Neutral Mutation hypothesis* or the *Neutral Theory of Molecular Evolution*. Despite much progress, how much of the genetic variability measured by molecular methods is produced by random genetic drift and how much by adaptive evolution is still not clear.

Controversies of neutral mutation hypothesis:

The hypothesis raised certain controversies such as -

1. Are certain types of mutation selectively neutral?
2. Whether the approximate rate-constancy observed was really true and could be used to support the neutral mutation hypothesis?
3. The emphasis on slightly deleterious mutations which are frequent. Whether they contribute significantly to polymorphism and evolution?

The controversy over the neutral mutation hypothesis had two strong impacts on molecular evolution and population genetics ie.

(a) It has become the general recognition that the effect of random drift cannot be neglected when the evolutionary dynamics of molecular changes is considered.

(b) It accelerated the fusion of molecular evolution and population genetics.

Study of molecular evolution

The study of molecular evolution uses methods such as PCR (polymerase chain reaction), DNA sequencing, and various techniques for isolating or localizing gene products such as RNA transcripts and proteins. The information obtained by these methods is used in evolutionary studies.

1.4.1 Gene evolution

Genes are perpetually added to and deleted from genomes during evolution. And the new genes rapidly change existing genetic systems that govern various molecular, cellular, and phenotypic functions. Thus, it is important to understand how new genes are formed and how they evolve to be critical components of the genetic systems that determine the biological **diversity of life**.

Genes evolve at different rates depending on the strength of selective pressure to maintain their function. **Exon shuffling** is a mechanism by which new genes are created. This can occur when two or more exons from different genes are combined together or when exons are duplicated. Exon shuffling results in new genes by altering the current intron-exon structure. This can occur by any of the following processes: transposon mediated shuffling, sexual recombination or illegitimate recombination. Exon shuffling may introduce new genes into the genome that can be either selected against and deleted or selectively favored and conserved.

Novel genes can also arise from **non-coding DNA**. For example, the origin of five new genes in the *D. melanogaster* genome from non-coding DNA.

Over hundreds of millions of years, a single gene can give rise not just to one new gene, but to hundreds via **gene duplication**. In humans around 400 genes are coding for smell receptors, for example, all of which derive from two original genes in a very early fish living around 450 million years ago. The evolution of this gene "family" has been a complicated process. The genome studies showed that, rather than steadily acquiring genes for new smell receptors, there have also been extensive losses of genes during the evolution of mammals – a process called **"birth-and-death evolution"**. This has led to great variation between mammals. For eg, dogs have more receptors than humans, with around 800 working smell-receptor genes and cows have even more, with over 1000 smell-receptor genes. According to **Masatoshi Nei** mammals only need a certain minimum number of

different olfactory receptors to have a good sense of smell. As long as animals have more olfactory receptor genes than they need, there is no natural selection and genes are gained and lost randomly.

The rate of evolution of a gene or mutation that is under selection will be very different. Similarly different genes with different functions or different parts of a gene with different functions will have different rates of evolution. Thus, different regions of DNA with different functional constraints will evolve at different rates.

1.4.2 Evolution of Gene families

A gene family can be defined as presence of two or more related genes in a genome. Or a gene family consists of groups of similar genes that have arisen by duplication from a common ancestral gene and that generally retain similar functions. For example, in olfactory receptor genes hundreds of different genes spread out over many chromosomes are present. All the copies are closely related (homologous) and they all descend from a common ancestral stalk following a gene duplication event. Thus, gene families consist of groups of similar genes that have arisen by duplication from a common ancestral gene and that generally retain similar function.

There are several gene families in eukaryotic genomes which appear to consist of repeated sequences including non-coding DNA and functional genes, e.g., globin gene family and histone gene family. The gene families have been formed gradually through a continued process of unequal crossing over, followed by differentiation and other mechanisms. Thus, evolution of gene families especially that of such non-coding repetitive DNA sequences are the result of molecular drive, which is responsible for origin and spread of biological variations.

The evolution of gene families has been studied for over 50 years. Three different modes of evolution (Fig. 1.2) have been studied which are as follows:

1. First mode of gene family evolution is called as **divergent evolution**. In divergent evolution, each of the genes evolve independently after the duplication. This is the most common mode of evolution for gene families, especially if the A gene and the B gene are separated in the genome (i.e., on different chromosomes). For example , the evolution of α-and β-globin genes in vertebrates.
2. Second mode of gene family evolution is called as **concerted evolution**. In

concerted evolution, after gene duplication, instead of showing two independent phylogenies, the A and B genes in each species are much more closely related to each other than to family members in any other species. For example, evolution of ribosomal RNA genes in all species. In concerted evolution the pair of genes (A and B) evolved in a concerted manner. They communicate to each other and when a mutation occurs in one gene it is transferred to the other so that both genes change in the same direction. Gene conversion represents the most important mechanism of concerted evolution. It is a form of recombination where the sequence of one gene "converts" the other. It explains how the two genes can communicate. Gene conversion can take place between any two homologous genes in the genome but it is much more common between two homologous genes that are adjacent to each other, especially if they are transcribed in the same direction as a result of tandem duplication. Therefore, concerted evolution is the process by which a series of nucleotide sequences or different members of a gene family remain similar or identical through time. It can be caused by unequal crossing over and gene conversion.

3. The third mode of gene family evolution is a combination of the patterns present in divergent evolution and in concerted evolution. If duplications of family members occur frequently then this gives rise to the birth of new genes. Newborn genes will closely resemble one another, as in concerted evolution. The number of family members does not keep expanding because some of the genes become inactivated—they become pseudogenes and they die. The resulting pattern of evolution will look like a mixture of divergent and concerted evolution. The mode is called **"birth-and-death**." In birth-and-death evolution, some genes survive in a lineage and some genes are lost. The birth and death of genes can be random or it can be under selection. In it, all members of the gene family in the ancestor will not show up in all species descending from the common ancestor, and that sometimes several members of the gene family will be much more closely related than those expected from divergent evolution.

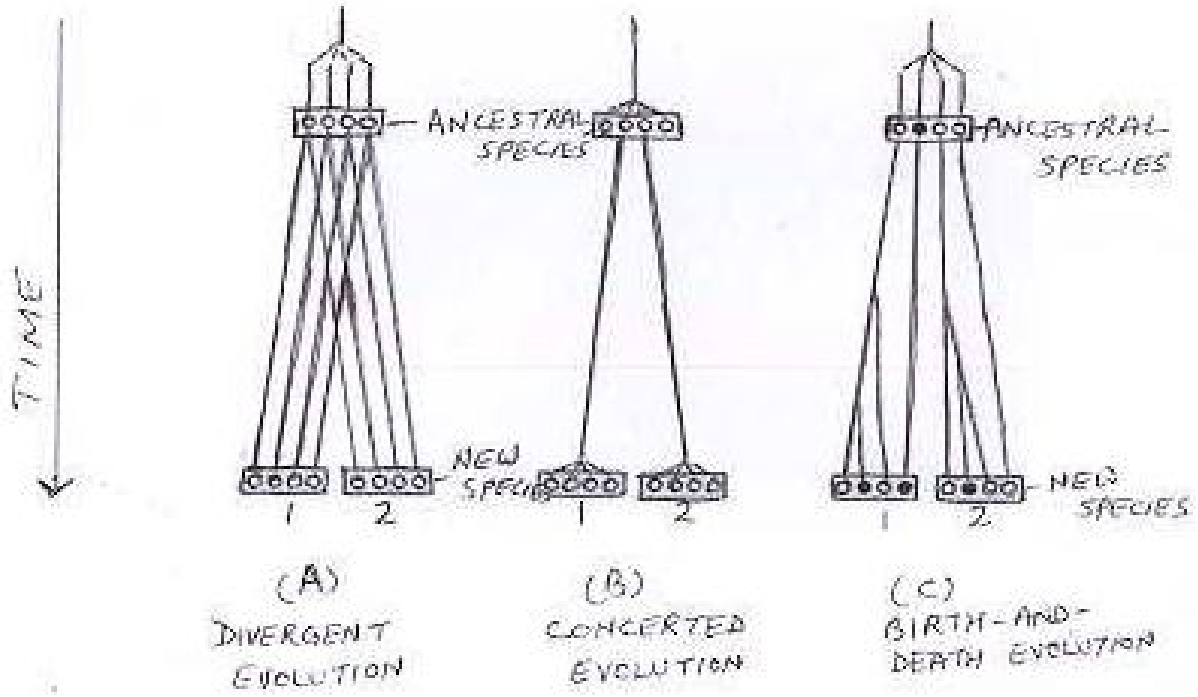


Figure 1.2: Different modes of gene family evolution

Concerted evolution is commonly observed for gene families which originated a long time ago, however there are many different types of multigene families, from uniform to diverse. The rate of homogenization by unequal crossing-over, gene conversion, etc. has been evolutionarily adjusted for each gene family. When new functions are needed by organisms, gene families may evolve into superfamilies, in which no further concerted evolution takes place, and each member of the family may acquire an indispensable function. For example -The homeo box-containing gene family.

In eukaryote genomes, many kinds of gene families are present. Gene duplication and conversion are thus, the sources of the evolution of gene families, including those with uniform members and those with diverse functions. According to the population genetics theory on identity coefficients among gene members of a gene family, the balance between diversification by mutation, and homogenization by unequal crossing over and gene conversion, plays an important role. Also, evolution of new functions is due to gene duplication followed by differentiation. Positive selection is necessary for the evolution of novel functions. However, many examples of current gene families show that both drift and selection are at work on their evolution.

1.4.3 Molecular drive

Molecular drive is a process of changing the average genotype of a population. It spreads new variant through a multigene family (homogenization) and through a sexual population (fixation). This process is different from chance, natural selection and genetic drift. It is a consequence of a variety of genomic mechanisms of turnover which have the effect of simultaneously spreading new variants. The cohesive dynamics of the process and the opportunity for molecular co-evolution, in molecular drive permit new ways of the progress of biological evolution. Molecular drive applies mainly, to long-term changes in the structures and functions of multigene families as well as of the meaningless repetitive DNA sequences as cohesive units of evolution, independent of natural selection and genetic drift. The genomic mechanisms of turnover, which underlie molecular drive, can result in the conservation or divergence of DNA sequences. Hence, observed differences in the levels of genetic conservation and divergence do not necessarily signify selection and drift, respectively. Molecular drive encompasses internal dynamics of eukaryotic genomes.

Thus, molecular drive is a process by which mutations are able to spread through a family (homogenization) and through a population (Fixation). A variety of mechanisms of non-reciprocal DNA transfer within and between chromosomes are responsible, such as-Gene conversion, Unequal crossing over, Transposition, Slippage replication, and RNA-mediated exchanges.

These mechanisms can induce a gain or loss of a variant gene in an individual's lifetime in their different ways, resulting in non-Mendelian segregation ratios. Any type of continued gain or loss leads to an accidental spread of one variant gene both within the family and the population. For example, a family of 100 genes may contain 99 wild-type copies and one mutant copy in a given individual. Anyone of the molecular drive mechanisms can lead to a change in ratios to either 98:2 or 100:0. In the former case, the additional variant copy may be on the same chromosome as the first variant, or on another homologous or non-homologous chromosome, depending on the chromosomal distribution of the family. If the gain is due to a non-reciprocal transfer between chromosomes, then, as a consequence of sex, the two copies will enter into two new individuals at the next generation, in each generation further random gain and loss may take place.

In the early stages, there is a high probability that the new variant is lost from the population. There is some probability that the variant copy would accidentally replace the wild-type copies in the family throughout the population. Or, in the case of a *de novo* amplification of a new gene among all individuals of the population, there will be an accidental accumulation of a new gene among all individuals of the population. The probability of, and the time taken for, homogenization and fixation are dependent on the population size, the size of the gene family and the rates of non-reciprocal transfer. Both the probability and the time are dramatically altered when there is a bias in a given turnover mechanism favoring a new variant, such as biased gene conversion. The meiotic disjunction of chromosomes and their redistribution in the next generation is the basic mechanism for spreading new copies of variant member genes. Therefore, homogenization and fixation are inextricably linked, because during sexual process the chromosomes are continually shuffled and assorted into new combinations at each generation.

Molecular Drive and Natural Selection:

Interaction between molecular drive and natural selection can be assessed by evaluating the phenotypic differences between individuals of a population at any given generation with respect to the extent of change in a gene family. Non-reciprocal transfer of DNA sequences results in the formation of a gene family. Consideration of the known slow rates of non-reciprocal exchange shows that the pattern of accumulation of a mutant gene in a Mendelian population is such that, at any given generation, most individuals have a similar copy number of the new mutation, This is because the sexual randomization of chromosomes between generations is very much faster than the rate (10^{-5} per generation) at which a new copy of the mutant gene can be formed in an individual by the genomic turnover mechanisms. So, the mean copy number of a mutant gene per individual in a population moves slowly from zero to full homogenization with a variance at each generation that is small in relation, to the total difference between the initial and final states.

The interaction between molecular drive and natural selection rests on knowing experimentally whether the cohesive change in phenotypes permits;

(a) The establishment of a new relationship between the organisms and then existing environments, and

(b) An internal adjustment (molecular co-evolution) between the interacting molecules involved with a given function during ontogeny.

Under constrained conditions of ecology and ontogeny of a species neither (a) nor (b) can take place. In such a case selection interferes in the process of molecular drive at some threshold level at which enough copies of the variant genes have accumulated. Over some defined, relatively short time-span, the population then would suffer a collective fate as the threshold is reached at each generation. These are the population never seen. Only those populations which are winners in the evolutionary game are observed which survived the molecularly driven changes in their gene families and an attendant effect on their ontogeny and their relationship with the environment.

1.4.4 Assessment of Molecular variation

The study of molecular evolution uses the following different methods and the information obtained by these methods are used in the assessment of molecular variation.

1. The physical basis of molecular variation

The hereditary information in all organisms is carried in DNA, except RNA viruses. So, a difference in any of the molecular markers used in study (and of genetically-based morphological, behavioral, or physiological traits) is associated with some difference in the physical structure of DNA, and molecular evolutionists study a variety of its aspects.

A. Nucleotide sequence

A difference in nucleotide sequence is the most obvious way in which two homologous stretches of DNA may differ. The differences may be in translated portions of protein genes (exons), portions of protein genes that are transcribed but not translated (e.g., introns, 5' or 3' untranslated regions), non-transcribed functional regions (e.g., promoters), or regions without apparent function.

B. Protein sequence

Because of redundancy in the genetic code, a difference in nucleotide sequence at a protein-coding locus may or may not result in proteins with a different amino acid sequence. Some loci code for RNA that has an immediate function without being translated to a protein, e.g., ribosomal RNA and various small nuclear RNAs.

2. Assessing molecular variation

The diversity of laboratory techniques used to assess molecular variation is even greater than the diversity of underlying physical structures. Various techniques involving direct measurement of aspects of DNA sequence variation are by far the most common techniques today. Some of the techniques that have been most widely used are as follows:

A. Immunological distance

Some molecules, notably protein molecules, induce an immune response in common laboratory mammals. The extent of cross-reactivity between an antigen raised to humans and chimps, for example, can be used as a measure of evolutionary distance. The immunological distance between humans and chimps is smaller than it is between humans and orangutans, suggesting that humans and chimps share a more recent common ancestor.

B. DNA-DNA hybridization

The rate and temperature at which DNA from two different species anneal shows the average percent sequence divergence between them. The percent sequence divergence can be used as a measure of evolutionary distance.

Immunological distances and DNA-DNA hybridization were once widely used to identify phylogenetic relationships among species. Neither is now widely used in molecular evolution studies.

1.5 Summary

Evolution is a change in the gene pool of populations which may arise due to genetic recombination, or mutations. If this change happens to be adaptively advantageous and the population is small then increased or relaxed selection pressure will disturb the genetic equilibrium of the existing gene frequencies. Consequently, this process of natural selection operating through differential reproduction will bring about a rapid or slow propagation of the genetic innovation throughout the population and a new trait will become established.

Thus, any change in the genetic equilibrium, that is, change in gene frequencies of a population through non-random mating, mutations and/or small population size (genetic drift) may cause an evolutionary alteration.

Many gene and supergene families exist in eukaryote genomes: gene families with uniform copy members like genes for ribosomal RNA, those with variable members like immunoglobulin genes, and supergene families such as those for various growth factor and hormone receptors. These examples show that gene duplication and subsequent differentiation are extremely important for the evolution of an organism. In particular, gene duplication is the primary mechanism for the evolution of complexity in higher organisms. Population genetic models for the origin of gene families with diverse functions showed that natural selection favors those genomes with more useful mutants in duplicated genes. Since any gene has a certain probability of degenerating by mutation, success versus failure in acquiring a new gene by duplication may be expressed as the ratio of probabilities of spreading of useful versus detrimental mutations in redundant gene copies. This shows that both natural selection and random drift are important for the origin of gene families. In addition, interaction between molecular mechanisms such as unequal crossing-over and gene conversion, and selection or drift is found to have a large effect on evolution by gene duplication.

A cluster of genes that has evolved from a single progenitor gene constitutes a gene family. Such families may expand or contract in size through a process of unequal crossing over. Sometimes a gene family evolves as a concerted unit.

1.6 Glossary

- **Evolution:** A term applied to those methods or processes and to the sum of those processes whereby organisms change through successive generations.
- **Gene pool:** The total genetic information present in all members of a species.
- **Genetic drift:** Fluctuation in the distribution of gene frequencies caused by isolation of non representative samples of the founding population.
- **Genotype:** The genetic constitution of an organism with respect to certain traits.
- **Molecular evolution:** Evolution viewed at a molecular level.

- **Mutation:** A change in the genetic material that in the homozygous condition, brings about a change in an individual 's phenotype.
- **Natural selection:** The evolutionary process by which each environment determines ("selects") which of the heritable variation arising in populations will, (because they best adapt their possessors to that environment), be the ones to be passed on to future generations.
- **Population:** A group of organisms of the same kind, usually of the same species, living in a given area.

1 .7 Self-Learning Exercise

Section -A (Very Short Answer Type)

1. Who proposed a mathematical expression for the distribution of gene and genotype frequencies in a Mendelian population?
2. What is the consequence of meiotic drive?
3. What are the conditions at which gene frequencies would remain constant during evolution?
4. What do you mean by Molecular evolution?
5. Define natural selection?
6. What is genetic drift?
7. What is Mutation?

Section -B (Short Answer Type)

1. Write a note on molecular drive.
2. Briefly explain about natural selection.
3. Mention the significance of Hardy-Weinberg law.
4. Define Mutation and its role in evolution.
5. Write down the role of isolation in species formation.
6. Define gene pool and gene frequency.

Section -C (Long Answer Type)

1. Discuss Hardy-Weinberg law of genetic equilibrium.

2. Enumerate the various destabilizing/evolutionary forces that tend to disturb the genetic equilibrium.
3. What is molecular evolution? Discuss in detail.
4. Discuss the role of genetic drift as one of the evolutionary force.
5. Explain in detail the conditions under which the gene frequency in the individuals of a population remains constant?

Answer Key of Section-A

1. **Hardy** and **Weinberg** (1908) proposed a mathematical expression for the distribution of gene and genotype frequencies in a Mendelian population.
2. Meiotic drive acts as significant means of evolution. With it, even most favorable genes would not be perpetuated if the chromosomes in which they were carried were systematically excluded from gametes.
3. In the absence of mutation, migration, and natural selection gene frequencies would remain constant during evolution.
4. Evolution viewed at a molecular leve is called as **Molecular evolution** .
5. The evolutionary process by which each environment determines ("selects") which of the heritable variation arising in populations will, (because they best adapt their possessors to that environment), be the ones to be passed on to future generations.
6. It is the fluctuation in the distribution of gene frequencies caused by isolation of non representative samples of the founding population.
7. It is a change in the genetic material that in the homozygous condition, brings about a change in an individual's phenotype.

1.8 References

- V.B.Rastogi : Organic evolution, Kedarnath Ramnath Meerut
- J.M. Smith : Evolutionary genetics ; Oxford University Press, New York
- A.P.Jha: Genes and Evolution; John Publication, New Delhi.
- R. Sisodia : Evolution and Population genetics; Paragon International Publishers.

Unit - 2

Quantifying genetic variability: genetic structure of natural populations, phenotypic variation, models explaining changes in genetic structure of populations, factors affecting human disease frequency

Structure of the Unit

2.0 Objectives

After going through this unit you will be able to understand:

- Methods for the quantification of genetic variability.
- Phenotypic variations and their types.
- Factors that changes the allelic or genetic frequencies of the population.
- Different factors which changes the human disease frequencies.

2.1 Introduction

Natural populations should maintain genetic diversity since it is essential for the long-term survival of species and provides the raw material for all evolutionary changes, allowing adaptation to environmental changes and thus decreasing extinction risk (Frankham et al. 2004).

Genetic diversity represents the total genetic variation among individuals within a population.

The concept of genetic markers is not a new one: in the nineteenth century, Gregor Mendel employed phenotype-based genetic markers in his experiments. The emergence of marker systems, for the last 40 years, closely tracked developments in molecular biology. Morphological markers were largely supplanted by biochemical markers and the latest markers were supplanted by the development of markers based on DNA polymorphisms.

A molecular marker is in essence a nucleotide sequence corresponding to a particular known or unknown physical location in the genome. Molecular markers and nucleotide sequence information allow uncovering of genetic diversity and also differentiation of individuals at the DNA level.

A prerequisite for the beginning of population studies is to detect the genetic diversity underlying phenotypic variation. Once genetic diversity is detected its distribution within and among local populations or spatial/temporal patterns can be analysed.

Several measures of genetic diversity have been developed over the years. The seminal measure of genetic diversity from molecular data has been the number of gene alternative forms or alleles at a given locus (NA), which is also known as

gene multiplicity. As multiplicity and abundance vary independently, genetic diversity can be expressed as the effective number of alleles (NE) (Kimura and Crow 1964). If the alleles show the same frequency, NE will be equal to NA.

However, NE will be a decreasing function if allelic frequency distribution is not uniform. The number of private alleles (NP) is a related measure and represents the number of alleles which can be found in only one population (Barton and Slatkin 1986).

In this way, NP is a simple measure of genetic distinctiveness. Since the number of detected alleles in a population depends on its size, it is not advisable to compare these genetic diversity parameters among local populations with different sizes. The allelic richness estimation (R) using the rarefaction method, is useful to compare the number of alleles between samples that differ in size, because it predicts the expected number of alleles if samples have the same size (Foulley and Ollivier 2006).

Forty years ago, Nei proposed the original measure of genetic diversity:

Nei's gene diversity index (h) (Nei 1973)

This parameter represents the probability that two alleles randomly and independently selected from a gene pool will represent different alleles. This index analyses allele frequency variation directly in terms of heterozygosity and can be applied to any population of all organisms (sexual or asexual, diploid or nondiploid) without consideration of the number of alleles at a given locus or the pattern of evolutionary forces because it has been formulated entirely in terms of alleles and genotypic frequencies in the population. In this way, the treatment of this index is biologically the most direct.

Finally, the extent of DNA polymorphism for a group of nucleotide sequences sampled in a population is measured by nucleotide diversity (π), which is defined as the average number of either nucleotide differences or substitutions per site.

2.2 Genetic structure of natural population

Genetic structure refers to any pattern in the genetic makeup of individuals within a population.

In the absence of genetic structure, one can infer little to nothing about the genetic makeup of an individual by studying other members of the population. When genetic structure is present, on the other hand, much can be inferred.

In trivial terms, all populations have genetic structure, because all populations can be characterised by their genotype or allele frequencies: if only 1% of a large sample of moths drawn from a single population have spotted wings, then it is safe to assume that any unknown individual is unlikely to have spotted wings.

A more complicated example arises in dense thickets of plants, where plants tend to be pollinated by near neighbours, and seeds tend to fall and germinate near the maternal plant. In such a scenario, plants tend to be more closely related to nearby plants than they are to distant plants; and yet they are more likely to breed with nearby plants than they are with distant plants. Thus an inbreeding cycle is created that perpetuates the pattern of plants being closely related to near neighbors. This is a form of genetic structure because one can infer much about the genetic makeup of any individual plant simply by studying plants in their immediate neighborhoods.

Our first step is to describe the genetic structure of a population; we need to do this before we can model what it would do over time. The genetic structure of a population is defined by the gene array and the genotypic array. To understand what these are, some definitions are necessary:

2.2.1 Definitions

- **Evolution**: a change in the genetic structure of a population
- **Population**: a group of interbreeding organisms that share a common gene pool;
- **Gene Pool**: sum total of alleles held by individuals in a population
- **Genetic structure**: Gene array and Genotypic array
- **Gene/Allele Frequency**: % of alleles at a locus of a particular type
- **Gene Array**: % of all alleles at a locus: must sum to 1.
- **Genotypic Frequency**: % of individuals with a particular genotype
- **Genotypic Array**: % of all genotypes for loci considered; must = 1.

2.2.2 Basic computations

Determining the Genotypic and Gene Arrays:

The easiest way to understand what these definitions represent is to work a problem showing how they are computed.

Consider the population shown to the right, in which there are 70 AA individuals, 80 heterozygotes, and 50 aa individuals. We can easily calculate the Genotypic Frequencies by dividing each of these values by the total number of individuals in the population. So, theGenotypic Frequency of AA = 70/200 = 0.35. If we account for all individuals in the population (and haven't made any careless math errors), then the three genotypic frequencies should sum to 1.0. The Genotypic Array would list all three genotypic frequencies: f(AA) = 0.35, f(Aa) = 0.40, f(aa) = 0.25. A Gene Frequency is the % of all genes in a population of a given type. This can be calculated two ways. First, let's do it the most obvious and direct way, by counting the alleles carried by each individual. So, there are 70 AA individuals.

	AA	Aa	aa	
Individuals	70	80	50	(200)
Genotypic Array	70/200 = 0.35	80/200 = .40	50/200 = 0.25	= 1
"A' alleles	140	80	0	220/400 = 0.55
'a' alleles	0	80	100	180/400 = 0.45

Each carries 2 'A' alleles, so collectively they are 'carrying' 140 'A' alleles. The 80 heterozygotes are each carrying 1 'A' allele. And of course, the 'aa' individuals aren't carrying any 'A' alleles. So, in total, there are 220 'A' alleles in the population. With 200 diploid individuals, there are a total of 400 alleles at this locus. So, the gene frequency of the 'A' gene = f(A) = 220/400 = 0.55. We can calculate the frequency of the 'a' alleles the same way. The 50 'aa' individuals are

carrying 2 'a' alleles each, for a total of 100 'a' alleles. The 80 heterozygotes are each carrying an 'a' allele, and the 140 AA homozygotes aren't carrying any 'a' alleles. So, in total, there are 180 'a' alleles out of a total of 400, for a gene frequency f(a) = 180/400 = 0.45. The gene array presents all the gene frequencies, as: f(A) = 0.55, f(a) = 0.45.

There is a faster way to calculate the gene frequencies in a population than adding up the genes contributed by each genotype. Rather, you can use these handy formulae:

f(A) = f(AA) + f(Aa)/2

f(a) = f(aa) + f(Aa)/2

So, to calculate the frequency of a gene in a population, you add the frequency of homozygotes for that allele with 1/2 the frequency of heterozygotes. In our example, this would be:

f(A) = 0.35 + 0.4/2 = 0.35 + 0.2 = 0.55

f(a) = 0.25 + 0.4/2 = 0.25 + 0.2 = 0.4

2.3 Phenotypic variation

Phenotypic variation (due to underlying heritable genetic variation) is a fundamental prerequisite for evolution by natural selection. It is the living organism as a whole that contributes (or not) to the next generation, so natural selection affects the genetic structure of a population indirectly via the contribution of phenotypes. Without phenotypic variation, there would be no evolution by natural selection.

The interaction between genotype and phenotype has often been conceptualized by the following relationship:

genotype (G) + environment (E) ⟶ phenotype (P)

A more nuanced version of the relationship is:

genotype (G) + environment (E) + genotype & environment interactions (GE) ⟶ phenotype (P)

Genotypes often have much flexibility in the modification and expression of phenotypes; in many organisms these phenotypes are very different under varying environmental conditions (see ecophenotypic variation). The plant*Hieracium*

umbellatum is found growing in two different habitats in Sweden. One habitat is rocky, sea-side cliffs, where the plants are bushy with broad leaves and expanded inflorescences; the other is among sand dunes where the plants grow prostrate with narrow leaves and compact inflorescences. These habitats alternate along the coast of Sweden and the habitat that the seeds of *Hieracium umbellatum* land in, determine the phenotype that grows.

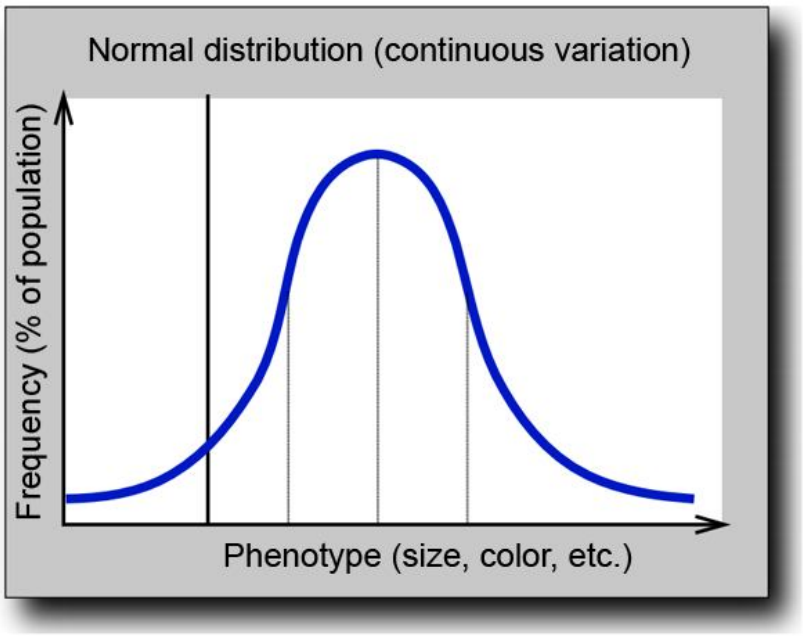
An example of random variation in *Drosophila* flies is the number of ommatidia, which may vary (randomly) between left and right eyes in a single individual as much as they do between different genotypes overall, or between clones raised in different environments.

The concept of phenotype can be extended to variations below the level of the gene that affect an organism's fitness. For example, silent mutations that do not change the corresponding amino acid sequence of a gene may change the frequency of guanine-cytosine base pairs (GC content). These base pairs have a higher thermal stability (*melting point*, see also DNA-DNA hybridization) thanadenine-thymine, a property that might convey, among organisms living in high-temperature environments, a selective advantage on variants enriched in GC content.

2.3.1 Types of Phenotypic variation: Continuous and Discontinuous

When a characteristic or phenotype normally exists in a range or gradient, it varies continuously(like shades of gray, as opposed to black and white). It's easy to think of examples of phenotypes that vary continuously, for example, height and skin color. In between the shortest person in the world and the tallest person in the world, any height is possible, not just 4 feet, 5 feet or 6 feet. And of course, skin comes in all kinds of shades, not just two or three.

If you would make a frequency graph of the range of heights or skin colors in a group of people, it would look like a bell curve, with intermediate phenotypes being the most common. This is a good way to recognize continuous variation.



Phenotypes that vary continuously fit in a bell curve or normal distribution. Intermediate phenotypes are the most common; extreme phenotypes are less common.

Discontinuous

In contrast, some phenotypes vary discontinuously. These phenotypes have 'black and white' differences: for example, you can have blood type A, B, AB or O, but there aren't any intermediate blood types in between. Another example is the ability to roll your tongue. Either you can or you can't, so this phenotype varies discontinuously.

2.3.2 Causes of Phenotypic variation

As mentioned above, phenotypes can be caused by genes, environmental factors, or both. When we say environmental factors, we aren't necessarily talking about the trees and the climate: environmental factors are things in an organism's surroundings or lifestyle that can influence it in various ways. For example, body weight in humans may be influenced by genes, but is also influenced by diet. In this case, diet is an example of an environmental factor. The effects that environmental factors have on phenotypes are hard to pin down, since there are so many possible factors to take into account.

A lot more is known about the relationship between genes and phenotypes. Let's take the example of hair color. Perhaps there is a gene in rabbits that codes for an

enzyme that, in turn, makes a brown-colored pigment in hair follicles. Some rabbits may have genetic differences that cause them to have more or less of this enzyme, or enzyme that works more or less efficiently to produce the pigment. We would expect these rabbits to have different phenotypes, e.g. lighter or darker brown hair, depending on these genetic differences.

Let's take one more example, this time in bacteria. Some bacteria may have a gene that codes for an enzyme that breaks down an antibiotic into a substance that isn't harmful anymore. If you treat these bacteria with the antibiotic, they'll survive: this phenotype is called antibiotic resistance. In contrast, bacteria without that particular gene will be susceptible to the antibiotic.

2.4 Changes in genetic structure of population

Violations of the Hardy-Weinberg equilibrium assumptions are evolutionary processes involved in changing allele frequencies. Multiple violations can occur simultaneously.

1. Mutation

Heritable changes in the DNA that occur within a locus.

Usually converts one allelic form of a gene to another.

A to *a* is called a **forward mutation ;** *a* to *A* is called a **reverse mutation**

The rate of mutation is generally low, but varies among loci and among species.

Certain genes modify overall mutation rates, and many environmental factors, such a chemicals, radiation, and infectous agents, may increase the number of mutations.

Genetic variation arises; then different alleles increase or decrease in frequency in response to evolutionary processes.

ultimately, mutation is the source of all new genetic variation. new combinations of alleles may arise through recombination.

Mutation provides the raw genetic material for evolution. Most mutations will be detrimental and will be eliminated from the population. A few mutations will convey some advantage to the individuals that possess that them and will spread through the population.

Whether the mutation is advantageous or detrimental, depends upon the specific environment, and if the environment changes, previously harmful or neutral mutations may become beneficial.

Mutations can change the frequencies of alleles

2. Genetic Drift

In small populations chance factors may produce large changes in allelic frequencies.random change in allelic frequency due to chance is called genetic drift or simply *drift* for short.

it is also called the sewall wright effect in honor of the population geneticist who championed its importance in the 1930's.

changes in allelic frequency resulting from random events (such as a typhoon) can have important evolutionary implications in small populations and can have important consequences for the conservation of a rare or endangered species.

If the actual number of progeny differs from the expected ratio due to chance, genotypes may not be in h-w proportions and changes in allelic frequencies may occur.

when the sample is small, the sampling error can be large. all genetic drift arises from such samplingerror.

Measuring genetic drift

drift is random so we cannot predict allelic frequencies after it has occurred. however, based on population size we can make predictions about the magnitude of drift.

ecologists often measure population size by counting the number of individuals, but not all individuals contribute gametes to the next generation.

if the sexes are equal and all individuals have an equal probability of producing offspring, the effective population size equals the number of breeding adults in the population.

however when males and females are not present in equal numbers the effective population size is

$n_e = 4 \times n_f \times n_m / n_f + n_m$

males as a group contribute ½ of all genes to the next generation and females as a group contribute the other half.

in a population of 70 females and 2 males, the 2 males are not genetically equivalent to 2 females:

each male contributes ½ x ½ = 0.25 of the genes to the next generation, whereas

each female contributes ½ x 1/70 = 0.007 of all genes.

the small number of males disproportionately influences what alleles are present in the next generation.

$n_e = (4 \times 70 \times 2) / 70 + 2 = 7.8$ or approximately 8 breeding adults.

genetic drift will occur in this population of 70 females and 2 males as if there were 4 breeding males and 4 breeding females. therefore genetic drift will have a much greater effect in this population than in one with 72 breeding adults equally divided between males and females.

other factors such as differential production of offspring, fluctuating population size, and overlapping generations can further reduce the effective population size. consider the complications it is quite difficult to measure the effective population size accurately.

the amount of variation among populations resulting from genetic drift is measured by the variance of allelic frequency:

$s_p^2 = pq/2n_e$

a more useful measure is the standard error of allelic frequency :

s_p = the square root of $pq/2n_e$

Causes of genetic drift:

all genetic drift occurs from sampling error, however there are several ways in which sampling error occurs in natural populations.

1. population size remains continuously small over many generations (due to occupation of marginal habitats, or when competition for resources limits population growth, and fragmentation of habitat due to human intervention such as clear cutting. populations may be spread out over a large geographical area, but is fragmented into many sub populations each showing genetic drift independently).

2. **founder effect** - occurs when a population is initially established by a small number of breeding individuals. although the population may subsequently grow in size and later consist of a large number of individuals, the gene pool of the population is derived from the genes present in the original founders (which may have been determined by chance). this has a profound effect on the gene pool in subsequent generations.
3. **bottleneck effect** - a form of genetic drift that occurs when a population is drastically reduced in size. some genes may be lost from the gene pool as a result of chance. this can be considered a form of founder effect, since the population is refounded by those few individuals that survive the reduction.

Effects of genetic drift:

1. causes the allelic frequencies of a population to change over time fixation of an allele extinction of an allele (the allele must then be reintroduced by mutation or migration) the rare allele is more likely to be lost.
2. reduction in genetic variation within populations
3. individual populations will not change in the same direction (populations diverge in their allelic frequencies). we expect more divergence in allele frequency among small populations than among large populations.

3. Migration

Many populations are not completely isolated, and exchange genes with other populations of the same species. Individuals migrating into a new population may introduce new alleles into the gene pool and alter the frequencies of existing alleles. Thus migration has the potential to disrupt H-W equilibrium and may influence the evolution of allelic frequencies within populations.

Migration usually implies movement of organisms, however in population genetics we are interested in movement of genes, which may or may not occur when organisms move. Movement of genes takes place only when organisms or gametes migrate and contribute their genes to the gene pool of the recipient population. This process is also referred to as gene flow.

1. Gene flow introduces new alleles to the population (because mutation is generally a rare event a mutant may arise in one population and not in another). Gene flow spreads unique alleles to other populations, and like mutation, is a source of genetic variation.

2. When the allelic frequencies of migrants and the recipient population differ, gene flow changes the allelic frequencies within the recipient population. Through exchange of genes, different populations remain similar, and thus migration is a homogenizing force that tends to prevent populations form accumulating genetic differences among them.
3. Migration among populations tends to increase the effective population size of the populations. Migration reduces divergence among populations, effectively increasing the size of the individual populations.
4. Since gene flow has large consequences for the maintenance of genetic diversity, this feature of population genetic structure needs to be taken into account by those interested in conserving genetic structure.

4. Natural Selection

mutation, migration, and genetic drift do not result in adaptation. adaptation is the process by which traits evolve that make organisms more suited to their immediate environment; these traits increase the organisms chance of surviving and reproducing. adaptation is responsible for the extraordinary traits seen in nature.

mutation, migration, and genetic drift all influence the pattern and process of adaptation, but adaptation arises chiefly from natural selection.

natural selection is the dominant force in the evolution of many traits and has shaped much of the phenotypic variation observed in nature.

selection in natural populations.....

1. directional
2. stabilizing selection
3. disruptive selection

fitness and the coefficient of selection....

we measure natural selection by assessing reproduction. it is measured in terms of darwinian fitness which is defined as the relative reproductive ability of a genotype. Violations of the Hardy-Weinberg equilibrium assumptions are evolutionary processes involved in changing allele frequencies. Multiple violations can occur simultaneously

often symbolized as *w,* and is also called the adaptive value of a genotype**.** **selection coefficient**, ***s,*** is a measure of the relative intensity of selection **against** a genotype. $s = 1\text{-}w$

1. no selection
2. against a dominant allele
3. against a recessive allele
4. favoring a heterozygote
5. favoring the homozygote

5. Simultaneous Effects Of Mutation And Selection

mutation can continuously reproduce the allele lost by selection. selection and mutation thus oppose each other!

6. Nonrandom Mating

Many populations do not mate randomly for some traits, and when nonrandom mating occurs, the genotypes will not exist in H-W equilibrium.

Positive assortative mating - when individuals with similar phenotypes mate preferentially.

Negative assortative mating - phenotypically dissimilar individuals mate more often than randomly chosen individuals.

Neither affects the allelic frequencies, but both may affect the genotypic frequencies if the phenotypes are genetically determined.

2.5 Factor affecting human disease frequency

The spread of disease in a population is affected by many factors. One such factor is the presence in the population of carriers. A carrier is someone (or something) that harbors the pathogen and spreads it to other individuals. (Pathogens are disease-causing organisms such as bacteria or fungi, or viruses). A carrier may or may not show symptoms of infection. For example, a person can harbor the human immunodeficiency virus (HIV) for years without knowing it, and without showing signs of acquired immune deficiency syndrome (AIDS), the fatal disease caused by this virus.

Human carriers can unknowingly transmit the pathogen to other humans. Some pathogens, e.g. cold and flu viruses, are transmitted through the air. Air transmission of the pathogen occurs when an infected person sneezes and a nearby person inadvertently breathes in the contaminated air, or when two such people kiss. The transmittal of other pathogens require more direct, prolonged, or intimate contact, such as sharing body fluids during sexual intercourse (which is one of the most common means of transmission of HIV). Sometimes the carrier is not human, but rather another organism such as a mosquito. For example, mosquitoes harbor the pathogens that cause malaria which is transmitted to a human via the mosquito's saliva when the mosquito sucks the human's blood.

When the carriers are non-human, the disease is sometimes easier to control. For example, one of the factors in finally eradicating the bubonic plague which swept through Europe several centuries ago, was the rising popularity of keeping cats as pets; the cats killed the rats that harbored the bacterium. Disease can spread in the absence of carriers. For example, anyone drinking water contaminated with the Giardia protozoan may become ill.

Population patterns (exactly where and how people live and interact with others) and densities (how many people are present within a defined area) are also important in the spread of disease. People living in close proximity, such as in the city, are more likely to spread infection than are those living in the country, where there is less person-to-person contract. One's ability to be reinfected is also a factor in the spread of disease. For example, the plaque bacterium usually can cause disease only once in an individual. After the bout with the disease, the person's immune system is able to eliminate the virus upon subsequent exposures to it, without the person suffering any symptoms.

2.5.1 Factor influencing incidence of disease in populations

Concepts of Incidence and Prevalence

How common is a given disease, such as diabetes, in a population? Well-established measures are used to answer this question.

The incidence rate is the number of new cases of a disease reported during a specific period (typically 1 year) divided by the number of individuals in the population. The denominator is often expressed as person-years. The incidence rate can be contrasted with the prevalence rate, which is the proportion of the

population affected by a disease at a specific point in time. Prevalence is thus determined by both the incidence rate and the length of the survival period in affected individuals.

For example, the prevalence rate of acquired immunodeficiency syndrome (AIDS) is larger than the yearly incidence rate because most people with AIDS survive for at least several years after diagnosis. Many diseases vary in prevalence from one population to another.

Cystic fibrosis is relatively common among Europeans, occurring about once in every 2500 births. In contrast, it is quiterare in Asians, occurring only once in every 90,000 births. Similarly, sickle cell disease affects approximately 1 in 600 American blacks, but it is seen much less frequently in whites. Both of these diseases are single-gene disorders, and they vary among populations because disease-causing mutations are more or less common in different populations. (This is in turn the result of differences in the evolutionary history of these populations.)

Nongenetic (environmental) factors have little influence on the current prevalence of these diseases. The picture often becomes more complex with the common diseases of adulthood. For example, colon cancer was until recently relatively rare in Japan, but it is the second most common cancer in the United States. Stomach cancer, on the other hand, is common in Japan but relatively rare in the United States. These statistics, in themselves, cannot distinguish environmental from genetic influences in the two populations.

However, because large numbers of Japanese emigrated first to Hawaii and then to the U.S. mainland, we can observe what happens to the rates of stomach and colon cancer among the migrants. It is important that the Japanese émigrés maintained a genetic identity, marrying largely among themselves. Among first-generation Japanese in Hawaii, the frequency of colon cancer rose several-fold—not yet as high as in the U.S. mainland but higher than that in Japan. Among second-generation Japanese on the U.S. mainland, colon cancer rates rose to 5%, equal to the U.S. average. At the same time, stomach cancer has become relatively rare among Japanese-Americans.

These observations strongly indicate an important role for environmental factors in the etiology of cancers of the colon and stomach. In each case, diet is a likely culprit—a high-fat, low-fiber diet in the United States is thought to increase the

risk of colon cancer, whereas techniques used to preserve and season the fish commonly eaten in Japan are thought to increase the risk of stomach cancer. It is interesting that the incidence

of colon cancer in Japan has increased dramatically during the past several decades as the Japanese population has adopted a more "Western" diet. These results do not, however, rule out the potential contribution of genetic factors in common cancers. Genes also play a role in the etiology of colon and other

Analysis of Risk Factors

The comparison just discussed is one example of the analysis of risk factors (in this case, diet) and their infl uence on the prevalence of disease in populations. A common measure of the effect of a specific risk factor is the relative risk. This quantity is expressed as a ratio:

Increased rate of the disease among individuals exposed to a risk factor

Incidence rate of the disease among individuals not exposed to a risk factor

A classic example of a relative risk analysis was carried out in a sample of more than 40,000 British physicians to determine the relationship between cigarette smoking and lung cancer. This study compared the incidence of death from lung cancer in physicians who smoked with those who did not. The incidence of death from lung cancer was 1.66 (per 1000 person-years) in heavy smokers (more than 25 cigarettes daily), but it was only 0.07 in the nonsmokers. The ratio of these two incidence rates is 1.66/0.07, which yields a relative risk of 23.7 deaths. Thus, it is concluded that the risk of dying from lung cancer increased by about 24-fold in heavy smokers compared with nonsmokers. Many other studies have obtained similar risk fi gures. Although cigarette smoking clearly increases one's risk of developing lung cancer (as well as heart disease, as we will see later), it is equally clear that most smokers do not develop lung cancer. Other lifestyle factors are likely to contribute to one's risk of developing this disease (e.g., exposure to cancer-causing substances in the air, such as asbestos fi bers).

In addition, differences in genetic background may be involved. Smokers who have variants in genes, such as CYP1A1 and GSTM1, that are involved in the metabolism of components of tobacco smoke are at significantly increased risk of developing lung cancer. Many factors can influence the risk of acquiring a common disease such as cancer, diabetes, or high blood pressure. These include

age, gender, diet, amount of exercise, and family history of the disease. Usually, complex interactions occur among these genetic and nongenetic factors. The effects of each factor can be quantifi ed in terms of relative risks. The following discussion demonstrates how genetic and environmental factors contribute to the risk of developing common diseases.

2.5.2 Principles of multifactorial inheritance

Basic Model

Traits in which variation is thought to be caused by the combined effects of multiple genes are polygenic ("many genes"). When environmental factors are also believed to cause variation in the trait, which is usually the case, the term multifactorial trait is used. Many quantitative traits (those, such as blood pressure, that are measured on a continuous numeric scale) are multifactorial. Because they are caused by the additive effects of many genetic and environmental factors, these traits tend to follow a normal, or bell-shaped, distribution in populations.

An example illustrates this concept. To begin with the simplest case, suppose (unrealistically) that height is determined by a single gene with two alleles, A and a. Allele A tends to make people tall, whereas allele a tends to make them short. If there is no dominance at this locus, then the three possible genotypes (AA, Aa, aa) will produce three phenotypes: tall, intermediate, and short, respectively. Assume that the gene frequencies of A and a are each 0.50. If we look at a population of individuals, we will observe the height distribution depicted in Figure 5-1 , A. Now suppose, a bit more realistically, that height is determined by two loci instead of one. The second locus also has two alleles, B (tall) and b (short), and they affect height in exactly the same way as alleles A and a. There are now nine possible genotypes in our population: aabb, aaBb, aaBB, Aabb, AaBb, AaBB, AAbb, AABb, and AABB. An individual may have zero, one, two, three, or four "tall" alleles, so now fi ve distinct phenotypes are possible .

Although the height distribution in our fi ctional population is still not normal compared with an actual population, it approaches a normal distribution more closely than in the single-gene case just described. From extension of this example, many genes and environmental factors influence height, each having a small effect. Then many phenotypes are possible, each differing slightly from the others, and the height distribution of the population approaches the bell-shaped curve.

It should be emphasized that the individual genes underlying a multifactorial trait such as height follow the mendelian principles of segregation and independent assortment, just like any other gene. The only difference is that many of them act together to infl uence the trait. More than 100 genes have now been shown to be associated with variation in human height. Blood pressure is another example of a multifactorial trait.

A correlation exists between parents' blood pressures (systolic and diastolic) and those of their children. The evidence is good that this correlation is partially caused by genes, but blood pressure is also infl uenced by environmental factors, such as diet, exercise, and stress. Two goals of genetic research are the identifi cation and measurement of the relative roles of genes and environment in the causation of multifactorial diseases.

Threshold Model

A number of diseases do not follow the bell-shaped distribution. Instead, they appear to be either present or absent in individuals, yet they do not follow the inheritance patterns expected of single-gene diseases. A commonly used explanation for such diseases is that there is an underlying liability distribution for the disease in a population. Those individuals who are on the "low" end of the distribution have little chance of developing the disease in question (i.e., they have few of the alleles or environmental factors that would cause the disease). Individuals who are closer to the "high" end of the distribution have more of the disease-causing genes and environmental factors and are more likely to develop the disease. For diseases that are either present or absent, it is thought that a threshold of liability must be crossed before the disease is expressed. Below the threshold, an individual appears normal; above it, he or she is affected by the disease.

A disease that is thought to correspond to this threshold model is pyloric stenosis, a disorder that presents shortly after birth and is caused by a narrowing or obstruction of the pylorus, the area between the stomach and intestine. Chronic vomiting, constipation, weight loss, and imbalance of electrolyte levels result from the condition, but it sometimes resolves spontaneously or can be corrected by surgery.

The prevalence of pyloric stenosis is about 3 per 1000 live births in whites. It is much more common in males than females, affecting 1 of 200 males and 1 of 1000

females. It is thought that this difference in prevalence reflects two thresholds in the liability distribution—a lower one in males and a higher one in females . A lower male threshold implies that fewer disease-causing factors are required to generate the disorder in males.

2.6 Summary

Genetic diversity represents the total genetic variation among individuals within a population. Genetic structure refers to any pattern in the genetic makeup of individuals within a population. Phenotypic variation (due to underlying heritable genetic variation) is a fundamental prerequisite for evolution by natural selection. The spread of disease in a population is affected by many factors. Phenotypic variations can be Continuous or Discontinuous. The cause of phenotypic Variations can be environmental , genetic or both. Changes in the genetic structure of population may be due to Mutation, Genetic Drift, Migration, Natural Selection or Non-random Mating. The factors affecting human disease frequency includes presence of carriers, population patterns and densities .The parameters of prevalence and incidence are used to study a disease in a given population. Environmental factors also affect incidence of a disease.Analysis of risk factors is important in controlling diseases .Principals of multifactoral inheritance can be explained by Basic Model or Threshold Model.

2.8 Glossory

- **Evolution**: a change in the genetic structure of a population
- **Population**: a group of interbreeding organisms that share a common gene pool;
- **Gene Pool**: sum total of alleles held by individuals in a population
- **Genetic structure**: Gene array and Genotypic array
- **Gene/Allele Frequency**: % of alleles at a locus of a particular type
- **Gene Array**: % of all alleles at a locus: must sum to 1.
- **Genotypic Frequency**: % of individuals with a particular genotype

- **Genotypic Array**: % of all genotypes for loci considered; must = 1.
- **Mutation**:Heritable changes in the DNA that occur within a locus.
- **Genetic Drift:** Random changes in the allelic frequency due to chance is called genetic drift.
- **Founder Effect:** Occurs when a population is initially established by a small number of breeding individuals.
- **Bottleneck effect:** A form of genetic drift that occurs when a population is drastically reduced in size.

2.9 Self-Learning Exercise

Section A

1. Explain Nei's diversity index.
2. Define Genetic Drift.
3. What is bottleneck effect?
4. Define Non-random mating.
5. List factors affecting human disease frequency.
6. What is migration?
7. Name the two types of phenotypic variations.

Section B:

1. How is genotypic and gene arrays determined?
2. Explain the causes of phenotypic variations
3. Distinguish between Incidence and Prevalance
4. Explain the basic model of multifactorial inheritance.

Section C:

1. Describe the genetic structures of natural populations.
2. Explain the changes in genetic structure of population.
3. Write an essay on principles of multifactorial inheritance

Unit-3

Molecular population genetics: patterns of change in nucleotide and amino acid sequences, ecological significance of molecular variations, emergence of non-Darwinism-neutral hypothesis

Structure of the Unit

3.0 Objectives

After going through this unit you will be able to understand:

- What is population genetics?

- What pattern have been followed by nucleotide and amino acid sequence during evolution?
- Which type of mutation reatined and which are eliminated during the course of evolution?
- What are the importance of variation at molecular level on the ecology?
- A brief description of kimura's neutral hypothesis?

3.1 Introduction

The study of evolutionary biology is commonly divided into two components:

study of the processes by which evolutionary change occurs and study of the patterns produced by those processes.

By ``pattern'' we mean primarily the pattern of phylogenetic relationships among species or genes. Studies of evolutionary processes often don't devote too much attention to evolutionary patterns, except insofar as it is often necessary to take account of evolutionary history in determining whether or not a particular feature is an adaptation. Similarly, studies of evolutionary pattern sometimes try not to use any knowledge of evolutionary processes to improve their guesses about phylogenetic relationships, because the relationship between process and pattern can be tenuous or at least that's the way it was in evolutionary biology when evolutionary biologists were concerned primarily with the evolution of morphological, behavioral, and physiological traits and when systematists used primarily anatomical, morphological, and chemical features (but not proteins or DNA) to describe evolutionary patterns. With the advent of molecular biology after the Second World War and its application to an increasing diversity of organisms in the late 1950s and early 1960s, that began to change. Goodman used the degree of immunological cross-reactivity between serum proteins as an indication of the evolutionary distance among primates. Zuckerkandl and Pauling proposed that after species diverged, their proteins diverged according to a ``molecular clock,'' suggesting one that molecular similarities could be used to reconstruct evolutionary history. In 1966, Harris and Lewontin and Hubby showed that human populations and populations of Drosophila pseudoobscura respectively, contained surprising amounts of genetic diversity.

In this course, we'll focus on advances made in understanding the processes of molecular evolution and pay relatively little attention to the ways in which inferences about evolutionary patterns can be made from molecular data.

Types of data used to asses molecular population genetics

Before we delve any further into our study of molecular evolution, it's probably useful to back up a bit and talk a bit about the types of data that are available to molecular evolutionists. Even though studies of molecular evolution in the last 10-15 years have focused on data derived from DNA sequence or copy number variation, modern applications of molecular markers evolved from earlier applications. Those markers had their limitations, but analyses of them also laid the groundwork for most or all of what's going on in analyses of molecular evolution today. Thus, it's useful to remind everyone what those groups are and to agree on some terminology for the ones we'll say something about. Let's talk first about the physical basis of the underlying data. Then we'll talk about the laboratory methods used to reveal variation.

3.1.1 The physical basis of molecular evolution

With the exception of RNA viruses, the hereditary information in all organisms is carried in DNA. Ultimately, differences in any of the molecular markers we study (and of genetically-based morphological, behavioral, or physiological traits) is associated with some difference in the physical structure of DNA, and molecular evolutionists study a variety of its aspects.

(i) Nucleotide sequence

A difference in nucleotide sequence is the most obvious way in which two homologous stretches of DNA may differ. The differences may be in translated portions of protein genes (exons), portions of protein genes that are transcribed but not translated (e.g., introns, 5' or 3' untranslated regions), non-transcribed functional regions (e.g., promoters), or regions without apparent function.

(ii) Protein sequence

Because of redundancy in the genetic code, a difference in nucleotide sequence at a protein-coding locus may or may not result in proteins with a different amino acid sequence. **Important note**: Don't forget that some loci

code for RNA that has an immediate function without being translated to a protein, e.g., ribosomal RNA and various small nuclear RNAs.

(iii) Secondary, tertiary, and quaternary structure

Differences in amino acid sequence may or may not lead to a different distribution of α-helices and β-sheets, to a different three-dimensional structure, or to different multisubunit combinations.

(iv) Imprinting

At certain loci in some organisms the expression pattern of a particular allele depends on whether that allele was inherited from the individual's father or its mother.

(v) Expression

Functional differences among individuals may arise because of differences in the patterns of gene expression, even if there are no differences in the primary sequences of the genes that are expressed.

(vi) Sequence organization

Particular genes may differ between organisms because of differences in the position and number of introns. At the whole genome level, there may be differences in the amount and kind of repetitive sequences, in the amount and type of sequences derived from transposable elements, in the relative proportion of G-C relative to A-T, or even in the identity and arrangement of genes that are present.

(vii) Copy number variation

Even within diploid genomes, there may be substantial differences in the number of copies of particular genes. In humans, for example, 76 copy-number polymorphisms (CNPs) were identified in a sample of only 20 individuals, and individuals differed from one another by an average of 11 CNPs.

It is worth remembering that in nearly all eukaryotes there are two different genomes whose characteristics may be analyzed: the nuclear genome and the mitochondrial genome. In plants there is a third: the chloroplast genome. In some protists, there may be even more, because of secondary or tertiary endosymbiosis. The mitochondrial and chloroplast genomes are typically inherited only through the maternal line, although some instances of biparental inheritance are known.

3.1.2 Revealing the molecular evolution

The diversity of laboratory techniques used to reveal molecular variation is even greater than the diversity of underlying physical structures. Various techniques involving direct measurement of aspects of DNA sequence variation are by far the most common today, so we'll mention only the techniques that have been most widely used.

(i) Immunological distance

Some molecules, notably protein molecules, induce an immune response in common laboratory mammals. The extent of cross-reactivity between an antigen raised to humans and chimps, for example, can be used as a measure of evolutionary distance. The immunological distance between humans and chimps is smaller than it is between humans and orangutans, suggesting that humans and chimps share a more recent common ancestor.

(ii) DNA-DNA hybridization

Once repetitive sequences of DNA have been ``subtracted out",[4] the rate and temperature at which DNA species from two different species anneal reflects the average percent sequence divergence between them. The percent sequence divergence can be used as a measure of evolutionary distance. Immunological distances and DNA-DNA hybridization were once widely used to identify phylogenetic relationships among species. Neither is now widely used in molecular evolution studies.

(iii) Isozymes

Biochemists recognized in the late 1950s that many soluble enzymes occurred in multiple forms within a single individual. Population geneticists, notably Hubby and Lewontin, later recognized that in many cases, these different forms corresponded to different alleles at a single locus, allozymes. Allozymes are relatively easy to score in most macroscopic organisms, they are typically co-dominant (the allelic composition of heterozygotes can be inferred), and they allow investigators to identify both variable and non-variable loci.Patterns of variation at allozyme loci may not be representative of genetic variation that does not result from differences in protein structure or that are related to variation in proteins that are insoluble.

(iv) RFLPs

In the 1970s molecular geneticists discovered restriction enzymes, enzymes that cleave DNA at specific 4, 5, or 6 base pair sequences, the recognition site. A single nucleotide change in a recognition site is usually enough to eliminate it. Thus, presence or absence of a restriction site at a particular position in a genome provides compelling evidence of an underlying difference in nucleotide sequence at that positon.

(v) RAPDs, AFLPs, ISSRs

With the advent of the polymerase chain reaction in the late 1980s, several related techniques for the rapid assessment of genetic variation in organisms for which little or no prior genetic information was available. These methods differ in details of how the laboratory procedures are performed, but they are similar in that they (a) use PCR to amplify anonymous stretches of DNA, (b) generally produce larger amounts of variation than allozyme analyses of the same taxa, and (c) are bi-allelic, dominant markers. They have the advantage, relative to allozymes, that they sample more or less randomly through the genome. They have the disadvantage that heterozygotes cannot be distinguished from dominant homozygotes, meaning that it is difficult to use them to obtain information about levels of within population inbreeding.

(vi) Microsatellites

Satellite DNA, highly repetitive DNA associated with heterochromatin, had been known since biochemists first began to characterize the large-scale structure of genomes in DNA-DNA hybridization studies. In the mid-late 1980s several investigators identified smaller repetitive units dispersed throughout many genomes. Microsatellites, which consist of short (2-6) nucleotide sequences repeated many times, have proven particularly useful for analyses of variation within populations since the mid-1990s. Because of high mutation rates at each locus, they commonly have many alleles. Moreover, they are typically co-dominant, making them more generally useful than dominant markers. Identifying variable microsatellite loci is more laborious than identifying AFLPs, RAPDs, or ISSRs.

(vii) Nucleotide sequence

The advent of automated sequencing has greatly increased the amount of population-level data available on nucleotide sequences. Nucleotide sequence data has an important advantage over most of the types of data

discussed so far: allozymes, RFLPs, AFLPs, RAPDs, and ISSRs may all hide variation. Nucleotide sequence differences need not be reflected in any of those markers. On the other hand, each of those markers provides information on variation at several or many, independently inherited loci. Nucleotide sequence information reveals differences at a location that rarely extends more than 2-3kb. Of course, as next generation sequencing techniques become less expensive and more widely available, we will see more and more examples of nucleotide sequence variation from many loci within individuals.

(viii) Single nucleotide polymorphisms

In organisms that are genetically well-characterized it may be possible to identify a large number of single nucleotide positions that harbor polymorphisms. These SNPs potentially provide high-resolution insight into patterns of variation within the genome. For example, the HapMap project has identified approximately 3.2M SNPs in the human genome, or about one every kb .

As you can see from these brief descriptions, each of the markers reveals different aspects of underlying hereditary differences among individuals, populations, or species. There is no single ``best'' marker for evolutionary analyses. In many cases in molecular evolution, the interest is intrinsically in the evolution of the molecule itself, so the choice is based not on what those molecules reveal about the organism that contains them but on what questions about which molecules are the most interesting.

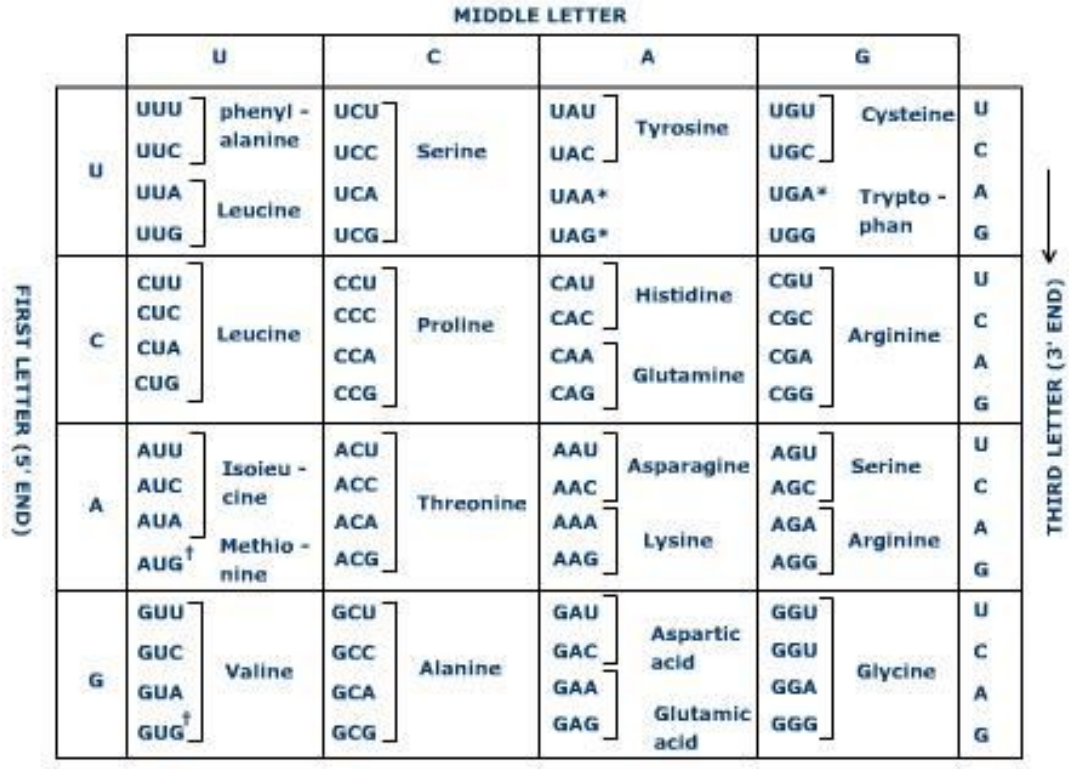
3.2 Patterns of nucleotide and amino acid substitution

Neutral theory of molecular evolution explains quite a bit about this pattern, but it also ignores quite a bit.

1. The derivations we did assumed that all substitutions are equally likely to occur, because they are selectively neutral. That isn't plausible. We need look no further than sickle cell anemia to see an example of a protein polymorphism in which a single amino acid difference has a very large effect on fitness. Even reasoning from first principles we can see that it doesn't make much sense to think that all nucleotide substitutions are created equal. Just as it's unlikely that you'll improve the performance of

your car if you pick up a sledgehammer, open its hood, close your eyes, and hit something inside, so it's unlikely that picking a random amino acid in a protein and substituting it with a different one will improve the function of the protein.

2. The genetic code of course, not all nucleotide sequence substitutions lead to amino acid substitutions in protein-coding genes. There is redundancy in the genetic code. Table 1 is a list of the codons in the universal genetic code.
3. Notice that there are only two amino acids, methionine and tryptophan, that have a single codon. All the rest have at least two. Serine, arginine, and leucine have six.



Moreover, most of the redundancy is in the third position, where we can distinguish 2-fold from 4-fold redundant sites (Table 2).

Codon	Amino Acid	Redundancy
CCU CCC CCA	Pro	4-fold

CCG		
AAU AAC	Asn	2-fold
AAA AAG	Lys	2-fold

Table 2: Examples of 4-fold and 2-fold redundancy in the 3rd position of the universal genetic Code of two nucleotides can be present in a codon for a single amino acid.

4-fold redundant sites are those at which any of the four nucleotides can be present in a codon for a single amino acid. In some cases there is redundancy in the first codon position, e.g, both AGA4 and CGA are codons for arginine. Thus, many nucleotide substitutions at third positions do not lead to amino acid substitutions, and some nucleotide substitutions at first positions do not lead to amino acid substitutions. But every nucleotide substitution at a second codon position leads to an amino acid substitution. Nucleotide substitutions that do not lead to amino acid substitutions are referred to as **synonymous substitutions**, because the codons involved are synonymous, i.e., code for the same amino acid. Nucleotide substitutions that do lead toamino acid substituions are **non-synonymous substitutions**.

3.2.1 Rates of synonymous and non-synonymous substitution

By using a modification of the simple Jukes-Cantor model, it is possible make separate estimates of the number of synonymous substitutions and of the number of non-synonymous substitutions that have occurred since two sequences diverged from a common ancestor.

If we combine an estimate of the number of differences with an estimate of the time of divergence we can estimate the rates of synonymous and nonsynonymous substitution (number/time). Table 3 shows some representative estimates for the rates of synonymous and non-synonymous substitution in different genes studied in mammals.

Two very important observations emerge after you've looked at this table for awhile.

The Locus	Non-synonymous rate	Synonymous rate
Histone		
H4	0.00	3.94
H2	0.00	4.52
Ribosomal proteins		
S17	0.06	2.69
S14	0.02	2.16
Hemoglobins & myoglobin	0.56	4.38
α-globin	0.78	2.58
β-globin	0.57	4.10
Myoglobin		
Interferons		
γ	3.06	5.50
α	1 1.47	3.24
β_1	2.38	5.33

Table 3: Representative rates of synonymous and non-synonymous substitution in mammalian genes . Rates are expressed as the number of substitutions per 109 years.

first - The rate of non-synonymous substitution is generally lower than the rate of synonymous substitution. This is a result of "sledgehammer principle."

Mutations that change the amino acid sequence of a protein are more likely to reduce that protein's functionality than to increase it. As a result, they are likely to lower the fitness of individuals carrying them, and they will have a lower probability of being fixed than those mutations that do not change the amino acid sequence.

The second observation is more subtle. Rates of non-synonymous substitution vary by more than two orders of magnitude: 0.02 substitutions per nucleotide per billion years in ribosomal protein S14 to 3.06 substitutions per nucleotide per billion years in γ-interferon, while rates of synonymous substitution vary only by a factor of two (2.16 in ribosomal protein S14 to 4.52 in histone H2). If synonymous substitutions are neutral, as they probably are to a first approximation,4 then the rate of synonymous substitution should equal the mutation rate. Thus, the rate of synonymous substitution should be approximately the same at every locus, which is roughly what we observe. But proteins differ in the degree to which their physiological function affects the performance and fitness of the organisms that carry them.

Some, like histones and ribosomal proteins, are intimately involved with chromatin or translation of messenger RNA into protein. It's easy to imagine that just about any change in the amino acid sequence of such proteins will have a detrimental effect on its function. Others, like interferons, are involved in responses to viral or bacterial pathogens.

It's easy to imagine not only that the selection on these proteins might be less intense, but that some amino acid substitutions might actually be favored by natural selection because they enhance resistance to certain strains of pathogens. Thus, the probability that a nonsynonymous substitution will be fixed is likely to vary substantially among genes, just as we observe.

3.2.2 Correlations with neutral theory

So we've now produced empirical evidence that many mutations are not neutral. Does this

- Most non-synonymous substitutions are deleterious. We can actually generalize thisassertion a bit and say that most mutations that affect function are deleterious.
- Most molecular variability found in natural populations is selectively neutral. If most function-altering mutations are deleterious, it follows that we are unlikely to find much variation in populations for such mutations. Selection will quickly eliminate them.
- Natural selection is primarily purifying. Although natural selection for variants that improve function is ultimately the source of adaptation, even at

the molecular level, most of the time selection is simply eliminating variants that are less fit than the norm, not promoting the fixation of new variants that increase fitness.

- Alleles enhancing fitness are rapidly incorporated. They do not remain polymorphic for long, so we aren't likely to find them when they're polymorphic.

As we'll see, even these revisions aren't entirely sufficient, but what we do from here on out is more to provide refinements and clarifications than to undertake wholesale revisions.

3.3 Ecological significance of molecular variations

Understanding the relationships between species richness, species diversity and community stability remains a key area of interest in population ecology. At the theoretical level a great deal of attention has been given to how the processes of competition and reproduction determine interactions amongst plant species in communities.

Maintaining stability in natural or managed vegetation is important for practical reasons relating to habitat and species preservation and rural sustainability. There is thus a motivation to bring theoretical results into practical use and, as a result, a considerable amount of research has been conducted in both the theoretical and applied aspects of vegetation ecology.

Recent developments in molecular methods for assessing diversity, particularly the analysis of randomly amplified polymorphic DNA fragments, have made it possible to obtain estimates of intra- and inter-specific diversity from nuclear genomes of natural plant populations with relative ease. However, methods for applying results of such analyses to improve our understanding the ecology of natural vegetation are still relatively poorly developed.

This is due, in part, to the difficulties which arise in relating molecular marker data to important ecophysiological traits in natural plant populations.

Contemporary Molecular Ecology has come to focus much more on the outer shell that Weiss envisaged; the interactions of the organism with 'the environment' in general and, of course, such interactions are the stuff of ecology itself.

Exciting recent developments in Molecular Ecology now provide scientists with a wide array of DNA tools by which to map and explore these interactions. Although collectively these techniques can assist in the resolution of a number of contemporary ecological problems, each of them has particular strengths and weaknesses and is applicable usually to a subset of problems.

Molecular Ecology has come to represent the use of DNA nucleotide sequence variation, nuclear genotypes and organelle haplotypes to gather information about natural populations. With the expansion of DNA tools there has been a dramatic increase in the application of ecological problems.

3.3.1 Genes as markers

Molecular Ecology uses genes essentially as markers to estimate ecologically important variables. In fact, there are two broad approaches that biologists take to the study of genetics. Typically, their focus is either to investigate genes in relation to the role they play in the development of organismic form (sensu, behaviour, morphology etc.). From this perspective, genes are typically conceived of as representing some sort of 'Central Directing Agency' in relation to ontogeny.

In contrast, Mendel himself used phenotypes of organisms (e.g., the shape of pea seeds) as markers to investigate the action of genes themselves. By examining the phenotypic patterns in different generations, Mendel was able to infer the action of genes at meiosis. Over the subsequent history of population genetics, biologists have used genes themselves as markers to investigate population level phenomena.

3.3.2 Genetic markers

Throughout the history of population biology these genetic markers have become progressively more precise and specific. We have moved from chromosome markers to isozymes and latterly, with the proliferation of DNA methods, to the direct examination of single locus, multilocus and nucleotide sequence variation.

An important use of genetic markers has been the detection of cryptic species - those that were indistinguishable on morphological grounds. There are many examples of the use of genetic markers in this way.

In recent years the use of more genetically variable markers such as multilocus minisatellite DNA approaches has enabled the determination of parentage in natural populations, a problem typically unresolvable using chromosome or even isozyme markers.

DNA sequence variation now allows the investigation of the evolution of such ecologically important molecules as the antifreeze proteins of Antarctic fish (Bargelloni et al., 1994).

Molecular ecology is an essential tool in ensuring the proper assessment of the risks of the release of genetically modified organisms (GMOs) (Williamson, 1992).

Our knowledge of the molecular processes that underlie evolutionary change has been, until recently, based on comparisons of the genes of living species. Unlike the remains of animals and plants themselves, DNA does not leave impressions in the rocks. However, ancient DNA, although degraded, can survive the ravages of time. To date, DNA from a number of extinct animals and plants has been amplified using PCR and the DNA sequence successfully determined. The oldest and most important instance is the woolly mammoth, a frozen carcass found in the permafrost of Siberia. This species is thought to have lived 40 000 years ago and the amplification and the subsequent analysis of its DNA represents a major advance for ancient DNA studies.

New studies in Molecular Ecology will enable the direct examination of changes in gene frequencies, not only across space, but over considerable periods of geological time - an achievement never before possible!

3.4 Neutral theory

The Neutral Theory of Molecular Evolution (NTME) propose that the overall pattern of DNA sequence volution can be explained by the combined forces of mutation, genetic drift and purifying selection. While the NTME does not discount adaptive evolution at some loci, it suggests that the evolutionary fate of the majority of loci are not under the influence of positive selection. Therefore, while the few mutations that are influenced by positive selection may have a great biological significance hey are too few to have any statistically significant effect on the overall patterns of DNA sequence evolution under the NTME.

Methods that utilize comparative genomics data, i.e. data from a few, closely related species, in combination with population-level data on polymorphism have proved to be particularly powerful in addressing the question of the relative proportion of mutations that are deleterious and beneficial. The reason for the success of these methods is that natural selection is expected to have stronger effects of rates of sequence divergence between the species that on levels of

polymorphism maintained within species. Methods that combine information from intraspecific levels of polymorphism with data on divergence between losely related species are therefore particularly suitable to infer the rates at which positive and negative selection are acting across the genome of an organism. In this project the goal is to quantify the relative importance of genetic drift and natural selection in shaping genome wide patterns of variation in Populus.

Neutral theory assumes that selection plays a minor role in determining the maintenance of molecular variants and proposes that different molecular genotypes have almost identical relative fitnesses; that is, they are neutral with respect to each other. The actual definition of selective neutrality depends on whether changes in allele frequency are primarily determined by genetic drift.

In a simple example, if s is the selective difference between two alleles at a locus, and if

s < 1/(2Ne), the alleles are said to be neutral with respect to each other because the impact of genetic drift is larger than selection. This definition implies that alleles may be effectively neutral in a small population but not in a large population. Neutral theory does not claim that the relatively few allele substitutions responsible for evolutionarily adaptive traits are neutral, but it does suggest that the majority of allele substitutions have no selective advantage over those that they replace.

Kimura also showed that the neutral theory was consistent with a molecular clock; that is, there is a constant rate of substitution over time for molecular variants. To illustrate the mathematical basis of the molecular clock, let us assume that mutation and genetic drift are the determinants of changes in frequencies of molecular variants. Let the mutation rate to a new allele be u so that in a population of size 2N there are 2Nu new mutants per generation. It can be shown that the probability of chance fixation of a new neutral mutant is 1/(2N) (the initial frequency of the new mutant).

Therefore, the rate of allele substitution k is the product of the number of new mutants

$$K = 2Nu\left(\frac{1}{2N}\right)$$ u

In other words, this elegant prediction from the neutral theory is that the rate of substitution is equal to the mutation rate at the locus and is constant over time.

Note that substitution rate is independent of the effective population size, a fact that may initially be counterintuitive. Thisindependence occurs because in a smaller population there are fewer mutants; that is, 2Nu is smaller, but the initial frequency of these mutants is higher, which increases the probability of fixation, 1/2N, by the same magnitude by which the number of mutants is reduced.

This simple, elegant mathematical prediction and others from the neutral theory provide the basis for the most important developments in evolutionary biology in the past halfcentury.One of the appealing aspects of the neutral theory is that, if it is used as a **null hypothesis**, then predictions about the magnitude and pattern of genetic variation are possible. Initially, molecular genetic variation was found to be consistent with that predicted from neutrality theory.

In recent years, examination of neutral theory predictions in DNA sequences has allowed tests of the cumulative effect of many generations of selection, and a number of examples of selection on molecular variants have been documented .

Factors influencing the evolutionary changes

Traditionally, population genetics examines the impact of various evolutionary factors on the amount and pattern of genetic variation in a population and how these factors influence the future potential for evolutionary change. Generally, evolution is conceived of as a forward process, examining and predicting the future characteristics of a population.

However, rapid accumulation of DNA sequence data over the past two decades has changed the orientation of much of population genetics from a prospective one investigating the factors involved in observed evolutionary change to a retrospective one inferring evolutionary events that have occurred in the past. That is, understanding the evolutionary causes that have influenced the DNA sequence

Variation in a sample of individuals, such as the demographic and mutational history of the ancestors of the sample, has become the focus of much population genetics research. In a determination of DNA variation in a population,a sample of alleles is examined. Each of these alleles may have a different history, ranging from descending from the same ancestral allele, that is, identical by descent, in the previous generation to descending from the same ancestral allele many generations before. The point at which this common ancestry for two alleles occurs is called coalescence.

If one goes back far enough in time in the population, then all alleles in the sample will coalesce into a single common ancestral allele. Research using the coalescent approach is the most dynamic area of theoretical population genetics because it is widely used to analyze DNA sequence data in populations and species.

3.5 Summary

With the exception of RNA viruses, the hereditary information in all organisms is carried in DNA. A difference in nucleotide sequence is the most obvious way in which two homologous stretches of DNA may differ. The diversity of laboratory techniques used to reveal molecular variation is even greater than the diversity of underlying physical structures viz. Immunological distance,DNA-DNA hybridization,Isozymes, RFLPs ,RAPDs, AFLPs, ISSRs, Microsatellites, Nucleotide sequence, Single nucleotide polymorphisms etc.Neutral theory of molecular evolution explains quite a bit about patterns of nucleotide and amino acid substitution. Molecular ecology has come to represent the use of DNA nucleotide sequence variation , nuclear genotypes and organelle haplotypes to gather information about natural populations .The Neutral Theory of Molecular propose that the overall pattern of DNA sequence volition can be explained by the combined forces of mutation , genetic drift and purifying selection.

3.6 Self-Learning Exercise

Section A

1. How nucleotide sequence can be helpful in studying molecular evolution?
2. Name 2 techniques used in studying molecular evolution.
3. What are microsatellites?
4. What are synonymous and non-synonymus substitution?

Section B

1. Explain the types of data used in the assessment of molecular population genetics.
2. How patterns of nucleotide and amino acid substitution useful in molecular evolution study?
3. Explain correlations with neutral theory

Section C

1. Write detailed note on “Ecological Significance of molecular variations”
2. Explain in detail the Neutral Theory.

Unit - 4

Genetics of quantitative traits in populations, genotype-environment interactions, inbreeding depression and heterosis, molecular analysis of quantitative traits, phenotypic plasticity

Structure of the Unit

4.0 Objectives

After going through this unit you will be able to understand:

- What are quantitative traits and how these are different from mendelian qualitative traits.
- What is gene environment interaction and their influence on to the phenotype
- How selfbreeding reduces the fitness of progeny by the phenomenon of inbreeding depression
- Mechanism of heterosis to overcome the inbreeding depression.
- What are quantitative trait loci(QTL's) and process of their mapping.
- How the environment changes the phenotype without change in genetic mekeup.

4.1 Introduction

All of the traits that we have studied to date fall into a few distinct classes. These classes can be used to predict the genotypes of the individuals. For example, if we cross a tall and short pea plant and look at F_2 plants, we know the genotype of s hort plants, and we can give a generalized genotype for the tall plant phenotype. Furthermore, if we know the genotype we could predict the phenotype of the plant. These type of phenotypes are called discontinuous traits.

Other traits do not fall into discrete classes. Rather, when a segregating population is analyzed, a continuous distribution of phenotypes is found. An example, is ear length in corn. Black Mexican Sweet corn has short ears, whereas Tom Thumb popcorn has long ears. When these two inbred lines are crossed, the length of the F_1 ears are intermediate to the two parents. Furthermore, when the F_1plants are intermated, the distribution of ear length in the F_2 ranges from the short ear Black Mexican Sweet size to the Tom Thumb popcorn size. The distribution resembles the bell-shaped curve for a normal distribution.

These types of traits are called continuous traits and cannot be analyzed in the same manner as discontinuous traits. Because continuous traits are often measured and given a quantitative value, they are often referred to as quantitative traits, and

the area of genetics that studies their mode of inheritance is called quantitative genetics.

Many important agricultural traits such as crop yield, weight gain in animals, fat content of meat are quantitative traits, and much of the pioneering research into the modes of inheritance of these traits was performed by agricultural geneticists. Many human phenotypes such as IQ, learning ability and blood pressure also are quantitative traits.These traits are controlled by multiple genes, each segregating according to Mendel's laws. These traits can also be affected by the environment to varying degrees.

The following are examples of quantitative traits that we are concerned with in our daily life.

- Crop Yield
- Some Plant Disease Resistances
- Weight Gain in Animals
- Fat Content of Meat
- IQ
- Learning Ability
- Blood Pressure

Quantitative genetics is that branch of population genetics which deals with phenotypes which vary continuously (such as height or mass), rather than with phenotypes and gene-products which are discretely identifiable (such as eye-colour, or the presence of a particular biochemical). Both employ the frequencies of different alleles of a gene in breeding populations (gamodemes), and combine them with concepts arising from simple Mendelian inheritance in order to analyze inheritance patterns across generations and descendant lines. While population genetics can focus on particular genes and their subsequent metabolic products, quantitative genetics focuses more on the outward phenotypes, and makes summaries only of the underlying genetics.

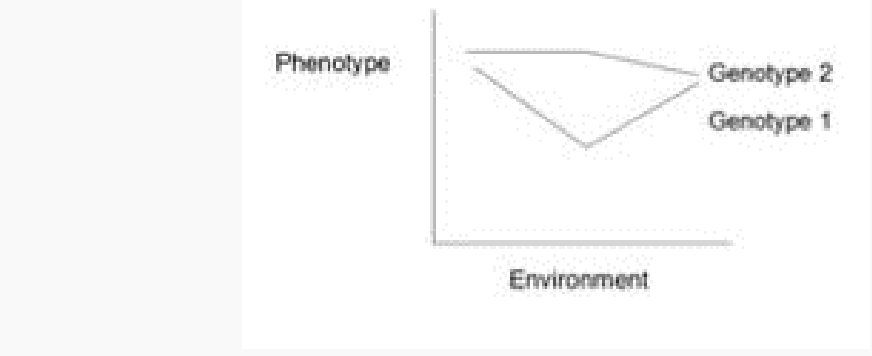
This, however, can be viewed as its strength, because it facilitates an interface with the biological macrocosm, including micro-evolution and artificial selection in plant and animal breeding. Both branches share some common history; and some mathematics: for example, they use expansion of the quadratic equation to

represent the fertilization of gametes to form the zygote. However, because of the continuous distribution of phenotypic values, quantitative genetics needs also to employ many other statistical methods (such as the effect, the mean and the variance) in order to link the phenotype to underlying genetics principles. Some phenotypes (attributes) may be analyzed either as discrete categories or as continuous phenotypes, depending on the definition of cut-off points, or on the metric used to quantify them

Mendel himself had to discuss this matter in his famous paper, especially with respect to his peas attribute tall/dwarf, which actually was "length of stem" Analysis of quantitative trait loci, or QTL, is a more recent addition to quantitative genetics, linking it more directly to molecular genetics.

4.2 Genotype environment interaction

Gene–environment interaction (or genotype–environment interaction or G×E) is when two different genotypes respond to environmental variation in different ways. A norm of reaction is a graph that shows the relationship between genes and environmental factors when phenotypic differences are continuous. They can help illustrate GxE interactions. When the norm of reaction is not parallel, as shown in the figure below, there is a gene by environment interaction. This indicates that each genotype responds to environmental variation in a different way.



This norm of reaction shows lines that are not parallel indicating a gene by environment interaction. Each genotype is responding to environmental variation in a different way.

Gene–environment interactions are studied to gain a better understanding of various phenomena. In genetic epidemiology, gene-environment interactions are useful for understanding some diseases. Sometimes, sensitivity to

environmental risk factors for a disease are inherited rather than the disease itself being inherited. Individuals with different genotypes are affected differently by exposure to the same environmental factors, and thus gene-environment interactions can result in different disease phenotypes. For example, sunlight exposure has a stronger influence on skin cancer risk in fair-skinned humans than in individuals with darker skin.

Nature versus nurture debates assume that variation in a trait is primarily due to either genetic differences or environmental differences. However, the current scientific opinion holds that neither genetic differences nor environmental differences are solely responsible for producing phenotypic variation, and that virtually all traits are influenced by both genetic and environmental differences. Statistical analysis of the genetic and environmental differences contributing to the phenotype would have to be used to confirm these as gene-environment interactions.

Examples

1. In Drosophila: A classic example of gene–environment interaction was performed on drosophila by Gupta and Lewontin in 1981. In their experiment they demonstrated that the mean bristle number on drosophila could vary with changing temperatures. As seen in the graph to the right, different genotypes reacted differently to the changing environment. Each line represents a given genotype, and the slope of the line reflects the changing phenotype (bristle number) with changing temperature. Some individuals had an increase in bristle number with increasing temperature while others had a sharp decrease in bristle number with increasing temperature. This showed that the norms of reaction were not parallel for these flies, proving that gene-environment interactions exist.
2. In plants: Seven genetically distinct yarrow plants were collected and three cuttings taken from each plant. One cutting of each genotype was planted at low, medium, and high elevations, respectively. When the plants matured, no one genotype grew best at all altitudes, and at each altitude the seven genotypes fared differently. For example, one genotype grew the tallest at the medium elevation but attained only middling height at the other two elevations. The best growers at low and high elevation grew poorly at medium elevation. The medium altitude produced the worst

overall results, but still yielded one tall and two medium-tall samples. Altitude had an effect on each genotype, but not to the same degree nor in the same way.

3. Phenylketonuria (PKU) is a human genetic condition caused by mutations to a gene coding for a particular liver enzyme. In the absence of this enzyme, anamino acid known as phenylalanine does not get converted into the next amino acid in a biochemical pathway, and therefore too much phenylalanine passes into the blood and other tissues. This disturbs brain development leading to mental retardation and other problems. PKU affects approximately 1 out of every 15,000 infants in the U.S. However, most affected infants do not grow up impaired because of a standard screening program used in the U.S. and other industrialized societies. Newborns found to have high levels of phenylalanine in their blood can be put on a special, phenylalanine-free diet. If they are put on this diet right away and stay on it, these children avoid the severe effects of PKU.[11] This example shows that a change in environment (lowering PKU consumption) can affect the phenotype of a particular trait, demonstrating a gene-environment interaction.
4. A functional polymorphism in the monoamine oxidase A (MAOA) gene promoter can moderate the association between early life trauma and increased risk for violence and antisocial behavior. Low MAOA activity is a significant risk factor for aggressive and antisocial behavior in adults who report victimization as children. Persons who were abused as children but have a genotype conferring high levels of MAOA expression are less likely to develop symptoms of antisocial behavior. These findings must be interpreted with caution, however, because gene association studies on complex traits are notorious for being very difficult to confirm.
5. In Drosophila Eggs:

   Contrary to the aforementioned examples, length of egg development in drosophila as a function of temperature demonstrates the lack of gene-environment interactions. The attached graph shows parallel reaction norms for a variety of individual drosophila flies, showing that there is not a gene-environment interaction present between the two variables. In other words, each genotype responds similarly to the changing environment

producing similar phenotypes. For all individual genetypes, average egg development time decreases with increasing temperatuer. The environment is influencing each of the genotypes in the same predictable manner.

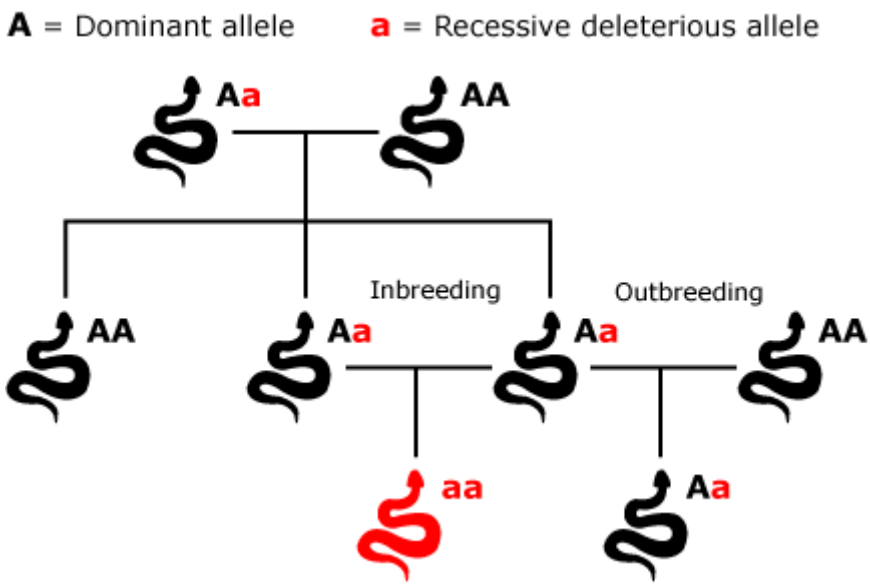
4.3 Inbreeding depression

Inbreeding depression is the reduced biological fitness in a given populationas a result of inbreeding - ie., breeding of related individuals. Population biological fitness refers to its ability to survive and reproduce itself. Inbreeding depression is often the result of a population bottleneck. In general, the higher the genetic variation or gene pool within a breeding population, the less likely it is to suffer from inbreeding depression.

Inbreeding depression seems to be present in most groups of organisms, but varies across mating systems. Hermaphroditic species often exhibit lower degrees of inbreeding depression than outcrossing species, as repeated generations of selfing is thought to purge deleterious alleles from populations.

4.3.1 Mechanism

Inbreeding (ie., breeding between closely related individuals) may on the one hand result in more recessive deleterious traits manifesting themselves, because the genomes of pair-mates are more similar: recessive traits can only occur in offspring if present in both parents' genomes, and the more genetically similar the parents are, the more often recessive traits appear in their offspring. Consequently, the more closely related the breeding pair is, the more homozygous deleterious genes the offspring may have, resulting in very unfit individuals. For alleles that confer an advantage in the heterozygous and/or homozygous-dominant state, the fitness of the homozygous-recessive state may even be zero (meaning sterile or unviable offspring).



The explanation for inbreeding depression lies in the evolutionary history of the population. Over time, natural selection weeds deletarious allele out of a population—when the dominant deleterious alleles are expressed, they lower the carrier's fitness, and fewer copies wind up in the next generation. But recessive deleterious alleles are "hidden" from natural selection by their dominant non-deleterious counterparts. An individual carrying a single recessive deleterious allele will be healthy and can easily pass the deleterious allele into the next generation.

When the population is large, this is generally not a problem—the population may carry many recessive deleterious alleles, but they are rarely expressed. However, when the population becomes small, close relatives end up mating with one another, and those relatives likely carry the same recessive deleterious alleles. When the relatives mate, the offspring may inherit two copies of the same recessive deleterious allele and suffer the consequences of expressing the deleterious allele, as shown in the example below. In the case of the Swedish adders, that meant stillborn offspring and deformities.

4.3.2 Inbreeding depression and natural selection

cannot effectively remove all deleterious recessive genes from a population for several reasons. First, deleterious genes arise constantly through mutation within a population. Second, in a population where inbreeding occurs frequently, most offspring will have some deleterious traits, so few will be more fit for survival than the others. It should be noted, though, that different deleterious traits are extremely

unlikely to equally affect reproduction – an especially disadvantageous recessive trait expressed in a homozygous recessive individual is likely to eliminate itself, naturally limiting the expression of its phenotype. Third, recessive deleterious alleles will be "masked" by heterozygosity, and so in a dominant-recessive trait, heterozygotes will not be selected against.

When recessive deleterious alleles occur in the heterozygous state, where their potentially deleterious expression is masked by the corresponding wild-type allele, this masking phenomenon is referred to as complementation .

In general, sexual reproduction in eukaryotes has two fundamental aspects: recombination during meiosis, and outcrossing. It has been proposed that these two aspects have two natural selective advantages respectively. A proposed adaptive advantage of meiosis is that it facilitates recombinational repair of DNA damages that are otherwise difficult to repair. A proposed adaptive advantage of outcrossing is complementation, which is the masking of deleterious recessive alleles . The selective advantage of complementation may largely account for the general avoidance of inbreeding .

4.4 Heterosis

Heterosis, hybrid vigor, or outbreeding enhancement, is the improved or increased function of any biological quality in a hybrid offspring. The adjective derived from heterosis is heterotic. An offspring exhibits heterosis if its traits are enhanced as a result of mixing the genetic contributions of its parents. These effects can be due to Mendelian or non-Mendelian inheritance.

Heterosis, also called hybrid vigour, the increase in such characteristics as size, growth rate, fertility, and yield of a hybrid organism over those of its parents. Plant and animal breeders exploit heterosis by mating two different pure-bred lines that have certain desirable traits. The first-generation offspring generally show, in greater measure, the desired characteristics of both parents. This vigour may decrease, however, if the hybrids are mated together; so the parental lines must be maintained and crossed for each new crop or group desired.

4.4.1 Epigenetic basis of heterosis

Since the early 1900s (as discussed in the article Dominance versus overdominance) two competing genetic hypotheses, not necessarily mutually

exclusive, have been developed to explain hybrid vigor. More recently, an epigenetic component of hybrid vigor has also been established.

The genetic dominance hypothesis attributes the superiority of hybrids to the masking of expression of undesirable (deleterious) recessive alleles from one parent by dominant (usually wild-type) alleles from the other (see Complementation (genetics)). It attributes the poor performance of inbred strains to the expression of homozygous deleterious recessive alleles. The genetic overdominance hypothesis states that some combinations of alleles (which can be obtained by crossing two inbred strains) are especially advantageous when paired in a heterozygous individual. This hypothesis is commonly invoked to explain the persistence of some alleles (most famously the Sickle cell trait allele) that are harmful in homozygotes. In normal circumstances, such harmful alleles would be removed from a population through the process of natural selection. Like the dominance hypothesis, it attributes the poor performance of inbred strains to expression of such harmful recessive alleles. In any case, outcross matings provide the benefit of masking deleterious recessive alleles in progeny. This benefit has been proposed to be a major factor in the maintenance of sexual reproduction among eukaryotes, as summarized in the article Evolution of sexual reproduction.

An epigenetic contribution to heterosis has been established in plants, and it has also been reported in animals. MicroRNAs (miRNAs), discovered in 1993, are a class of non-coding small RNAs which repress the translation of messenger RNAs (mRNAs) or cause degradation of mRNAs. In hybrid plants, most miRNAs have non-additive expression (it might be higher or lower than the levels in the parents). This suggests that the small RNAs are involved in the growth, vigor and adaptation of hybrids.

It was also shown that hybrid vigor in an allopolyploid hybrid of two Arabidopsis species was due to epigenetic control in the upstream regions of two genes, which caused major downstream alteration in chlorophyll and starch accumulation. The mechanism involves acetylation and/or methylation of specific amino acids in histone H3, a protein closely associated with DNA, which can either activate or repress associated genes.

4.5 Molecular analysis of quantitative traits

The improvement of quantitative traits has been an important goal for many plant breeding programs. With a pedigree breeding program, the breeder will cross two parents and practice selection until advanced-generation lines with the best phenotype for the quantitative trait under selection are identified. These lines will then be entered into a series of replicated trials to further evaluate the material with the goal of releasing the best lines as a cultivar. It is assumed that those lines which performed best in these trials have a combination of alleles most favorable for the fullest expression of the trait.

This type of program, though, requires a large input of labor, land, and money. Therefore plant breeders are interested in identifying the most promising lines as early as possible in the selection process. Another way to state this point is that the breeder would like to identify as early as possible those lines which contain those QTL alleles that contribute to a high value of the trait under selection. Plant breeders and molecular geneticists have joined efforts to develop the theory and technique for the application of molecular genetics to the identification of QTLs.

Molecular makers associated with QTLs are identified by first scoring members of a random segregating population for a quantitative trait. The molecular genotype (homozygous Parent A, heterozygous, or homozygous parent B) of each member of the population is then determined. The next step is to determine if an association exists between any of the markers and the quantitative trait.

The most common method of determining the association is by analyzing phenotypic and genotypic data by one-way analysis of variance and regression analysis. For each marker, each of the genotypes is considered a class, and all of the members of the population with that genotype are considered an observation for that class. (Data is typically pooled over locations and replications to obtain a single quantitative trait value for the line.) If the variance for the genotype class is significant, then the molecular marker used to define the genotype class is considered to be associated with a QTL. For those loci that are significant, the quantitative trait values are regressed onto the genotype. The R^2 value for the line is considered to be the amount of total genetic variation that is explained by the specific molecular marker. The final step is to take those molecular marker loci that are associated the quantitative trait and perform a multiple regression analysis.

From this analysis, you will obtain an R^2 value which gives the percentage of the total genetic variance explained by all of the markers.

Quantitative trait loci (QTLs) are stretches of DNA containing or linked to the genes that underlie a quantitative trait. Mapping regions of the genome that contain genes involved in specifying a quantitative trait is done using molecular tags such as AFLP or, more commonly, SNPs. This is an early step in identifying and sequencing the actual genes underlying trait variation. Quantitative traits refer to phenotypes (characteristics) that vary in degree and can be attributed topolygenic effects, i.e., product of two or more genes, and their environment.

4.5.1 The quantitative traits

Polygenic inheritance refers to inheritance of a phenotypic characteristic (trait) that is attributable to two or more genes and can be measured quantitatively.Multifactorial inheritance refers to polygenic inheritance that also includes interactions with the environment. Unlike monogenic traits, polygenic traits do not follow patterns of Mendelian inheritance (separated traits). Instead, their phenotypes typically vary along a continuous gradient depicted by a bell curve.

An example of a polygenic trait is human skin color variation. Several genes factor into determining a person's natural skin color, so modifying only one of those genes can change skin color slightly or in some cases moderately. Many disorders with genetic components are polygenic, including autism, cancer, diabetes and numerous others. Most phenotypic characteristics are the result of the interaction of multiple genes.

4.5.2 Impotance

1. A quantitative trait locus (QTL) is a region of DNA that is associated with a particular phenotypic trait. These QTLs are often found on different chromosomes. Knowing the number of QTLs that explains variation in the phenotypic trait tells us about the genetic architecture of a trait. It may tell us that plant height is controlled by many genes of small effect, or by a few genes of large effect.
2. Another use of QTLs is to identify candidate genes underlying a trait. Once a region of DNA is identified as contributing to a phenotype, it can be sequenced. The DNA sequence of any genes in this region can then be

compared to a database of DNA for genes whose function is already known.

3. In a recent development, classical QTL analyses are combined with gene expression profiling i.e. by DNA microarrays. Such expression QTLs (eQTLs) describe cis- and trans-controlling elements for the expression of often disease-associated genes. Observed epistatic effects have been found beneficial to identify the gene responsible by a cross-validation of genes within the interacting loci with metabolic pathway- and scientific literature databases.

4.5.3 QTL mapping

For organisms whose genomes are known, one might now try to exclude genes in the identified region whose function is known with some certainty not to be connected with the trait in question. If the genome is not available, it may be an option to sequence the identified region and determine the putative functions of genes by their similarity to genes with known function, usually in other genomes. This can be done using BLAST, an online tool that allows users to enter a primary sequence and search for similar sequences within the BLAST database of genes from various organisms. It is often not the actual gene underlying the phenotypic trait, but rather a region of DNA that is closely linked with the gene.

Another interest of statistical geneticists using QTL mapping is to determine the complexity of the genetic architecture underlying a phenotypic trait. For example, they may be interested in knowing whether a phenotype is shaped by many independent loci, or by a few loci, and do those loci interact. This can provide information on how the phenotype may be evolving.

In order to begin a QTL analysis, scientists require two things.

1. First, they need two or more strains of organisms that differ genetically with regard to the trait of interest. For example, they might select lines fixed for different alleles influencing egg size (one large and one small).

2. Second, genetic markers that distinguish between these parental lines. Molecular markers are preferred for genotyping, because these markers are unlikely to affect the trait of interest. Several types of markers are used, including single nucleotide polymorphisms (SNPs), simple sequence

repeats(SSRs, or microsatellites), restriction fragment length polymorphisms (RFLPs), and transposable element positions.

Then, to carry out the QTL analysis, the parental strains are crossed, resulting in heterozygous (F_1) individuals, and these individuals are then crossed using one of a number of different schemes . Finally, the phenotypes and genotypes of the derived (F_2) population are scored. Markers that are genetically linked to a QTL influencing the trait of interest will segregate more frequently with trait values (large or small egg size in our example), whereas unlinked markers will not show significant association with phenotype (Figure 1)

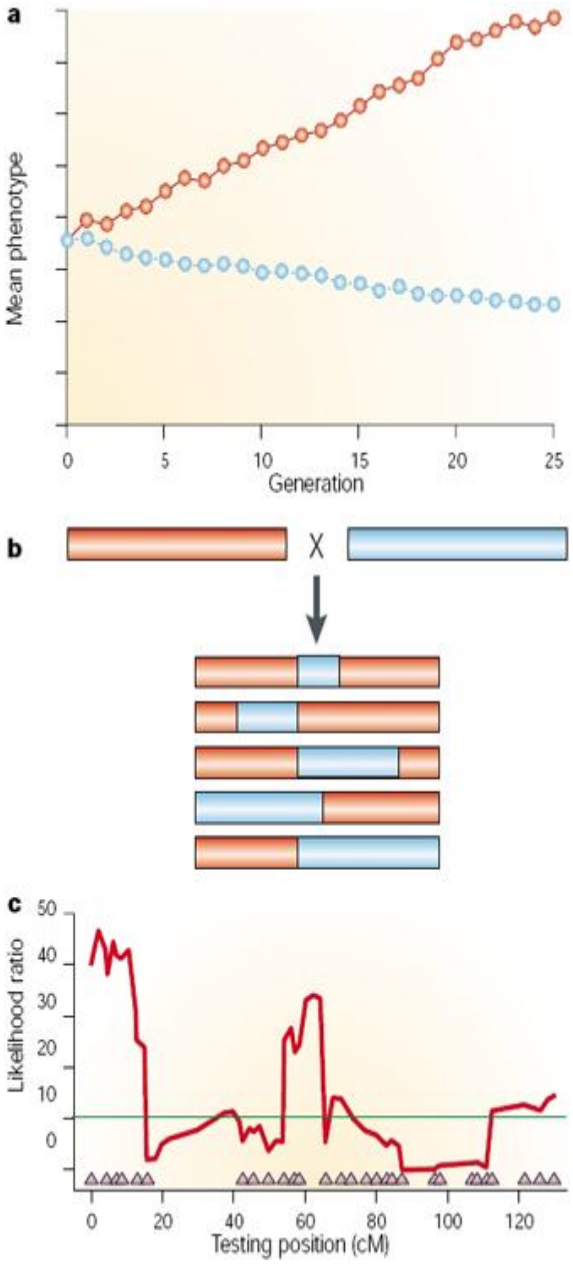


Figure 1: Quantitative trait locus mapping.
a) Quantitative trait locus (QTL) mapping requires parental strains (red and blue plots) that differ genetically for the trait, such as lines created by divergent artificial selection. b) The parental lines are crossed to create F1 individuals (not shown), which are then crossed among themselves to create an F2, or crossed to one of the parent lines to create backcross progeny. Both of these crosses produce individuals or strains that contain different fractions of the genome of each parental line. The phenotype for each of these recombinant individuals or lines is assessed, as is the genotype of markers that vary between the parental strains. c) Statistical techniques such as composite interval mapping evaluate the probability that a marker or an interval between two markers is associated with a QTL affecting the trait, while simultaneously controlling for the effects of other markers on the trait. The results of such an analysis are presented as a plot of the test statistic against the chromosomal map position, in recombination units (cM). Positions of the markers are shown as triangles. The horizontal line marks the significance threshold. Likelihood ratios above this line are formally significant, with the best estimate of QTL positions given by the chromosomal position corresponding to the highest significant likelihood ratio. Thus, the figure shows five possible QTL, with the best-supported QTL around 10 and 60 cM.

alleles; the backcross progeny would have anywhere from four to eight upper-case alleles.

A principal goal of QTL analysis has been to answer the question of whether phenotypic differences are primarily due to a few loci with fairly large effects, or to many loci, each with minute effects. It appears that a substantial proportion of the phenotypic variation in many quantitative traits can be explained with few loci of large effect, with the remainder due to numerous loci of small effect.

For example, in domesticated rice (Oryza sativa), studies of flowering time have identified six QTL; the sum of the effects of the top five QTL explains 84% of the variation in this trait . Once QTL have been identified, molecular techniques can be employed to narrow the QTL down to candidate genes . One important emerging trend in these analyses is the prominent role of regulatory genes, or genes that code for transcription factors and other signaling proteins. For instance, in rice, three flowering time QTL have been identified at the molecular level, and all of these loci encode regulatory proteins known from studies of Arabidopsis thaliana .

Another consistent trend in looking at QTL across traits and taxa is that phenotypes are frequently affected by a variety of interactions (e.g., genotype-by-sex, genotype-by-environment, and epistatic interactions between QTL), although not all QTL studies are designed to detect such interactions. Indeed, several complex

traits in the fruit fly Drosophila melanogaster have been extensively analyzed, and this research has indicated that the effects of such interactions are common . For example, detailed examination of life span in D. melanogaster has revealed that many genes influence longevity . In addition, significant dominance, epistatic, and genotype-by-environment effects have also been reported for life span.

4.6 Phenotypic Plasticity

Phenotypic plasticity is the ability of an organism, a single genotype, to exhibit different phenotypes in different environments. Such plasticity is nearly ubiquitous in nature and occurs in various animal and plant phenotypes, including behavior, physiology, and morphology.

Phenotypic plasticity may be observed as both adaptive and nonadaptive responses to the biotic or abiotic environment.

The continuity across levels of traits and across environmental boundaries arises because the mechanisms that enable the plastic responses at each of these levels are fundamentally the same. First, all responses are stimulated by a signal from the environment, whether the result is a change in protein production, physiological activity, growth or behavior. Second, all environmental signals,internal or external, must be received and processed at the level of individual cells.

The hierarchical levels transcriptome and proteome, which may be relatively unfamiliar to most ecologists, in order to focus attention on the fundamental processes that underlie plastic responses of all types. The transcriptome refers to those genetic sequences, in a cell at a specific time, that have been transcribed but not yet translated. The proteome then indicates the products of translation that are available for cellular metabolism or signaling. Why separate these processes that we have all learned as the single phrase ‘transcription-and-translation’? Because (in a specific cell at a specific time) not all genes are necessarily transcribed, nor are all transcripts necessarily translated, nor indeed are all proteins/polypeptides necessarily functionally active.

Cells and organisms can exert a remarkable degree of control over these processes. Differential control over the processes of transcription, translation and activation occurs dynamically in an integrated and coordinated manner as a response to internal and environmental influences.

4.6.1 Examples of phenotypic plasticity

Plants

Phenotypic plasticity in plants includes the allocation of more resources to the roots in soils that contain low concentrations of nutrients and the alteration of leaf size and thickness. *Dandelion* are well known for exhibiting considerable plasticity in form when growing in sunny versus shaded environments. The transport proteins present in roots are also changed depending on the concentration of the nutrient and the salinity of the soil. Some plants, Mesembryanthemum crystallinum for example, are able to alter their photosynthetic pathways to use less water when they become water- or salt-stressed.

Animals

As compared with plants, animals clearly show less plasticity in gross morphology. Nonetheless, developmental effects of nutrition and temperature have been demonstrated. Other generalities include the following: behavior is very plastic; in vertebrates, skeletal muscle is more plastic than the lung; skeletal muscle is more plastic in mammals than in lizards; snake guts are very plastic; carp are very plastic as a species. For example, in the Speckled Wood butterfly, the males of this species have two morphs. One with three dots on its hind wing, and one with four dots on its hind wings. The development of the fourth dot is dependent on environmental conditions, more specifically location, and the time of year.

Temperature

Plastic responses to temperature are essential among ectothermic organisms, as all aspects of their physiology are directly dependent on their thermal environment. As such, thermal acclimation entails phenotypic adjustments that are found commonly across taxa, such as changes in the lipid composition of cell membranes. Temperature change influences the fluidity of cell membranes by affecting the motion of the fatty acyl chains of glycerophospholipids. Because maintaining membrane fluidity is critical for cell function, ectotherms adjust the phospholipid composition of their cell membranes such that the strength of van der Waals forces within the membrane is changed, thereby maintaining fluidity across temperatures.

Parasitism

Infection with parasites can induce phenotypic plasticity as a means to compensate for the detrimental effects caused by parasitism. Commonly, invertebrates respond to parasitic castration or increased parasite virulence with fecundity compensation in order to increase their reproductive output, or fitness. For example,water fleas (Daphnia magna), exposed to microsporidian parasites produce more offspring in the early stages of exposure to compensate for future loss of reproductive success.

Hosts can also respond to parasitism through plasticity in physiology aside from reproduction. House mice infected with intestinal nematodes experience decreased rates of glucose transport in the intestine. To compensate for this, mice increase the total mass of mucosal cells, cells responsible for glucose transport, in the intestine. This allows infected mice to maintain the same capacity for glucose uptake and body size as uninfected mice.

4.6.2 When the phenotypic plasticity is selected and favored

A perfectly plastic genotype (i.e., one that converts immediately to the optimal phenotype when conditions change) will always be favored to evolve whenever environments are heterogeneous. However, two limitations can probably be recognized in its description, namely the capacity for immediate change and the production of the optimal phenotype. Theoretical investigations have delineated sets of conditions favoring the evolution of generalist (i.e., plastic) vs. specialist genotypes. The outcome of selection in these studies has hinged upon several factors: the likelihood of environmental change, the predictiveness of the environmental cues (i.e., the correlation between the cue and future conditions), whether variation is spatial or temporal, and whether a change in conditions will be encountered within (fine-grained) or between (coarsegrained) generations.

Plasticity is selected against:

(1) if the environmental change is rare,

(2) if reliable environmental cues are lacking,

(3) if the environment fluctuates more rapidly than the typical response time,

(4) if a single phenotype is optimal in both environments,

(5) if the environment is spatially coarse-grained, and the organism can choose

its habitat.

Plasticity is favored:

(1) if environmental change is frequent,

(2) if environmental cues are reliable,

(3) if environmental variation is temporally or spatially fine-grained,

(4) if environmental variation is temporally coarse-grained with predictive cues (polyphenism),

(5) if environmental variation is temporally fine-grained in a predictable sequence (heteroblasty)

4.7 Summary

Gene-environment interaction is when two different genotypes respond to environmental variation in different ways. Gene-environment interactions are studied to gain a better understanding of various phenomenon. Inbreeding depressions is the reduced biological fitness in a given populations a result of inbreeding- i.e. breeding of related individuals. Hetrosis, hybrid vigour or outbreeding enhancement , is the improved or increased function of any biological quality in a hybrid offspring.Quantitative trait loci(QTLs) are stretches of DNA containing or linked to the genes that underline a quantitative trait. Polygenic inheritance refers to inheritance of a phenotypic characteristic (trait) that is attributable to two or more genes and can be measured quantitatively. Multifactorial inheritance refers to polygenic inheritance that also includes interactions with the environment. Phenotypic plasticity is the ability of an organism , a single genotype , to exhibit different phenotypes in different environments.

4.8 Self-Learning Exercise

Section A

1. What is genotype environment interactions?
2. Define inbreeding depressions.
3. Explain heterosis.
4. What is QTL mapping?

Section B

1. Explain four examples of gene-environment interactions.
2. Enumerate the mechanism of Inbreeding depression.
3. Explain the relation between inbreeding depression and natural selection.
4. Give the epigenetic basis of hetrosis.

Section C

1. Write detailed notes on Molecular analysis of quantitative traits.
2. Write an essay on phenotype plasticity.

Unit -5

Genetics of speciation: phylogenetic and biological concept of species, patterns and mechanisms of reproductive isolation, models of speciation (Allopatric, Sympatric, Parapatric)

Structure of the Unit

5.0 Objectives

After going through this unit you will be able to understand

- Species concept
- Pattern and mechanism of reproductive isolation
- Different models of speciation

5.1 Introduction

Species Concept

When we observe an organism the immediate reaction of our brain is to know about it . What is this organism? Having background of biological sciences, we want to know about its designation. Designation of an organism can best understood by two words the first is Genus and the second is species. The diversity of life around us is because we are surrounded by different type of species. Plant, and animal microorganism species.

Darwin was the first to propose natural mechanisms for evolutionary transformations of species and laid groundwork for scientific study of speciation.

I look at the term species, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other…. In short, we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience. This may not be a cheering prospect; but we shall at least be freed from the vain search for the undiscovered and undiscoverable essence of the term species.

– Charles Darwin, 1859

The Origin of Species

5.2 Genetics of Speciation

Now we know how the term was used by the naturalist in their description about species. Most modern biologists treat species as a fundamental natural unit. They Identify research organisms at species level and communicate by Linnaean binomial system for an international Uniformity. In practice, most species are recognized phonetically by consistent differences in easily identifiable morphological characters called traits. Species are also diagnosed by obvious reliable characters possessed by almost all individuals and not possessed by individuals of other species. Due to similarities in characters individuals of a species breed and produce reproductively capable offsprings. Therefore, species is also defined as reproductively breeding group of individuals in a population. Thus,

Species are sets of populations of organisms held together by reproduction and genetic compatibility.

Species are kept separate from other species to the extent that they remain reproductively isolated from one another under normal conditions. Species is a unit in a population that participate in evolutionary processes (gene flow, adaptation, etc.). In the course of time under the evolutionary process organisms copulate ,genes replicate, variations are produced, species evolve that results into speciation. You have seen Indian Bull frog (*Hoplobatrichus tigerinus)* and Common Indian toad (*Duttaphrynus melanostictus)* Now you can understand that *Hoplobatrichus tigerinus* and and *Duttaphrynus melanostictus represents two* different genera. But Common Indian toad *Duttaphrynus melanostictus and Duttaphrynus stomaticus,* although one genus but two different species

5.2.1Biological and Phylogenetic concept of species

(1) Biological species concept means interbreeding individuals. This concept defines species in terms of interbreeding individuals.

Species are groups of interbreeding natural populations that are reproductively isolated from other such groups. This is the Most widely accepted species concept and Predates Darwin. Biological species concept explains why members of species (i) resemble one another, and (ii) differ from other species.

Taxonomists usually identify species by morphology, not by testing their reproductive competence. Although, morphological characters shared among the individuals of a species are indicators of interbreeding. It is always not necessary that individuals of a species are morphologically uniform. Sometimes individuals of a species show different degree of variations in their morphological characters, called polytypic or morpho- variants.

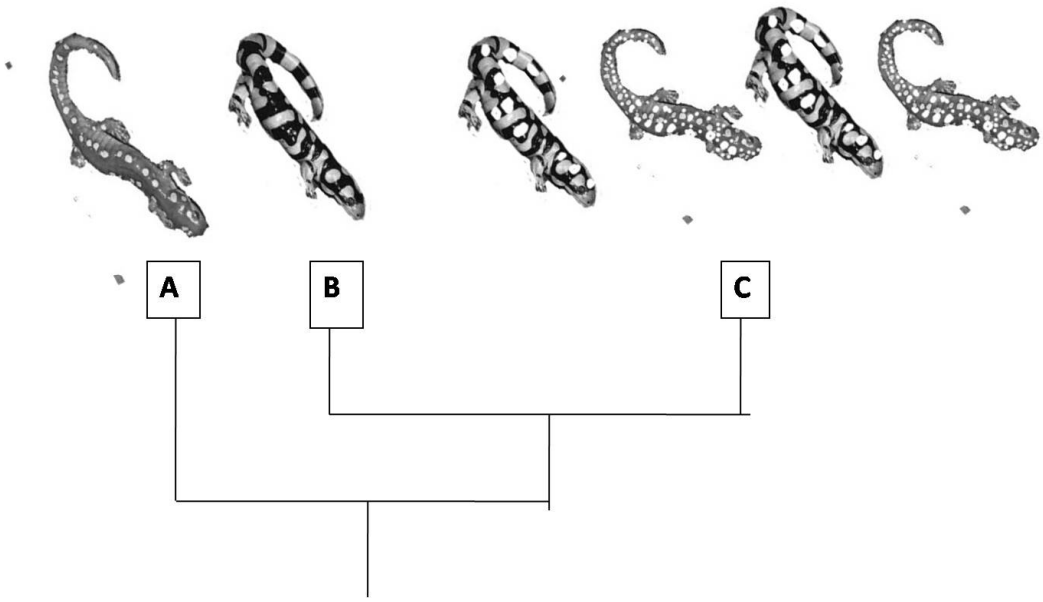
The Biological Species concept does not stand well under certain conditions. In nature, there are lots of situations where it is difficult to apply this definition. For example, many bacteria reproduce mainly asexually. The bacterium shown at right is reproducing asexually, by binary fission. The definition of a species as a group of interbreeding individuals cannot be easily applied to organisms that reproduce only or mainly asexually. Similarly, many plants, and some animals, form hybrids in nature. Hooded crows and carrion crows look different, and largely mate within

their own groups but in some areas, they hybridize. Should they be considered the same species or separate species?

The biological species concept has its limitations ,although, it works well for many organisms and has been very influential in the growth of evolutionary theory. In order to address some of these limitations, many other "species concepts" have been proposed, such as phylogenetic species concept etc.

Phylogenetic species concepts •

Phylogenetic species concept: a species is a "tip" on a phylogeny, that is, the smallest group of organisms that share an ancestor and can be distinguished from other such groups. Under this definition, a ring species is a single species that encompasses a lot of phenotypic variation. In this example, *Ensatina* salamander lineages A and B are separate species. Each has a common ancestor that individuals of other species do not. Even though it has diversified a lot, Lineage C is a single species, according to the phylogenetic species concept. None of the subspecies of Lineage C has a single common ancestor separate from the other such species.



5.2.2 Patterns and mechanism of Reproductive isolation

Mechanism that prevent gene exchange have been broadly termed **isolating mechanism .** In other word all factors that prevent gene exchange in between

different species are covered under the broad spectrum of isolation mechanism. Such geographically separated populations , obviously do not have the opportunity for gene exchange and remain reproductively isolated.

The term isolating mechanisms be restricted to those that prevent gene exchange among populations in the same geographic locality. Isolating mechanisms can be classified into two broad categories (1)Those that operate before fertilization can occur (**pre mating**) and (2) those that operate afterward (**post mating**) .

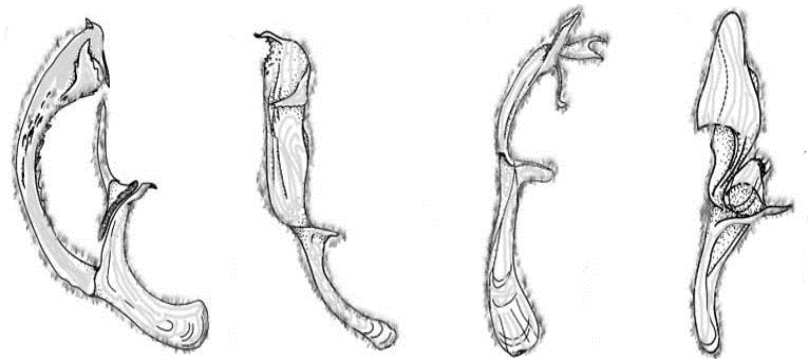
Pre mating isolating mechanisms

1) **Seasonal or habitat isolation:** Potential mates do not meet because they flourish in different seasons or in different habitats. some plant species such as the spiderworts *Tradescantia canaliculata* and *T. subaspera* for example, are sympatric throughout their geographical distribution, yet remain isolated because their flowers boom at different seasons. Also one species grow in sunlight and the other in deep shade.
2) **The evolution of different mating location, mating time, or mating rituals:** Genetically-based changes to these aspects of mating could complete the process of reproductive isolation and speciation. For example, bowerbirds construct elaborate bowers and decorate them with different colors in order to woo females. If two incipient species evolved differences in this mating ritual, it might permanently isolate them and complete the process of speciation.

3) **Behavioral or sexual isolation:** The sex of two species of animals may be found together in the same locality , but their courtship patterns are sufficiently different to prevent mating. The distinctive song of many birds, the special mating calls of certain frogs, and the sexual display of many animals are generally attractive only the mates of the same species. Numerous plants have floral displays that attached only certain insect pollinators. Even where the morphological differences between two species is minimal, behavioral differences may suffice to prevent cross fertilization. Thus *D. melanogaster* and *D. simulans*, designated as sibling species because of their morphological similarity, will normally not mate with each other even when kept together in a single population cage.

4) **Mechanical isolation:** Mating is attempted, but fertilization cannot be achieved because of difficulty in fitting together male female genitalia. This type of incompatibility, long through to be a primary isolating mechanism in animals is no longer considered important. There is little evidence that mating in which the genitalia are markedly different are even seriously attempted, although some exceptions exist among fruitfly *Drosophila* species and some other group.

Lack of "fit" between sexual organs: Hard to imagine for us, but a big issue for insects with variably-shaped genitalia.



Male genitalia of four different species of Drosophila

These Drosophila male genital organs illustrate just how complex insect genitalia may be.

Post mating mechanism - These prevent the success of an inter populational cross even though mating has taken place.

1**)** **Gametic mortality:** In this mechanism, either sperm or egg is destroyed because of interspecific cross. Pollen grains in plant, for example, may be unable to grow pollen tubes in the styles of foreign species. In some *Drosophila* crosses it has been shown that an insemination reaction takes place in the vaginal of the female that causes swelling in this organ and prevents successful fertilization of the egg.

2) **Zygotic mortality and hybrid inviability:** The egg is fertilized but the zygote either does not develop, or it develops into an organism with reduced viability. Numerous instances of this type of incompatibility have been observed in both plants and animals. In an experiment crosses between 12 frog species of the genus *Rana* and found the wide range of inviability. In some crosses, no egg cleavage could be observed; in others, the cleavage and the blastula stages were normal but gastrulation failed; and instill others, early development was normal but later stages failed to develop.

3) **Hybrid sterility:** The hybrid has normal viability but is reproductively deficient or sterile. This is exemplified in the mule (progeny of a male donkey and female horse) and many other hybrids. Sterility in such case may be caused by interaction between the cytoplasm from the source and the chromosomes from the other. In general, the barriers separating species are not confined to a single mechanism. The *Drosophila* sibling species *D. pseudoobscura* and *D. persimilis* are isolated from each other by habitat (*persimilis* usually lives in cooler regions and at higher elevations) courtship period (*persimilis* is usually more active in the morning, *pseudoobscura* in the evening) and the mating behavior (the females prefer males of their own species).Thus, although the distribution range of these two species overlap throughout a large area of habitat, these isolation mechanisms are sufficient to keep the two species apart. However, even when cross-fertilization occurs between these two species, gene exchange is still impeded, since the F_1 hybrid male is completely sterile and the progenies of fertile F_1 females backcrossed to males of either species show markedly lower viabilities than the parental stocks.

5.2.3 Models of speciation (Allopatric, sympatric, parapatric)

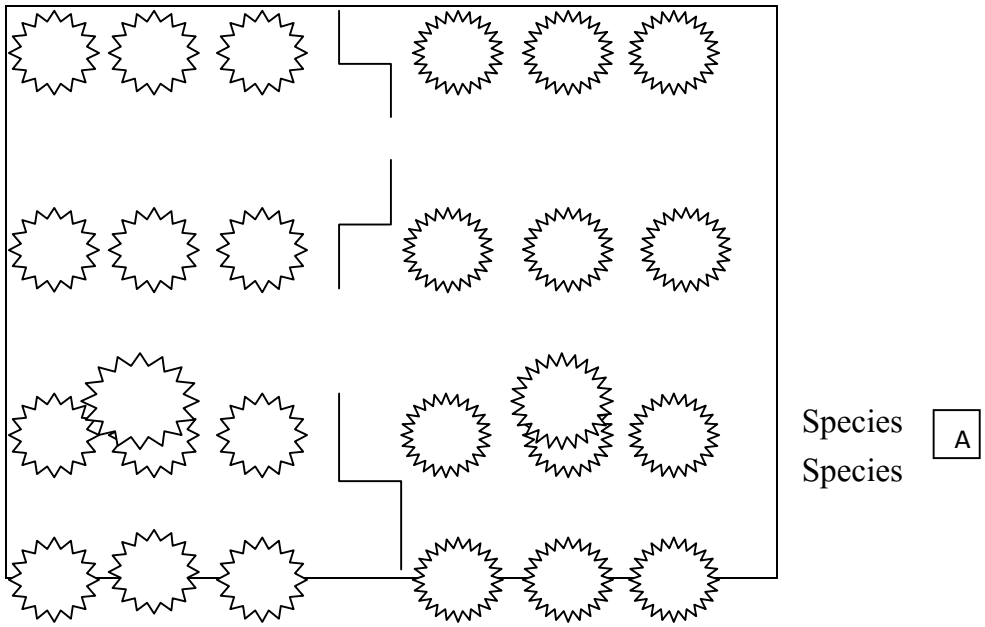
Speciation

The process of splitting a genetically homogenous population into two or more populations that undergo genetic differentiation and eventual reproductive isolation is called speciation. The key to speciation is the evolution of genetic differences between the incipient species. For a lineage to split once and for all, the two incipient species must have genetic differences that are expressed in some way that causes mating between them to either not happen or to be unsuccessful. These need not be huge genetic differences. A small change in the timing, location, or rituals

of mating could be enough. But still, some difference is necessary. This change might evolve by natural selection or genetic drift. Reduced gene flow probably plays a critical role in speciation. Modes of speciation are often classified according to how much the geographic separation of incipient species can contribute to reduced gene flow. There are three important models of speciation namely allopatric, sympatric and parapatric speciation.

Allopatric Speciation:-

In the absence of gene flow between geographically separate populations, daughter species form gradually by divergence. The formation of two or more species often requires geographical isolation of subpopulations of the species. Only then can natural selection or perhaps genetic drift produce distinctive gene pools. It is no accident that the various races (or "sub-species") of animals almost never occupy the same territory. Their distribution is allopatric ("other country"). Allopatric speciation is just a fancy name for speciation by geographic isolation, discussed earlier. In this mode of speciation, something extrinsic to the organisms prevents two or more groups from mating with each other regularly, eventually causing that lineage to speciate. Isolation might occur because of great distance or a physical barrier, such as mountain ranges, a desert or river.



Sympatric Speciation

Daughter species arise from a group of individuals within an existing population. Sympatric speciation refers to the formation of two or more descendant species from a single ancestral species all occupying the same geographic location..The sympatric species can arise either due to changes in the chromosome number or due to introgressive hybridization and polyploidy. The change in the chromosome number may occur by polyploidy, aneuploidy, haploidy or translocation. Examples of origin of species by these methods are quite frequent in plants but few in animals. For example Different species of Drosophila have different number and apperance of Chromosome.

i) *Drosophila melanogaster* and *D. ricanusame* have 4 pairs of chromosomes.

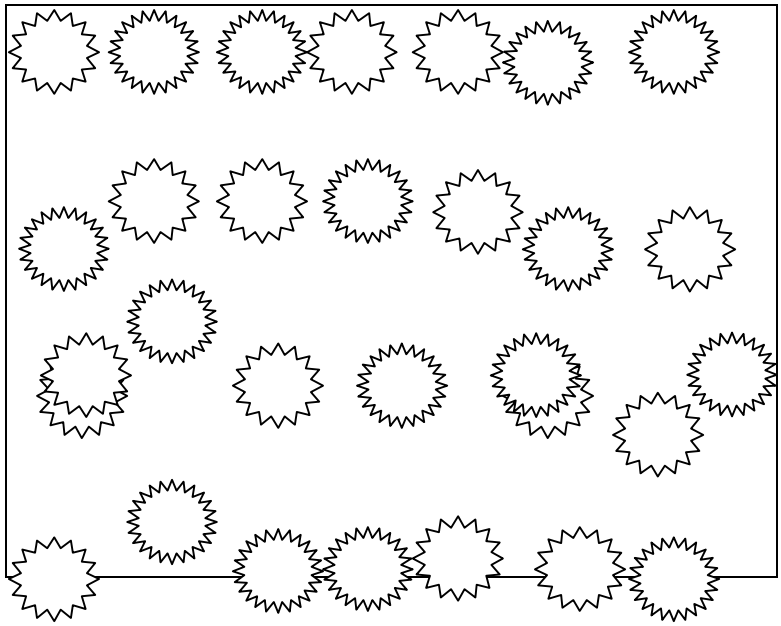
ii) *D. virilis* has 6 pairs of chromosome.

iii) *D. pseudo-obscura* and *D. persimilis* posses 6 pairs of chromosome, and

iv) *D. willistoni* possesses three pair of chromosomes.

The chromosomal composition of *D.*virilis is regarded to be of ancestral type, which possesses five pair of road shaped and one pair of dot like chromosome . The chromosomal complement of *D.* virilis by the translocation between X-chromosome and one of the autosomes so that these possess a pair of V-shaped , three pair of rod shaped and a pair of dot like chromosomes. The chromosome complement *of D. melanogasrte* could be derived by two translocation between two pair of V-shaped, a pair of rod shaped and a pair of dot like chromosome.

Unlike the alloptaric mode, sympatric speciation does not require large-scale geographic distance to reduce gene flow between parts of a population. How could a randomly mating population reduce gene flow and speciate? Merely exploiting a new niche may automatically reduce gene flow with individuals exploiting the other niche. This may occasionally happen when, for example, herbivorous insects try out a new host plant.



The best example is that about 200 years ago, the ancestors of apple maggot flies laid their eggs only on hawthorns but today, these flies lay eggs on hawthorns (which are native to America) and domestic apples (which were introduced to America by immigrants and bred). Females generally choose to lay their eggs on the type of fruit they grew up in, and males tend to look for mates on the type of fruit they grew up in. So hawthorn flies generally end up mating with other hawthorn flies and apple flies generally end up mating with other apple flies. This means that gene flow between parts of the population that mate on different types of fruit is reduced. This host shift from hawthorns to apples may be the first step toward sympatric speciation. In less than 200 years, some genetic differences between these two groups of flies have evolved.

Gene flow has been reduced between flies that feed on different food varieties, even though they both live in the same geographic area

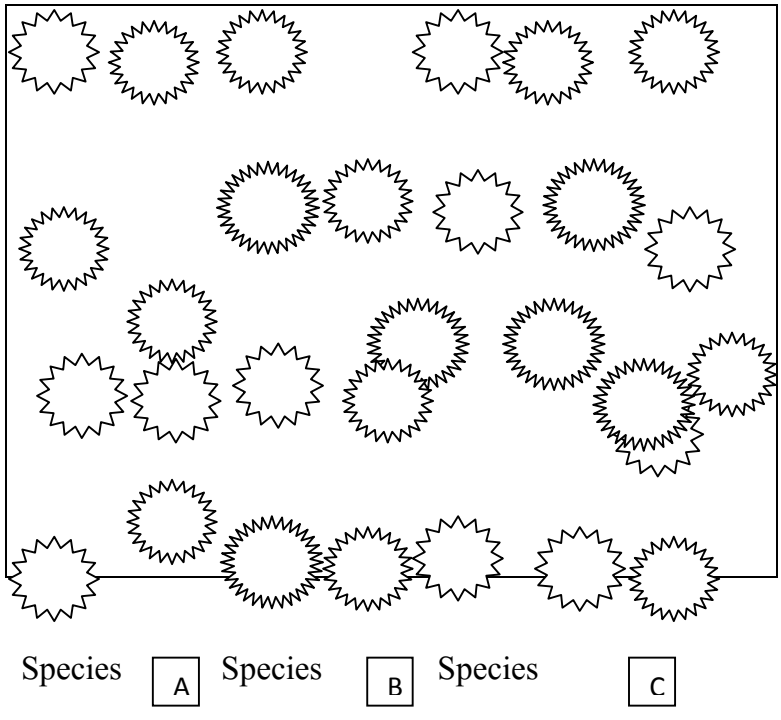
However, biologists question whether this type of speciation happens very often. In general, selection for specialization would have to be extremely strong in order to cause the population to diverge. This is because the gene flow operating in a randomly-mating population would tend to break down differences between the incipient species.

Parapatric Speciation:- Parapatric (para- near) speciation is the development of reproductive isolation among the members of a continuous population in the

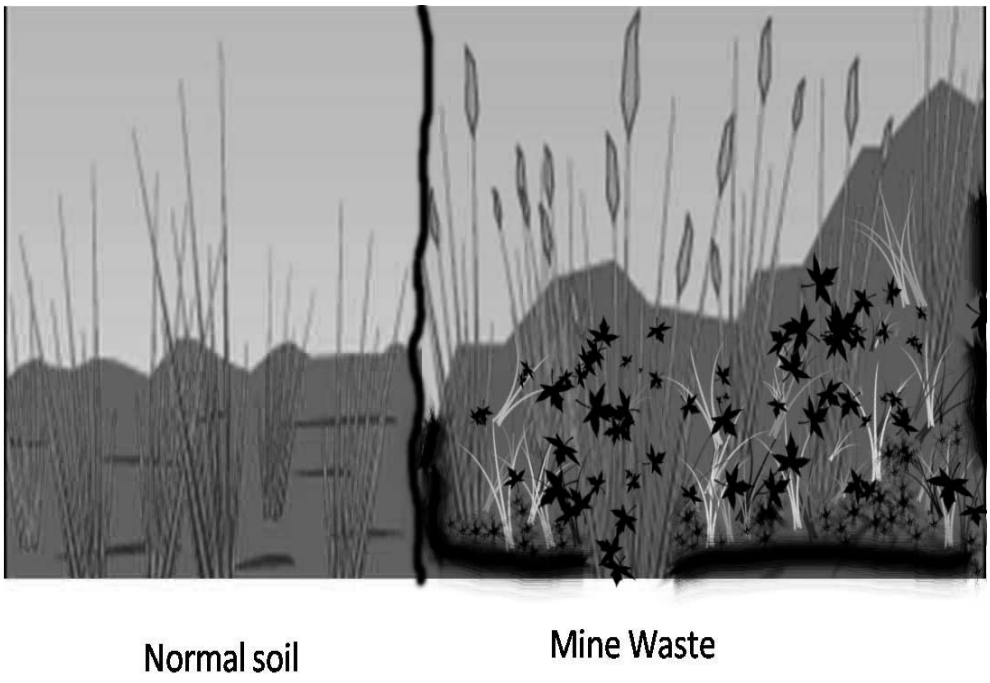
absence of geographical barriers. Chromosomal aberrations lead to partial reproductive isolation in the individuals of a population in some areas of its distribution. The lower fertility with further addition of structural changes in the chromosomes finally established reproductive isolation forming new species.

Parapatric speciation has been described for the land snail of the genus *Partula* on the island of Moorea near Tahiti. Eleven species of *Partula* described from Moorea, even though the island is only 15 Km. wide and there are no geographic barriers. These species fall into two groups. Among snail belonging to *Partula suturalis* complex, there are both sinistral and dextral individuals. The dextral and sinistral forms occur in different areas, these exhibits restricted hybridization , while in some other areas, the reproductive isolation is in the process of being established. The area of reproductive isolation are represented by jagged lines and are not isolated by any geographic barriers.

In parapatric speciation there is no specific extrinsic barrier to gene flow.The population is continuous, but nonetheless, the population does not mate randomly. Individuals are more likely to mate with their geographic neighbors than with individuals in a different part of the population's range. In this mode, divergence may happen because of reduced gene flow within the population and varying selection pressures across the population's range.



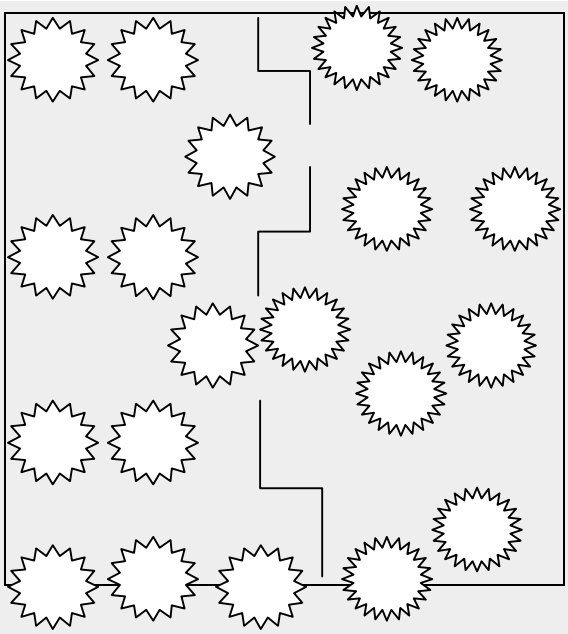
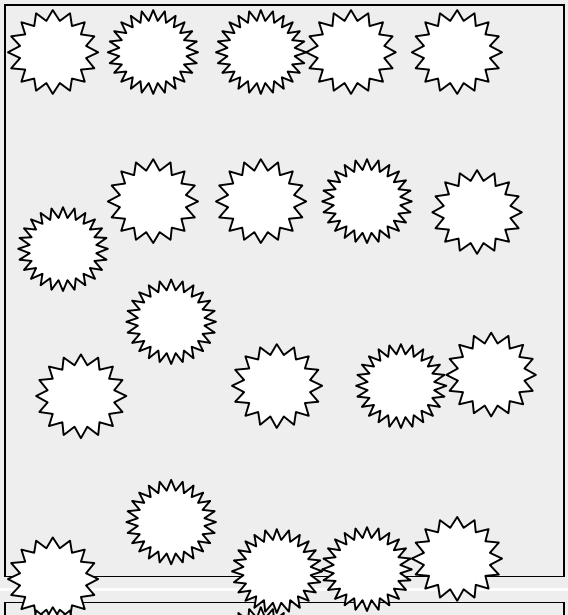
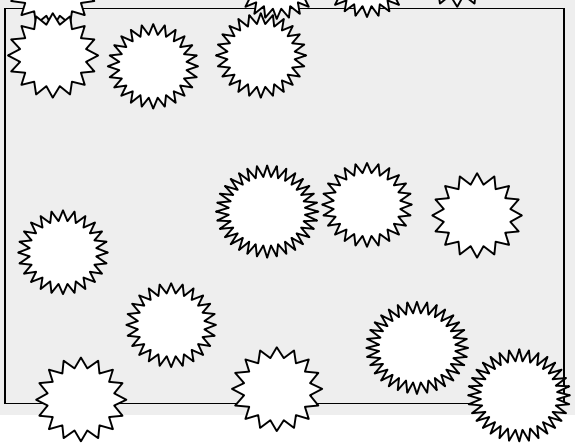
Parapatric speciation may also be observed in a grass species *Anthoxanthum odoratum.* Some of these plants live near mines where the soil has become contaminated with heavy metals. The plants around the mines have experienced natural selection for genotypes that are tolerant of heavy metals. Meanwhile, neighboring plants that don't live in polluted soil have not undergone selection for this trait. The two types of plants are close enough that tolerant and non-tolerant individuals could potentially fertilize each other so they seem to meet the first requirement of parapatric speciation, that of a continuous population. However, the two types of plants have evolved different flowering times. Although continuously distributed, different flowering times have begun to reduce gene flow between metal-tolerant plants and metal-intolerant plants. This change could be the first step in cutting off gene flow entirely between the two groups.

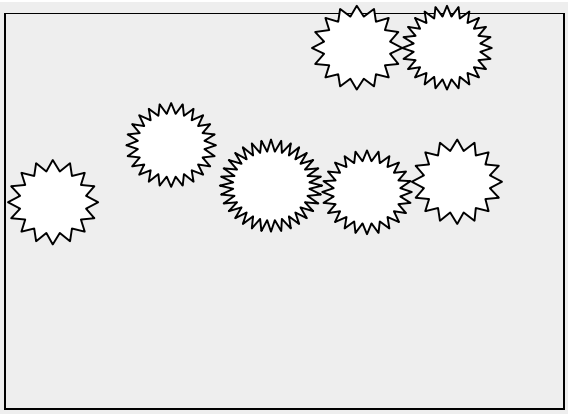


Tabular presentation of modes of Speciation- Table 1

The following table compares some of these speciation modes.

Mode of Speciation	New species formed from	Diagrammatic representation

Allopatric (allo = other, patric = place)	geographically isolated populations	
Sympatric (sym = same, patric = place)	within the range of the ancestral population	
Parapatric (para = beside, patric = place)	a continuously distributed population	



5.3 Summary

As per Biological Species Concept ,Species are groups of interbreeding natural populations that are reproductively isolated from other such groups. Phylogenetic species concept defines a specie as a “tip” on a phylogeny ,that is, the smallest group of organisms that share an ancestor and can be distinguished from other such groups .Mechanisms that prevent gene exchange have been broadly termed isolating mechanisms and can be categorized into pre-mating and post-mating isolating mechanisms .There are three modes of speciation namely Allopatric,Sympatric and Parapatric Speciation.

5.4 Glossary

- **Speciation:**The process of splitting a genetically homogenous population into two or more populations that undergo genetic speciation and eventual reproductive isolation is called speciation.
- **Allopatric Speciation(allo=other;patric=place):**In this mode of speciation new species are formed from geographically isolated populations.
- **Sympatric Speciation(sym=same;patric=place):**In this mode new species are formed within the range of the ancestoral populations
- **Parapatric Speciation(para=besides;patric=place):**In this mode species are formed within a continuosly distributed population.

5.5 Self Learning Exercises

Section A:

1. Define Species as per Phylogenetic Species Concept.
2. What is Speciation?
3. Name the three mode of Speciation.

Section B:

1. Explain the Biological concept of Species in detail .
2. Write about the allopatric mode of speciation with examples.
3. What is parapatric speciation?Explain.

Section C:

1. Write a detailed note on Speciation .
2. How will you distinguish between the three modes of speciation?

Unit - 6

Adaptation diversity & nature of adaptation: adaptive radiations & occupation of new environments & niches: mimicry and coloration

Structure of the Unit

6.0 Objectives

After going through this unit you will be able to understand :

- What is adaptive diversity?
- The nature and types of adaptation.
- Different types of adaptive radiations
- How the occupation of new environment is done?

6.1 Adaptive Diversity

In order to survive in an environment all living beings are equipped with certain characters called adaptive features. With the help of these adaptive features organism survive and reproduce to propagate their progeny, and gets adjusted and fitted with the environmental conditions. Therefore, adaptation is the morphological, anatomical, physiological and behavioral modifications in an organism to adjust itself in a particular environment. For example birds are

equipped with flying adaptations.These are provided with wings, light body weight etc. suited to fly. Similarly Fishes are provided with streamline body, fins etc which help in swimming. In the course of time environment keeps on changing. The organisms should acquire adaptive features for their survival in the changed scenario of environment or prepare for extinction if fails to adapt. Now it should be clear to you what is adaptation and what are adaptive features.

Types of adaptations

There are four major types of adaptations:

1. Structural
2. Physiological
3. Protective and
4. Animal association adaptation

Structural - These adaptations include structural organization of an organism with respect to the specific conditions of environment in which organism is inhabiting . Organisms live in varieties of habitats. In order to survive in such specific conditions there is need to develop some adaptive features which can help in survival of that organism. Such adaptations with reference to environmental conditions or habitats are as follows:

Cursorial adaptations: Organisms which are required to run swiftly on land require elongated long bones and reduced digits for example Horse, ungulates etc.. The tarsals and metatarsals are raised from the ground. In ungulates (hoofed) animals there is unguligrade progression. These walk on modified nail or hoof.

Fossorial adaptations: Burrowing and cave dwelling organisms require slender, poorly pigmented body, functionless eyes,, increased sense of smell and touch etc . The body contour is spindle shaped or fusiform so that very little resistance to subterranean passage.In these head tapers anteriorly to form a kind of snout for burrowing e.g. snakes, caecilians (Ichthyophis) etc..

Scansorial adaptations: Those animals which require climbing in their routine activities needs strong pectoral girdles, much elongated proximal segments of limbs, , prehensile foot and tail well developed claws and adhesive pads eg, monkeys, apes, sloths , Hyla etc.

Aerial adaptations (Volant adaptations) Those animals which use power of flight for various activities possess many structural modification which support them this

particular mode of life. They have streamlined body with reduced air resistance. Body is covered with feathers. Bones are light weight and pneumatized. Wings are used for flight.Wing muscles are very strong.Humerus and sternum are highly specialized and well developed. Sense of smell, touch and vision are more developed. E.g. Birds

Desert adaptations. Those terrestrial animals which live in desert conditions possess many adaptive features to survive under extreme xerophytic conditions. The desert habitats are characterized by scarcity of water and high temperature during summers. The animals living in desert have protection from heat, their skin is modified in a manner that does not permit moisture loss. For example lizards living in desert are having compactly covered scales in the body wall. The eyes, years and nostrils are well protected from dust .

Aquatic adaptations: The structural modifications in the organisms that support aquatic mode of life are called aquatic adaptations. Fishes are the best example of aquatic adaptations. Although, many amphibians, reptiles, birds and mammals are generally terrestrial or aerial but some representative cases too represent aquatic mode of life. For swimming fishes are characterized by stream lined body. Fins help in swimming. Gills help in respiration.

Parasitic adaptations. Those organisms which remain parasites (external or Internal) on some other organisms body, acquire many adaptive features to survive in this mode of life. These modifications include poorly developed alimentary canal, highly developed reproductive system, poor sense organs, eg Ascaris, Taenia etc.

Physiological adaptations.

Physiological adaptations are those functional changes in the body of an organism which help them to survive under changed environmental conditions related to physiological functions. For example Scoliodon (Elasmobranch) accumulates urea in its body to balance osmotic concentration with reference to marine water where they live. Birds and mammals maintain a constant body temperature by changes in many metabolic pathways and cardiovascular modifications.

Protective adaptations

Animals live in an environment that provides them food and shelter. But such habitats are also occupied by other animals and some acts as their natural enemies

in the food chain. Those adaptations which protect animals from their natural enemies are called protective adaptations. Such adaptations reduce risk from the enemies. Natural enemies use hunting tactics either visually or non visually. Therefore protective adaptation are also divided into two categories, i) Visual adaptations and ii) Non visual adaptations.

i) Visual adaptations- when animals hide them selves from their predator animals using body colouration similar to the surrounding,and makes them indistinguishable from the background, such colour harmony tones are called visual adaptations. For example polar bears, foxes, rabbits are white in colour and are invisible in polar regions dominated by ice. Some snakes ,frogs praying mantis etc makes their colour similar to their surrounding and are normally not distinguished from the back ground by their natural enemies.

Mimimicry- some animals imitates themselves not only with the colour of the surrounding environment but also in appearance, surface patterns, shape and other details. Such adaptations keeps them away from their natural enemies. Leaf insect *Phylium* looks very similar to their surrounding leaf patterns and similarly stick insect *Carausis* acquire shape of bunch of small sticks. We will discuss mimicry in detail in other section in this unit.

Visual adaptations are not only protective in nature but some times these are also used to hunt the pray. Tiger skin pattern often mix with the surrounding and ungulates may confuse and becomes pray. Some animals use visual adaptations fo warning. Newts and some red coloured frogs warns their predators with their bright colours.

ii) Non visual adaptations- some animals use non visual adaptive featurs to get away from their enemies. Some toad species leave very repelling odours to protect from their enemies. Some mollusks cover their body with hard shell. Spines on the body of hedgog is also another example of non visual adaptations.

Adaptive radiations:

The living organisms exhibit placticity in their organization. The animals belonging to different unrelated taxonomic groups when occupy the same habitat

in an ecosystem, exhibit some common features. Such taxonomically different groups sharing common features due to similar habitat becomes convergent groups. On the other hand when similar taxonomic groups moves in different habitats and acquire features as per the requirement of the diversified environment becomes different from their stock taxonomic group. Such organisms which have common taxonomic group but diverged because of migration to different habitats are called divergent groups.

Therefore, each isolated region when large in size and carrying varied topographic features, soil characteristics, vegetation etc will give rise to a diversified fauna in due course of time. The larger the region, more diverse the conditions, greater will be the varieties of species. Such radiating features from a stock population producing diversity of organisms is called adaptive radiations i.e. radiating trend with diversified adaptations with respect to varied conditions of the habitats. Therefore, adaptive radiation is evolution in several specialized directions starting from a common ancestral stock or entry of different organisms of original stock from different regions to a new adaptive zone to give rise similar characteristic features of body organization .

Divergent pattern

Adaptive radiations in Mammals.

The limb structure of placental mammals provides a classical example of adaptive radiation or evolutionary diversgence.The ancestry of all present day types of mammals can be traced back to primitive insect eating,five toed,short legged creature that walked with the soles of its flat feet.Its pentadactylus limbs were not modified for any particular type of locomotion.These lived on land and formed ancestors to modern mammnals.which now have occupied five different habitat with modifications in their limb structure.

1. One limb lead to arboreal modification seen in tree dwelling forms like squirrels and primates.
2. Another line acquired aerial modification found in animals adapted for flight
3. The third line presented cursorial modification.The term refers to those mammals adapted to fast running(horse,dog,deer etc)
4. The next line diverged to fossorial or burrowing mammnals.

5. The last line lead to aquatic adaptations.

In all these lines mammals exhibit limb structure modified for some particular mode of locomotion and these limb types could easily be deprived from one common type represented by short pentadactyle limbs of terrestrial mammals.This example suggest that adaptive radiation is a process by which a relatively generalized ancestral group gives rise to many relatively more specialized descendants

Adaptive radiations in finches

A similar divergent evolution has occurred in the Mesozoic aves and is also illustrated by the ground finches of Galapogos island situated on the equator some 900 Km west of Equador.Finches in general posses stout conical beaks adapted for crushing seeds.But they have undergone great diversification in their feeding habits and accordingly in the shape and size of their beak

(1) The ground finches of the subgenus Geospiza exhibited great variations in their beak structure.Although chiefly seed crushers the size of the beak is correlated with the size of the seeds they eat
(2) Warbler finch has a slender warbler like beak and is insectivores in habit
(3) Vegetarian tree finch have a short thick and somewhat parrot like beak and feed upon leaves buds and fruits
(4) Woodpecker finches have a stout and straight but long beak and are completely insectivores.These search bark and leaf clusters and bore into the woods like a woodpecker
(5) Insectivores finches have a beak very similar to a vegetarian finch but they feed upon beetles and other insects
(6) Cactus ground finches have a long somewhat decurved beak and a split tongue.It probes the flowers of prickly pear cacatus for nectar and feeds upon the soft pulp of the cactus

Convergent pattern

Entry into a new adaptive zone

Adaptive radiations also includes the entry of a new lineage into a new adaptive zone followed by a tremendous burst of evolutionary activities. In the convergent processes animals from their parental stock acquires some new adaptations that removes them from their ancestral adaptive zones. The new adaptive zones are

characterized by some specific features, and the speces inhabiting such areas acquire some features suitable to those conditions. This make a parental stock to converge in an adaptive zone from different zones, therefore, the phenomenon is referred o as convergent pattern of adaptive radiations.For example spiny rayed teleost fishes, characin fishes of South Africa, frogs geckonid lizards, finches and rodents etc.

Causes of adaptive radiations

Species explore new habitats for food, safety and better breeding grounds. These are the main motives for adaptive radiations. Organisms keeps on exploring those habitats where these can survive with the adaptive features. These prefer those areas which are unoccupied and with no or limited competitions.

Evolutionary significance of Adaptive radiations. We have discussed while explaining the adaptive radiations that this phenomenon helps in generation of new genera and species. Therefore, adaptive radiation plays significant role in the evolutionary processes

Occupation of new environment and Niches

Nich: The term Nich or ecological Nich was first described by Grinell in 1917 but it was clearly explainrd by Charles Elton in 1927 as, ecological nich represents functional status of an organism in its community. According to Elton Nich is an animals place in the biotic environment or community with relation to food and protection from enemies. Recently Odum has compared habitat o an organism to its address and the nich to its profession.

Mimicry and Colouration

Mimicry is the superficial but close resemblance of one organism to another or to natural objects among which it lives, that secures its concealment, protection or some other advantage.

Kinds of Mimicry :- It can be classified into three categories:-

1. Protective Mimicry- This kind of mimicry includes imitation, in which organisms mimic either some other organism or the natural object in form, color or behavior and thus protect themselves from predators. This could be of two types.

A. Concealing mimicry- Organism conceal or camouflage themselves either by altering their coloration to fit background or search a background which matches their color.

i.) Changes in color to match the background

a.) White crab, Cryptolithodes, harmonizes perfectly with white pebbles on beach that it is very difficult to recognize it. Its smooth, round shape, texture and white color almost resemble the pebbles.

ii.) Changes the place to match the background

Where an animal is not capable of changing its color it moves to seek out a suitable background for resting.

b.) The caterpillar of geometrid pine moth, *Bupalus piniarius* is always green with white longitudinal stripes because it always sits o pine needles. This protects them from the predatory birds which prey upon them.

iii.)Mimicking the living or dead objects

a.) The stick insect or walking stick exhibits close resemblance with the twigs in having slender body, attenuated limbs, sympathetic coloration and slow movement.

b.) Leaf insect *Phyllium* possesses flattened and expanded body and limbs. It is green in color and possesses irregular yellowish spots which simulate the fungus or rust grown upon a leaf.

B. Warning mimicry- It includes all those cases of protective mimicry in which the nonpoisonous and harmless organisms mimic the poisonous and harmful organisms and the palatable forms resemble and advertise to be non-palatable. This is helpful in self defence, because by imitating these are able to delude and frighten the enemy and escape themselves.

a.) Certain non-venomous coral snakes of the family Colubridae exhibit color pattern of venomous coral snakes belonging to the family Elapidae. These have alternating bands of bright red and black color edged with yellow.

b.) The non-venomous snake of genus *Heterodon* flattens its head, produces frequent hissing and strikes to advertise as if it is dangerous.

C. Crypic structures- Many animals exhibit structures, which resembles to a strong predator so that the predator of that speices escapes after watching such structures.

a.) Butterflies, Caterpillar possess black spots on their wings on body. They appear terrified eyes, when the predator watches such eyes get away of fearful appearance of such eyes-like spots.

b.) Dummy head- In some animals like lantern fly from thyland has particularly convincing the structures on the posterior end resembling antennae, black eyes and black beak are actually the appendages of the wing's tips. Such a structure threatens the predator.

c.) The snake *Cylindrophis rufus,* when faced with opponents or prey hold their head still and move the tip of their tail as to divert attention of the animal away from the genuine head.

2. Aggressive mimicry- This is exhibited by carnivorous animals, which either conceal themselves so that these are not easily recognized from their surrounding or allure the prey. These are of two types-

A. Concealing mimicry- The animals develop cryptic colors so as to blend with surrounding or mimic to its model and are not differentiated easily.

a.) Spiders resemble in shape and color to the flowers on which they live so that these are not easily distinguished from the flowers. Thus by hiding themselves among the flowers these are able to prey.

B. Alluring mimicry- In this mimicry, animal possesses some lure to attract its prey, whereas it blends itself with the surroundings. The misleaded animals fall victim and form the prey of the mimic.

a.) Certain spiders mimic the flowers of orchid and the insects lured to collect honey are devoured.

3. Pretending of death (Conscious mimicry)- Certain animals exhibit conscious imitation and on the approach of danger behave as if they are dead bodies.

a.) American opossum, *Didelphis virginiana* becomes unconscious and simulates as dead.

Occupation of new environment & Niche:-

The term "Niche" was described firstly by Grinnell (1917). According to Odum, Grinnell thought of the niche mostly in terms of the microhabitat or what we now

call the spatial niche. According to Charles Elton (1927) "functional status of an organism in its community".

Odum compared the habitat of an organism to its "address" and the niche to its "profession". The habitat defines where an organism lives, its niche describes its mode of life in that habitat. A habitat obviously can contain more tha one niche.

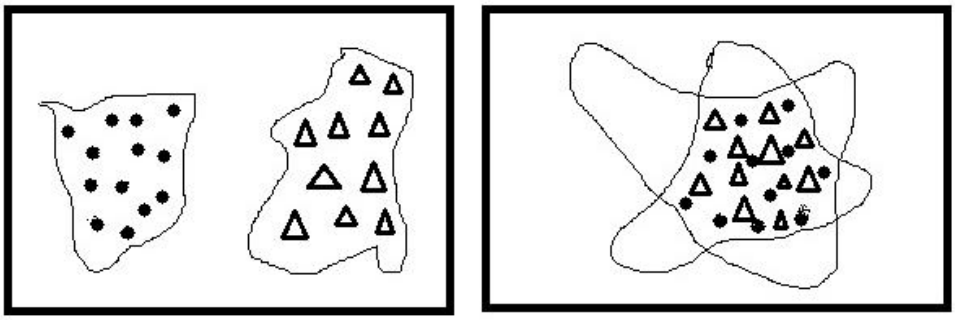
Eg. Zebra and African lion both roam the veldt or tropical grassland but they don't have the same niche. The zebra is herbivorous, that lives of the veldt vegetation. Where the lion is carnivorous that feeds on veldt herbivores, including zebra.

Aspects- Three aspects of ecological niche:-

1. Spatial or habitat niche
2. Trophic niche
3. Multidimensional niche

1. Spatial niche- This concept represent the ultimate distribution. In a particular habitat share by the several species. Each of species may be confined to its own microhabitat.
   Eg. O'Neill (1967) has given example of microhabitat segregation in seven species of millipedes. All of the species live in the same general habitat, the forest floor of maple-oak forest and all belong to the same basic trophic level, that is, they are detritus feeders. But each of the seven species predominates in a different microhabitat.
2. Trophic niche- Sometimes two species may live in same habitat but they occupy different trophic niche because different in food habits.
   Eg. Two aquatic birds *Notonecta* and *Corixa* live in the same pond but occupy different trophic niches.
3. Multidimensional or hypervolume niche- Hutchinson made a distinction between the "fundamental niche" the maximum abstractly inhabited hypervolume when the species does not face competition with others and the "realized niche".
   Eg. Figure represents schematically the hypervolume concept of the ecological niche. The background dots represent environmental factor. The irregular polygon enclose set of factor that are operationally significant for the species population. In "A", two species occupy nonoverlapping niches, while in "B", niche of two species overlap to such extent that serve

competition for shared resources, that may result in elimination of one species or a divergence of niches as indicated by arrows.



A B

Unit -7

Biostatistics

Structure of the Unit

7.1 Objectives

After going through this unit you will be able to understand:

- Introduction, Definition, Characteristics of Biostatistics
- Importance, Application, Scope of Biostatistics
- Statistical Terms & Symbols

- Methods of Representation of Statistical Data
- Tabular, Graphical & Diagrammatic Representation of Data

7.2 Introduction

Statistics has wide application in almost all sciences – social as well as physical such as biology, medicine, agriculture, veterinary, economics, psychology, ethology, business management etc. It plays a major role in bioscience because data of bioscience are of a variable nature. It is very difficult to draw a concrete conclusion from biological experiments because of inherent differences between two individuals. Homozygous twins are even not exactly same in physiology and behaviour.

Most of the happenings in life depends upon counting or measurements. Plants and animals obtained by any hybridization experiment agree with Mendel's law or not, can only by concluded by statistical test i.e. χ^2 test or low blood pressure has no meaning unless it is expressed in numbers.

Blood pressure, pulse rate, Hb%, rate of reproduction, rate of transpiration, action of a drug on individual or a group etc. varies not only from person to person but also from group to group. The extent of this variability in a character is by way of chance, i.e. biological or normal is revealed by statistical methods.

7.2.1 Definition

Biostatistics is used when tools of statistics are applied to the data that is derived from life science.

7.2.2 Characteristics of Bio-Statistics

(1) Statistics is the aggregate of facts.
(2) Statistics is numerically expressed.
(3) Statistics is usually affected by multiplicity of causes and not by single cause.
(4) Statistics must be related to some field of inquiry.
(5) Statistics should be capable of being related to each other, so that some cause and effect relationship can be established.
(6) The reasonable standard of accuracy should be maintained in statistics.

7.3 Importance of Bio-statistics

(1) Statistics help in presenting large quantity of data in a simple and grouped form.

(2) It gives the methods of comparison of data.

(3) In enlarges individual mind.

(4) It helps in finding the conditions of relationship between the variables.

(5) It proves useful in almost every sphere of human activities.

7.4 Application and Uses of Biostatistics

Biostatistics is applied and used in different branches of bioscience.

7.4.1 (I) In Physiology and Anatomy

(1) To define what is normal or healthy in a population and to find limits of normality in variables.

(2) To find the difference between the mean and proportion of normal at two places or in different periods.

(3) To find out correlation between two variables X and Y such as height and weight.

7.4.2 (II) In Pharmacology

(1) To know that action of drug-a drug given to animals and humans to observe the changes produced are due to the drug or by chance.

(2) To compare that action of two different drugs or two successive dosages of the same drug.

(3) To find out the relative potency of a new drug with respect to a standard drug.

7.4.3 (III) In Medicine

(1) To compare the efficacy of a particular drug. For this, the percentage of cured and died in the experiment and control groups is done.

(2) To find out an association between two attributes such as cancer and smoking.

(3) To identify signs and symptoms of a disease of syndrome. Cough and typhoid is found by chance and fever is found in almost every case.

7.4.4 (IV) In community medicine and Public health

(1) To test usefulness of sear and vaccines in the field-the percentage of attacks or deaths among the vaccinated subjects is compared with that

among the unvaccinated ones to find whether the difference observed is statistically significant.

(2) In epidemiological studies – the role of causative factors is statistically tested.

(3) In public health, the measures adopted are evaluated.

7.4.5 (V) In Genetics

Biostatistics is used in studying the genetics. Mendel's law's of inheritance is tested by x^2 test. Hardy Weinberg law is tested by biostatistical methods.

7.5 Scope of Biostatistics

Use of statistical methods are constantly increasing in biological sciences. The development of biological theories are closely associated with statistical methods. Heredity, one of the recent branches of biology is mainly based on biostatistics. Therefore, for the students of biology, the knowledge of biostatistics is a must.

7.6 Statistical Terms and Symbols

Basic statistical terms : Population, Sample, Data. Observation. Parameter and Statistic. Characteristic. Attribute/Variable. Statistical error. Subscript and summations. Functions of statistics. Array. Class interval, Class size, Class mark. Freq. distribution.

Symbols : $\overline{X}$, μ, x, f, δ, σ, 't', χ^2, Z, γ, ρ, S

7.7 Terms and Symbols : Definition

Term is a word used to explain a particular identity. **Symbol** is a mark or sign with a particular meaning. The use of terms and symbols allow statisticians to deal with general expression and general results. By doing so, considerable saving in space and time takes place.

7.8 Statistical Terms

7.8.1 Population

Population may be defined as **"an entire group of organisms of one species, occupying a definite area or study elements-persons, things or measurements having some common fundamental characteristics". in statistics population is a well defined group which is being studied**. Suppose one has to study the incidence of helminth infection in rabbits. For this purpose 100 rabbits are

collected randomly and brought in the laboratory. Here 100 rabbits are population for this experiment and result will represent the universal population of rabbit. In other words in statistics, population always means the total number of individual of individual observations from which inferences are to be made at a particular time.

7.8.2 Sample :

The selected part of a population is population is known as sample. For example, all patients of AIDS of the world represents a population, whereas individual observations on 10 or 20 or 30 (any convenient number) patients from the population refer to a sample.

7.8.3 Data:

Data is a collection of observations expressed in numerical figures. Data is always in collective sense and never be used singular. The data in statistics are generally based on individual observations. The Hb% of 10 patients suffering from Kalaazar was measured as 10.2, 9.6, 8.8, 10.7, 9.9, 10.8, 11.3, 9.5, 8.9, 8.8, mg/100 ml. Here 10.2, 9.6.....8.8 mg/100 ml. are a set of values for an event i.e. Hb% and is called data.

7.9 Primary and Secondary Data

7.9.1 Primary data

The data which are collected directly by an investigator for the first time for a specific purpose are called as **primary data.** These are raw data or data in original nature, and directly collected from population. The collection of primary data may be made through either by complete enumeration or sampling survey methods.

7.9.2 Secondary data

If data are collected and used by any other agency at least once then such data are termed as secondary data.

Note : In scientific researches only primary data are used.

7.10 Qualitative and quantitative data

(a) ***Qualitative*** : According to quality attributes the data is called qualitative. For example, lions of **Gir sanctuary** of Gujarat State are to be classified in respect to one attribute say sex, in two groups, one is of **male** and the other is of **female.**

(b) ***Quantitative*** : According to magnitude the data is called quantitative. For example, chickens of a poultry farm may be classified on the basis of their growth rate. Quantitative data may also be classified into two types :

(i) *Continuous.* Values of variate **do not exhibit any breaks or jumps.** For example the increasing length and weight of a child.

(ii) *Discrete*. Values of variate **vary by infinite jumps.** For example the oxygen consumption of rat **(*Rattus rattus*)** of different weight groups was measured as 500 cc/h/100 ml, 600cc/h/100 ml, 620 cc/h/100 ml, 680 cc/h/100 ml and so on.

7.11 Observation

Measurement of an event is called observation. For instance blood pressure, temperature of body, oxygen consumption etc. are events whereas, 160 mm and 80 mm. (upper and lower pressure), 106^{o}F, 65 kg/hour/100 ml are their respective observations. The source that gives observations such as object, person etc. are called observational units. In biostatistics statistics the term individuals or subjects is used for observational units.

7.12 Parameter and Statistics

A value calculated from a defined population such as Mean (μ for population), Standard deviation (σ), Standard error of difference of mean a ($\overline{X}_1$- $\overline{X}_2$) etc. are called a **parameter.** It is a constant value because it covers all members of the population. Familiar examples are mean height, birth rate (fecundity) and mortality rate etc. of any one species of animals or plants. The quantity calculated to represent a **character of population** is known as **parameter** whereas quantity calculated to present the **character of the sample** is called statistics. The former is the constant quantity whereas the latter is variable. In other words, we can say that the numerical quantities which characterise a population (in respect of any variable) are called parameter. For example; if the variable is in length and the measurement of length is taken for the entire population, the mean length can be regarded as parameter. But the mean length of the sample ($\overline{X}$) can be regarded as statistic. Standard deviation of sample (s) and Standard error of difference of mean [s($\overline{X}_1$ - $\overline{X}_2$)] of sample are statistic.

In biological experiments values calculated from a large sample is often applied to population and may be a valid estimate of population. Therefore, in biostatistics, though not desirable, parameter and statistics are often used as synonyms.

7.13 Characteristic

The term 'characteristic' means a quality possessed by an individual i.e., object, item of population. Height, weight, age, Hb%, VO_2 etc. are characteristics.

In statistics, characteristics are of two types.

(1) Non-measurable 'characteristics' is called **Attributes.**

(2) Measurable 'characteristics' is called **Variables.**

7.13.1 Attributes : Attributes are the non-measurable characteristics which not be numerically expressed in terms of unit. These are qualitative object. For example : sex, illiteracy etc.

7.13.2 Variables. Variables are the measurable characteristics which can be numerically expressed in terms of some unit. These are quantities which are capable of being measured by quantitative methods directly. An individual observation of any variable is known as variate. If we measure the height of some individuals of a population and obtain some values, the obtained values is variable. For example height and length in cm, weight in g, Hb in %, oxygen consumption in VO_2/100 ml etc. of individuals.

A variable is a symbol, such as *X, Y, Z* etc., that may take any value in some specified set of numbers. Whatever the value of the variable actually observed is the actual value and is denoted by the $X_1, X_2, X_3, ..., X_n$ or $Y_1, Y_2, Y_3, ..., Y_n$ or $Z_1, Z_2, Z_3, ..., Z_n$. The possible values of a variable are those values that the variable may possibly take. For example, the “weight of fishes in a pond" may take any value between $X_1, ...X_n$ *g.* Both X_1 and X_n are inclusive.

Depending on the break or continuity, variables are of two types :

(a) Discrete variable is one which cannot take all the values and there is a gap between one value and the other. For example, the number of zooplanktons in 5 ponds were obtained as 98, 305, 387, 105 and 208. So, in this case number of zooplankton is the discrete variable. Number of persons in a family, no. of books in. a library are discrete variable, because they cannot take functional value. One cannot say that there are 3.5 persons in my family or there are

500.6 books in a library. The discrete variable may take any integer value from 0 to $\propto$.

(b) ***Continuous variable is*** the one which can take any values and there is no interval. For example, the weight and height of human being is a continuous variable because it may take any value. Height of students in a class may be 120 cm, 120.2 cm, 120.5 cm, 120.7 cm, 120.9 cm and so on. Measurement of Hb%, VO_2 consumption etc. present continuous variable. Generally speaking 2 discrete variable take integer value while continuous variables take fractional values.

7.14 Statistical error

In statistical terminology, the word 'error' is used in special sense. Error shows the extent to which the observed value of a quantity exceeds the true value. Error = Observed value - True value.

7.14 Subscripts and Summations

Subscripts and suffixes are used to distinguish between values of a variable. If the weight of one species of fish be symbolized by X then the weight values of the different fishes of the same species may be denoted by X_1, X_2, X_3 ....., X_n, where X_i *(i = 1, 2, 3, ..., n)* denotes the weight of the i^{th} class.

Usually it becomes essential to add (sum up) a series of values of variable. The total weight of the obtained fishes from a pond is then $X_1 + X_2 + X_3 + ... + X_n$. This is a very common way of writing down the total. It may be simplified by writing X_i where the ($\square\square$ Greek letter symbolizes the operation of summation and tells us to sum up from $i = 1$ (which is the first value) to $i = n$ (which is the last value) of X. We shall still more simplify the notation by writing the sum as $\square X$ with the implication that the limits over which the summation is to be taken are obvious.

While dealing with $\square$ sign, one has to be particular about the difference between $\square X^2$ and $(\square X)^2$. The first symbol adds up the squared values of X whereas the second symbol squares the summed values of X. For example, if X takes the values 1, 2, 3, 4 and 5 then $\square X^2 = 1 + 4 + 9 + 16 + 25 = 51$ while $(\square X)^2 = (1 + 2 + 3 + 4 + 5)^2 = (15)^2 = 225$.

7.16 Array : The presentation of data in ascending or descending order of magnitude is called array.

7.17 Class interval : Each group into which the raw data is condensed is called a class interval. Each class is bounded by upper and lower figures called class limits.

7.18 Class Size : The difference between the true upper limit and true lower limit of a class is called the size of the class interval.

7.19 Class mark :

$$\text{Class mark of a class} = \frac{\text{upper class limit} + \text{lower class limit}}{2}$$

7.20 Frequency : The number of times a value of the variable occurs is called the frequency.

7.21 Frequency distribution : The law data or ungrouped data unless they are arranged in a systematic way, do not give clear idea of the subject-matter. Biostatisticians, therefore, arrange the raw data in ascending or descending order of their magnitudes. This arrangements is known as an array. Arrayed data is kept in such a manner that each variable exhibit its repetition number i.e. frequency of variable. This is called frequency distribution, as detailed in chapter 4.

7.22 Range : Range of the arrayed data is the difference between the maximum and minimum observation in a variable.

7.23 Cumulative frequency : The cumulative frequency corresponding to any value of the variable (or a class) is the sum of all the frequencies.

7.24 Symbols

Following important symbols are used is biostatistics

1.	Summation	Σ (It is capital Greek letter pronounces as sigma)
2.	(i) Mean or average	$\bar{X}$ (read as X bar) Obtained from sample
	(ii) Mean of population	μ

3.	Observed number	O
4.	Expected number	E
5.	Degree of freedom	Df
6.	Number of groups or classes	K
7.	Probability	P
8.	Deviation	χ
9.	Frequency	F
10.	Mode	M_o
11.	Median	M
12.	Geometric mean	GM
13.	Mean deviation	δ
14.	Standard deviation	$\square$ (of hypothetical population) S (of observed sample or S.d.)
15.	Standard Error of Mean	SE_M
16.	Standard Error of standard deviation	SE_σ *or* SE_s
17.	Student's test or 't' ratio	*'t'* test
18.	Chi-square test	χ^2
19.	Variance	σ^2 or S^2
20.	Width of class interval	I
21.	Correlation coefficient	R
22.	Number of observation	N or n
23.	The no. of Standard deviation from the	Z

	mean	
24.	Pearson's correlation coefficient	γ
25.	Spearman's correlation coefficient	ρ (Pronounced as Rho)

7.25 Methods of Representation of Statistical Data

Statistical data are presented in three processes :

(1) Tabular presentation
(2) Graphical presentation
(3) Diagrammatic presentation

7.25.1 Tabular presentation

(1) The logical and systematic presentation of numerical data in rows and columns designed to simplify the presentation and facilitate comparison is termed as tabulation.
(2) Tabulation is thus a form of presenting quantitative data in condensed and concise form so that the numerical figures are capable of easy and quick perception by the mind.

Definition

Tabulation may be defined as the logical and systematic presentation of numerical data in rows and columns designed to simplify the presentation and facilitate comparison.

Advantages of tabulation :

(1) It enables the significance of data readily understood and leaves a lasting impression than taxual impression.
(2) It facilitates quick comparison o statistical data shown between rows and columns.
(3) A tabular arrangement of data makes the errors and omissions readily detactable.
(4) Repetition of explanatory terms and phrases can be avoided.
(5) Concise tabular presentation of data clearly reveals the characteristics of data.

Types of table

All types of tables may be reduced to two major types :

(1) Simple tabulation

(2) Complex tabulation

1. **Simple tabulation** : It contains data in respect of only one characteristic. Hence, this type of table is also known as one-way table. It has got two factors placed in relation to each other. Following simple table 1 is made when length of 57 Papaya plants of a garden was measured in cm.

Table 1.

Length of plant	1-5	6-10	11-15	16-20	21-25	26-30
No. of plants	2	5	10	11	9	20

2. **Complex tabulation** : In a complex table, two or more characteristics are shown. Following complex table 6 contains two co-ordinate group. i.e. male and female.

Table-2.

Length	**No. of Papaiya plants**	
	Male	**Female**

If the members of a co-ordinate group is further classified on the basis of other attributes then the table is called complex multiple table. Following complex table 3 contains two co-ordinate group i.e. male and female which is further classified : Papaya plants are classified into two groups, according to sex and further classified on the basis of infection, then it will be a case of complex multiple table.

Table 3.

Length	**No. of Papaya plants**			
	Male		**Female**	
	Infected	**Not infected**	**Infected**	**Not infected**

Parts of a table

Title

(1) This is a brief description of contents of the table along with time, place and category of items if required.

(2) The title should be clear and precise.

(3) The title should be at the top of the table.

Stub

(1) The extreme left part of the table where description of the rows are shown is called stub.

(2) This must be precise and clear.

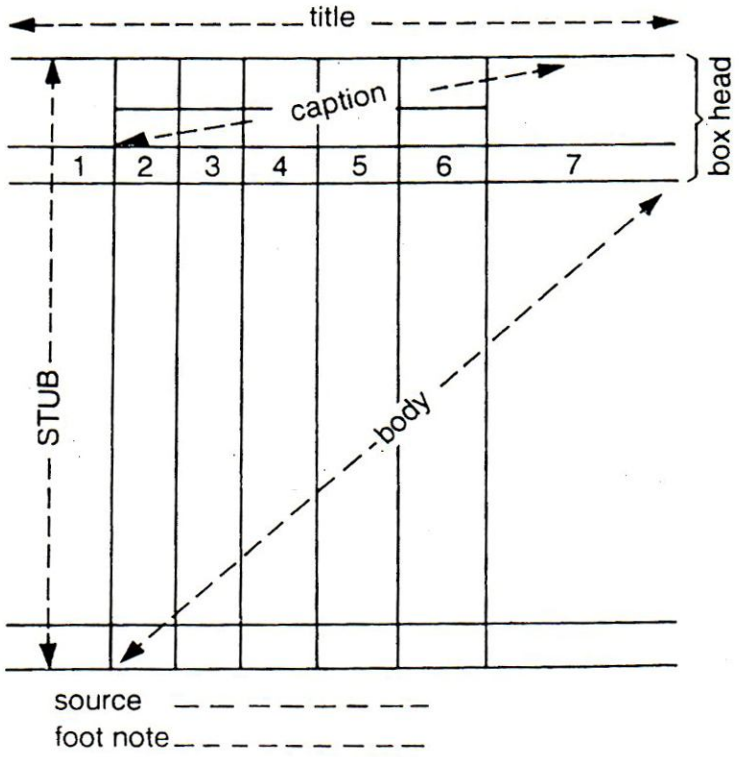


Fig. 1. Different parts of table

Caption and boxhead

(1) The upper part of the table which shows the description of columns and subcolumns is called **caption.**

(2) The whole of the upper part including caption, units of measurement and column number if any called **boxhead**.

Body

(1) It is the main part of the table except the title stub and captions.

(2) This contains numerical information which are arranged in the table according to the description of the rows and columns given the stub and caption.

Source and foot note

(1) It is customary that source of data from which information has been arrived should be given at the end of the table.

(2) Foot note is the part below the body where the source of data and any explanation are shown.

Statistical Table

Statistical table is a systematic arrangement of quantitative data under appropriate heads in rows and columns. After the data have been collected, they should be tabulated that it put the form of a table, so that whole information can be had at a glance.

Essential features of a good table

(1) A table must have a title giving clear and precise idea about the contents of the table.

(2) Units of measurements adopted in a table must be shown clearly in the top of the column.

(3) An investigator must prepare the table proportionate in length and breadth.

(4) For comparison, column of relevant figures must be kept as close as possible.

(5) Distinction is preferred in columns and sub columns. It can be made by distinct ruling (viz. double ruling, single ruling etc).

(6) Totals of columns may be shows in the bottom of the table. In cases where row totals are useful, they should also be shown.

(7) Table must contain necessary details.

(8) Source of information must be disclosed at the end of the table.

(9) Any ambiguous or confusing entry in the table should bear a special note at the end of the table for experiment.

(10) The arrangement of items in the table should have a logical sequence.

7.25.2 Graphic Representation of Data

Introduction

Graphic representation of data is widely used to present data in a simple, clear and effective manner. A graph is a visual form of presentation of data where comparisons can be made between two or more phenomena. It is rightly said the

wandering of a line is more powerful in its effect in mind than a tabulated statement.

Advantages of graphical representation

(1) It is easy to understand.
(2) The data can be presented in a more attractive form.
(3) It shows the trend and tendency of values of the variable.
(4) It exhibit clear cut relationship between two or more sets of figures.
(5) It has the universal applicability and is helpful in assimilating the data readily and quickly.

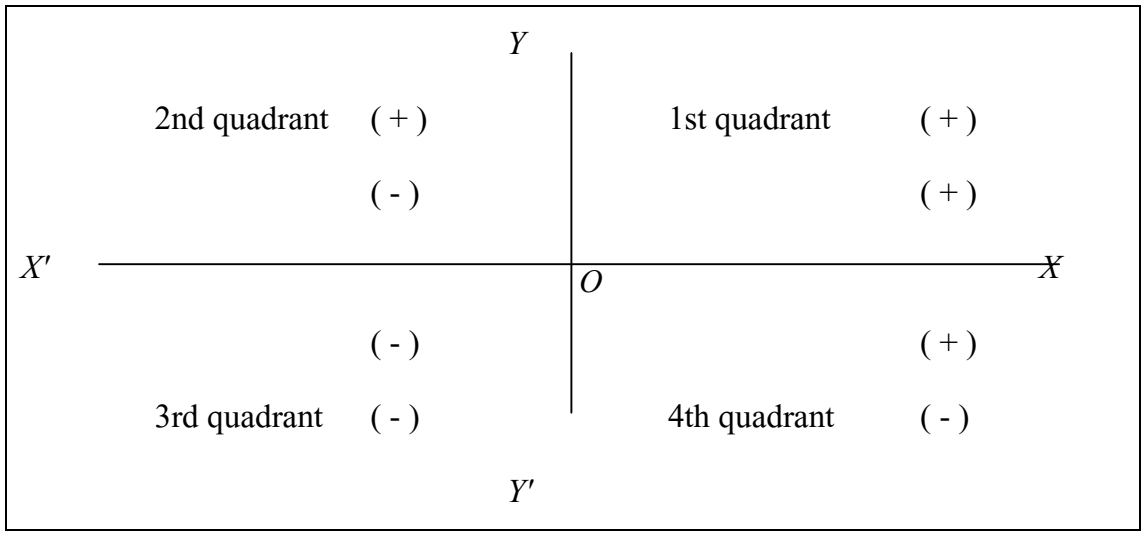
Disadvantages of graphical representation

(1) It does not show details or all the facts.
(2) Graphical representation can reveal only the approximate position.
(3) It takes a lot of time to prepare graph.

Methods of Preparation of Graph

Some basic knowledge is essential to prepare frequency graph. It is prepared with the help of two lines. The horizontal line is called **abscissa** i.e., X-axis representing independent variable and the vertical line called **ordinate** i.e., *Y*-axis representing dependent variable. The meeting point of *X* and *Y*-axis is called zero (o) or origin point.

The right part of ***'X'* axis** from the zero point (0) is positive (+) and left is negative (-). Likewise the upper part of **'*Y*' axis** from zero point is positive while the lower part is negative. *'X'* and *'Y'* axis intersect each other at '0' point and graph is divided into 4 parts. Each part is called quadrant. Upper right part is called first quadrant where *'X'* and *'Y'* both axis are positive. Upper left part is called second quadrant. Here *'X'* axis is negative (−) and *'Y'* axis is positive (+). The lower left part is called third quadrant where both *'X'* and *'Y'* axis is negative. The lower right part is known as fourth quadrant where *'X'* axis is positive (+) and *'Y'* axis is negative (−). Mostly first quadrant is used for graphical representation of statistical data, where both axes are positive.



Two axes *'X'* and *'Y'* intersecting each other on *'O'* point producing 4 quadrant.

Units of representation

Appropriate unit bar line is required to present the statistical data in graph. Suppose, one has to show large numbers such as 500, 1000, 2000 and above on graph, then he has to consider 1 cm long line on graph as 500 units bar line. Now 1 cm is divided 5 times to represent 500 unit bar line. 1 cm is denoted as 5 ; 2 cm as 10 ; 3 cm as 15 and so on. For number 750 a point in between 5 and 10 is mentioned. The same method can be adopted to represent any number.

Types of Graphic Representation

Grouped data can be represented graphically in following ways :

(1) Histogram
(2) Frequency polygon
(3) Frequency curve
(4) Relative frequency map
(5) Cumulative frequency curve or ogive and
(6) Scatter or dot diagram

1. **Histogram** : This graph is used for continuous frequency distribution. The width of the class interval are marked along with the *X*-axis, or abscissa. On these lengths, rectangles of areas proportional to the frequencies of the respective class intervals are erected.

If the class intervals are of equal lengths, then the heights of the rectangles are proportional to the corresponding frequencies and for unequal class interval, the heights of the rectangles are proportional to the ratios of the frequencies to the width of the corresponding class.

Following grouped data is obtained in an observation of "rate of reproduction" of 50 fishes of a species. Make a histogram, frequency polygon and frequency curve with the help of data provided.

Class intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	3	4	7	8	9	9	2	6	2

Solution: *OX*-axis 1 cm = 10 class interval denoting the rate of reproduction.
OY-axis 1 cm = 1 frequency the frequency of rate of reproduction.
The frequency of 1st class interval 0-10 is 3 which is being represented by 3 cm = 30 small squares on *OY* axis because 10 small squares = 1 frequency. In the same fashion rectangle for each interval and frequency is plotted and finally a histogram of the above frequency distribution is shown in figure 2.

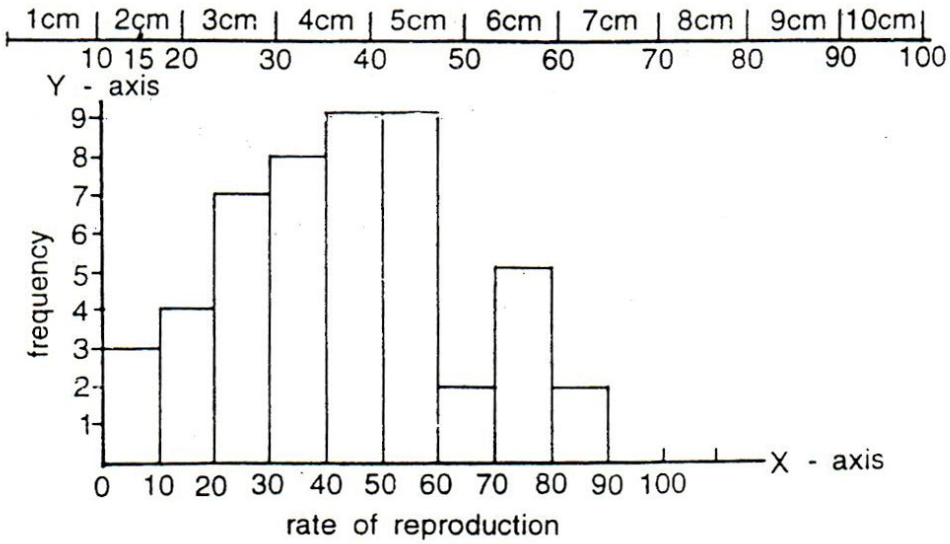


Fig. 2 : Histogram showing rate of reproduction and their frequency of 50 fishes of a species.

2. **Frequency polygon and 3. Frequency Curve** : The values of the variable for an ungrouped data are taken as the abscissae and their frequencies are taken as the ordinates. For a grouped data, the mid-points of the class intervals are taken as the abscissae. Then a **frequency polygon** is obtained by joining the plotted points by the straight lines. If the class intervals are of small length, then the plotted points are joined by free hand. The curve so obtained is known as frequency curve.

Frequency ploygon for equal class intervals can be obtained by joining all the mid-points of the upper sides of rectangles of the histogram, by straight lines. It gives a shape of polygon i.e., a figure with many angles. Figure 3 is plotted with the help of same above data.

4. **Relative frequency map** can be plotted with the help of relative frequency table. Relative frequency is the proportion of all observations.

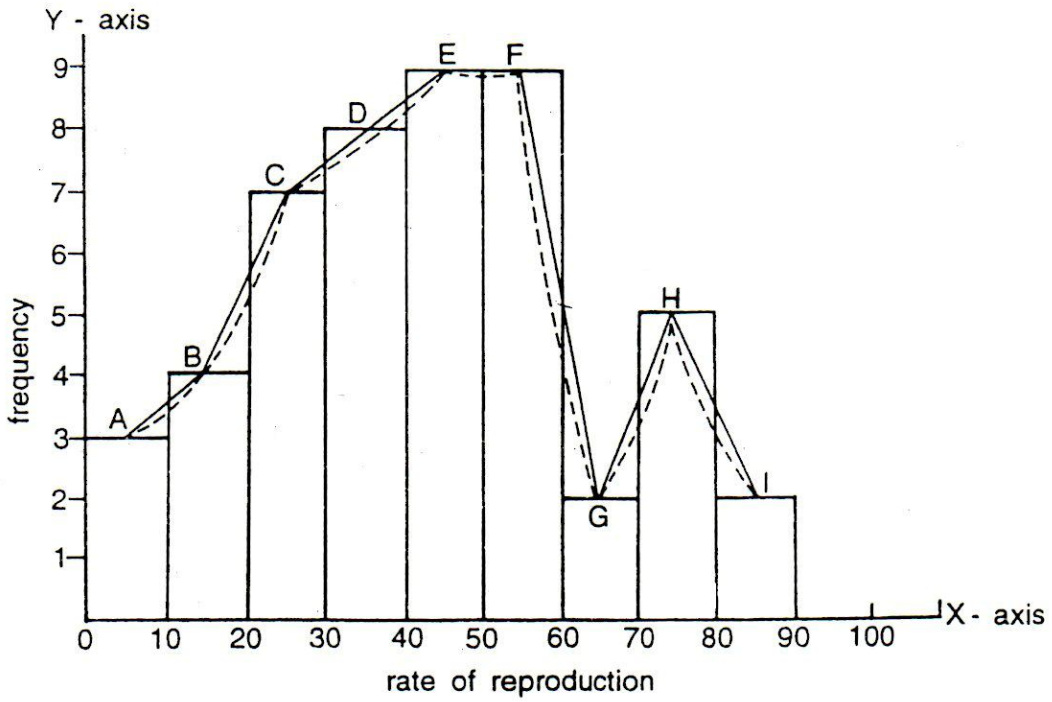


Fig. 3. Frequency polygon and frequency curve showing rate of reproduction and their frequency of 50 fishes of a species of fish.

Unbroken lines joining mid points A, B, C, D, E, F, G, H and I of rectangle show the frequency polygon.

Broken lines joining mid points A to I of rectangle represent the frequency curve. It is drawn by joining the mid-points of class intervals of upper horizontal lines of rectangle by free hand.

determined by dividing the number of observation in category by the total number of all observations. The relative frequency is generally calculated after the frequency distribution is obtained. The table 4 suggests the method of calculation of relative frequencies.

Table 4.

Age of albino rats in months X	No. of albino rats in the category (frequency) f	Relative frequency Rf.
1-3	7	7÷43=0.16
4-6	8	8÷43=0.19
7-9	12	12÷43=0.28
10-12	10	10÷43=0.23
13-15	6	6÷43=0.14
	$\Sigma f = 43$	Σ Relative frequency = 1

Relative frequency (Rf) is calculated with the help of following formula : Rf $= f / \Sigma f$

Relative frequency map (Fig. 4) have been plotted in a graph whose relative frequency are shown in the vertical axis and age groups in the horizontal axis.

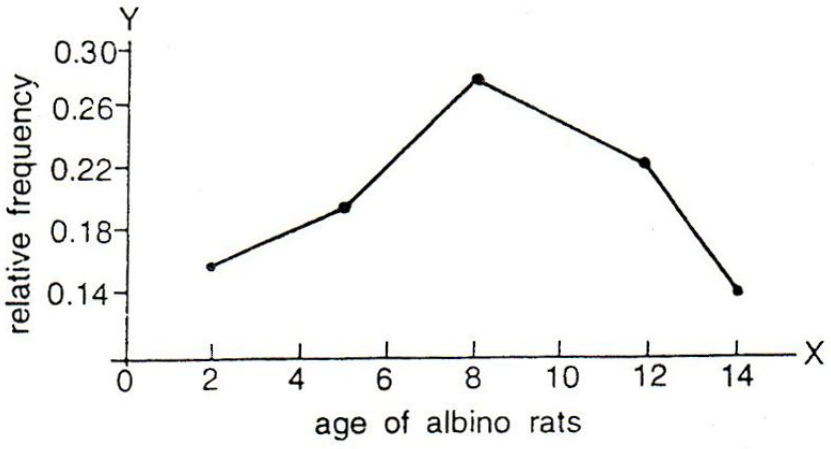


Fig. 4. Plotted values of table 7 showing relative frequency.

5**. Scatter or dot diagram** is prepared after cross tabulation in which frequencies of at least two variables have been cross classified. This graphic presentation shows the nature of correlation between two variable characters X and Y.

Worked example : Values of *X* variable and *Y* variable of 8 groups of fishes is given below to draw scatter or dot diagram :

X	13.9	15.7	15.8	17.5	18.1	19.9	22	23.8
Y	5	5.9	6.4	7.3	7.8	8.1	8.7	8.9

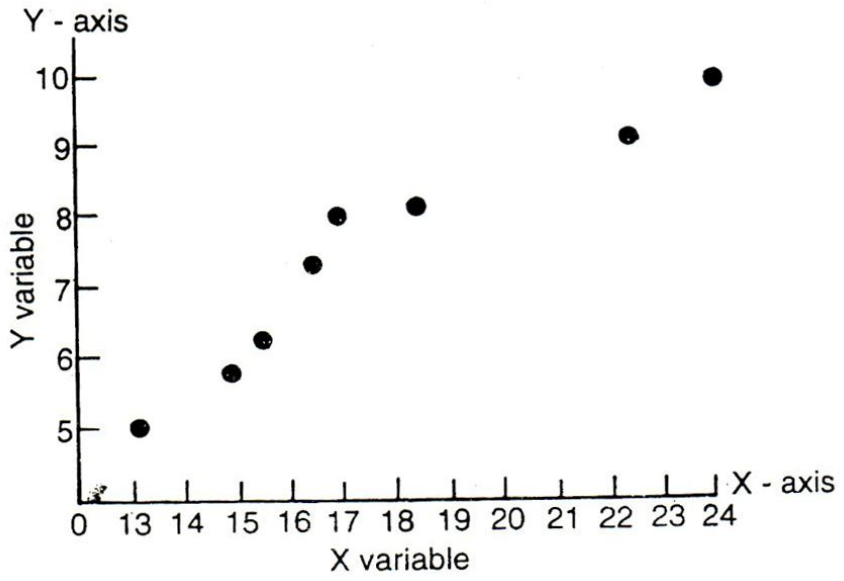


Fig. 5. Scatter diagram showing that no line can be drawn from the data.

Solution : The characters are read on the base and vertical axes and the perpendiculars drawn from these readings meet to give one scatter point. When two axes are at right angles to each other, they are called orthogonal axes.

X variable and *Y* variables are taken on the *X* and *Y* axes respectively.

Significance of graphic representation

Graphical representation of frequency distribution is important because :

(1) It is simplest method of presentation of data.
(2) They give clear cut and attractive view.
(3) They make comparison of variables easy.
(4) They are helpful in ascertaining certain statistical measures.
(5) They save time and energy.

Limitations of graphic representation

(1) A graph simply shows tendency and fluctuations, and not the actual values.
(2) Complete accuracy is not possible on a graph.
(3) Graphs cannot be quoted in support of some statement.
(4) Only a few characteristics can be depicted on a graph. However, in the case of many figures, it is difficult to follow the graph.

7.25.3 Diagrammatic Representation of Data

Introduction

Besides groupwise classification, tabulation and graphic representation, the data can also be presented by diagrams. Diagrams help biostatisticians to visualize the meaning of a numerical complex at a glance.

Types of diagrams

Important types of diagrams used for presentation of data are given below :

(1) Line diagrams
(2) Bar diagrams
(3) Pie-diagrams or pie chart

[I] Line diagrams

This is the simplest type of diagram. For diagrammatic representation of data, the frequencies of the discrete variable can be presented by a line diagram. The variable is taken on the X-axis, and the frequencies of the observation on the Y-axis. The straight lines are drawn whose lengths are proportional to the frequencies.

Worked example : The frequency distribution of a discrete variable (Rate of reproduction of 50 fishes) is given in the following table 5.

Table 5.

Rate of reproduction	10	20	30	40	50	60	70	80	90
Frequency	3	4	7	8	9	9	2	6	2

The line diagram is given in figure 8 of the data presented in above table 5.

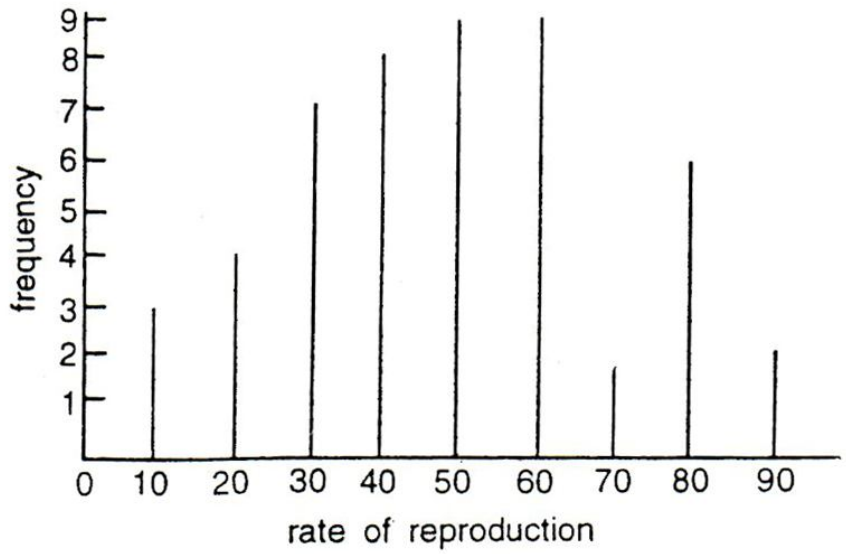


Fig. 6. Line diagram.

[II] Bar diagram

Bar diagrams are one dimensional diagrams because the length of the bar is important, and not the width. In this case the rectangular bars of equal width is drawn.

The following points should be taken into consideration while constructing a bar diagram. (i) They may be in the shape of horizontal or vertical bars. (ii) The width of the bars should be uniform throughout the diagram. (iii) The gap between one and the other bar should be throughout uniform.

Bar diagrams are of four types . (1) Simple bar diagrams (2) Divided bar diagrams (3) Percentage bar diagrams (4) Multiple bar diagrams.

1. **Simple bar diagram**. A simple bar diagram is used to represent only one variable. As one bar represents only one figure, there are as many bars as the number of figures. For example a simple bar diagram (Fig. 7) is drawn taking data of following example.

Worked example : Oxygen consumption is cc/kg/h in different months of a year in a species of fish was obtained as below. Draw a simple bar diagram.

Months	J-98	F-98	M-98	A-98	M-98	J-98	J-98	A-98	S-98	O-98	N-98	D-98	J-99
O_2	67	74	84	85	100	105	95	90	90	78	74	64	62

consumed

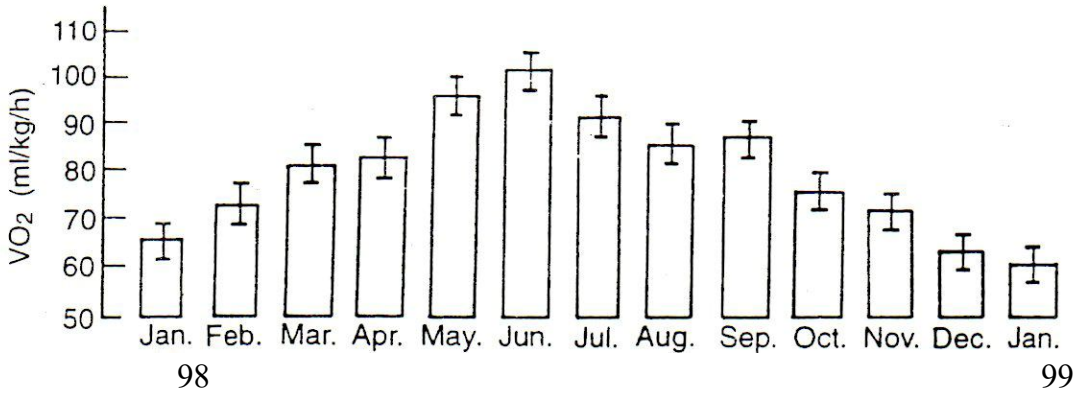


Fig. 7. Simple bar diagram showing oxygen consumption in a species of fish recorded in different months of a year.

2. **Divided bar diagram** : The frequency is divided into different components and such a diagrammatic representation is called a divided bar diagram. Suppose we have to show the average production of four species of fishes in different years, the data can be represented by divided bar diagram. Each bar then would be divided into four parts and each part would represent the mean production of each fish species (Fig. 8).

Average catch in metric tonnes of ***Wallago, Catla, Cirhinna*** and ***Clarius*** for the year 1993-94, 1994-95, 1995-96 and 1996-97 in India was as follows (Hypothetical data).

Years	**Wallago**	**Catla**	**Cirhinna**	**Clarius**	**Total**
1993-94	1383	634	513	400	2930
1994-95	2021	1383	521	313	4238
1995-96	1914	1413	551	900	4578
1996-97	2664	1636	424	265	4989

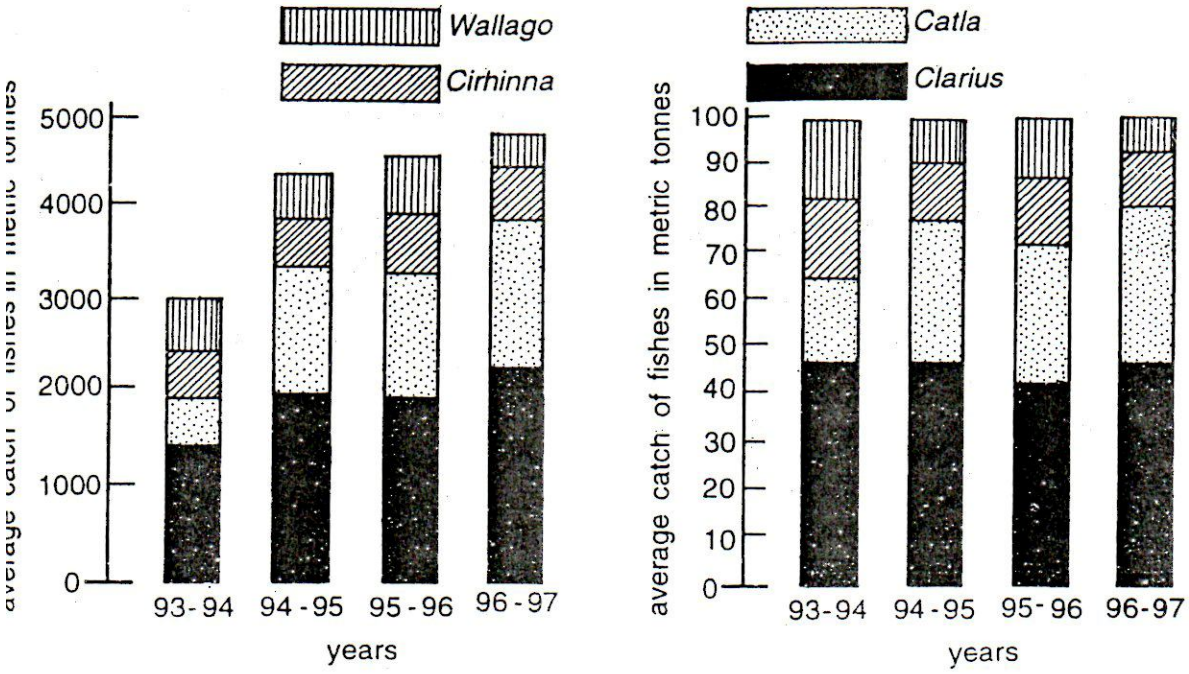


Fig. 8. Divided bar diagram representing the production of fishes of four species in four different years.

Fig. 9. Percentage divided bar diagram representing the data of fish production of 4 species in 4 different years.

3. **Percentage bar diagram** : The length of bars is kept equal to 100 and the divisions of the bar correspond to the percentage of different components. Each component of the bar diagram indicates the average catch of fishes. Above percentage divided bar diagram (Fig. 9) is drawn to represent the above data.

4. **Multiple bar diagram** : When a comparison between two or more related

Months	J-95	F-95	M-95	A-95	M-95	J-95	J-95	A-95	A-95	O-95	N-95	D-95	J-96
RBCs (Lac/mm^3)	2.01	2.01	2.08	2.12	2.25	2.46	2.27	1.87	1.91	2.30	2.19	2.12	2.04
Hb% (mg/100ml)	8.5	8.6	8.8	9.1	11.7	12.6	11.8	9.7	9.6	12.2	11.8	10.9	8.6

PCV (%)	14.1	14.1	14.1	14.4	16.6	19.6	26.2	24.9	14.4	14.5	25.6	24.2	14.0

variable has to be made, then multiple bar diagrams are preferred. The technique of simple bar diagrams can be extended to represent two or more sets of interrelated data in a diagram.

Worked example : Value of three haematological parameters viz. RBCs count, Hb% and PCV of a species of fish was studied for 13 months (Between Jan 95 to Jan 96). Data obtained is given below to draw a multiple bar diagram.

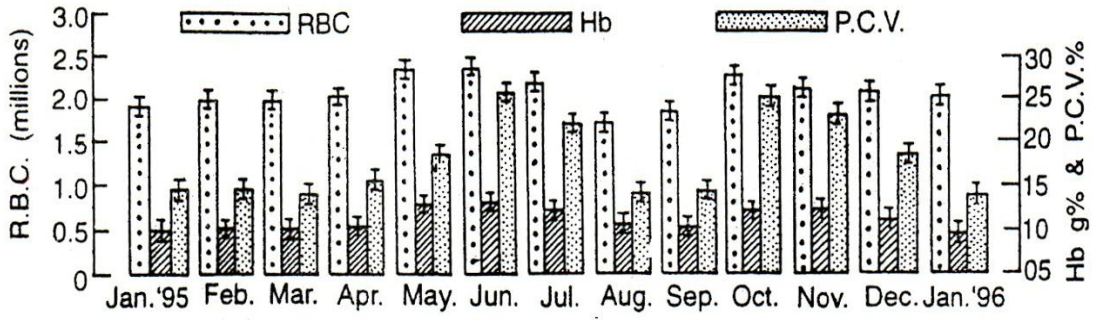


Fig. 10. Multiple bar diagram showing different blood parameters such as RBCs count, Hb%, PCV (%) in a species of fish for 13 months.

5. **Proportional bar diagram**. Here bar represents the proportion of variable. In a medical survey "monthly distribution of new and repeat patient in a hospital during different months of a year (1999), following results were obtained.

Table 6.

Months	J-99	F-99	M-99	A-99	M-99	J-99	J-99	A-99	S-99	O-99	N-99	D-99
No. of new causes	2658	2052	2179	1980	1714	2153	2203	2123	2642	3055	3869	2557

No. of repeat cases	1114	1470	1610	1351	1175	1434	1494	1536	1772	1965	2360	1894
Total	3772	3522	3789	3331	2889	3587	3697	3659	4414	5020	6229	4451

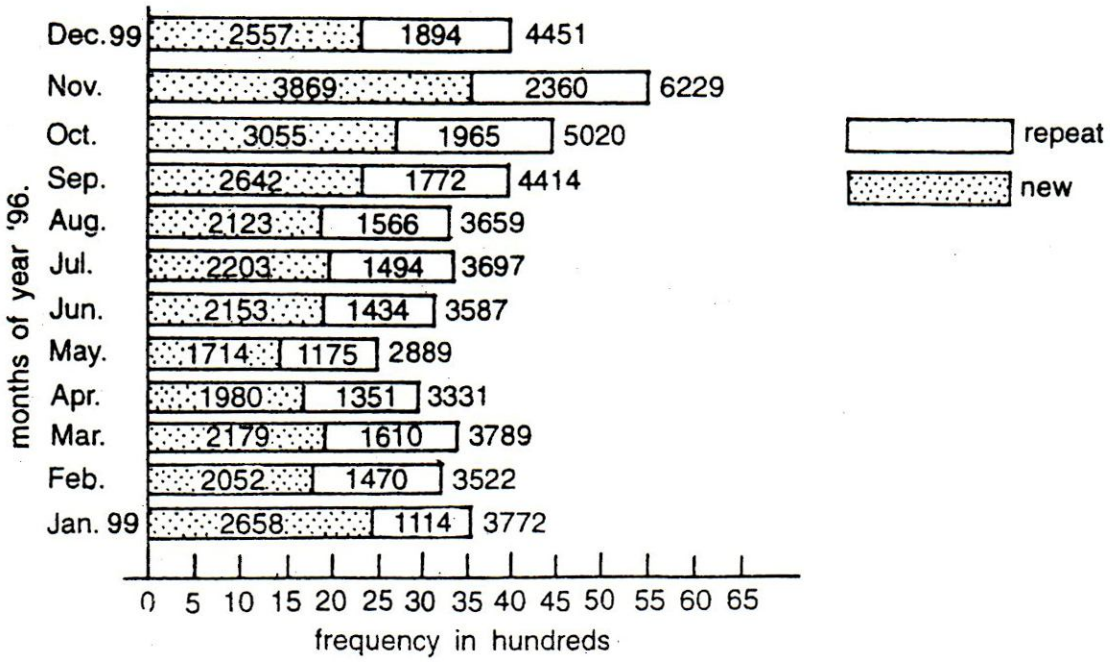


Fig. 11. Proportional bar diagram showing new and repeat outdoor patients given in table 13.

7.26 Summary

Statistics has wide application in almost all sciences-social as well as physical such as biology medicine, agriculture, veterinary, economics, psychology, ethology, business management etc. It plays a major role in bioscience because data of bioscience are of a variable nature. It is very difficult to draw a concrete conclusion from biological experiments because of inherent differences between two individuals.

It statistics various terms like population, data sample, variable, observation, parameter etc. and various symbols are used. Collected observations are called data. Data can be primary or secondary type, Qualitative or Quantitative data. These data are represented in Tabular, Graphical and Diagrammatically.

7.27 Self Learning Exercise

1. What do you mean by Data, Population, Sample, Variable, Parameter, Frequency distribution, Cumulative frequency, Primary data and Secondary data.
2. Explain the following notations or symbols :
   Σ, %, $\overline{X}$, x, f, P, O, E, δ, σ, k, $S\overline{E}_M$, SE_σ, SE_S
3. What is difference between population in general and population is biostatistics. Substantiate with suitable examples.
4. Define quantitative and qualitative variables. Give some examples of quantitative and qualitative variables.
5. What do you mean by population and sample. Describe method of sampling with suitable examples.
6. Define : (i) Random sampling, (ii) Non-random sampling, (iii) Quantitative data, (iv) Qualitative data and (v) Attribute.
7. Define continuous and discrete variables. Find whether the variable is continuous or discontinuous in the following cases :
8. Draw histogram, frequency polygon and cumulative frequency curve with the help of data mentioned i the following two tables :

Table A

Class interval	**Frequency**
1-10	3
11-20	14
21-30	21
31-40	25
41-50	40

Table B

Class interval	**Frequency**
24.5-29.5	3
30.5-35.5	6
36.5-41.5	14
42.5-47.5	20
48.5-53.5	25

51-60	40
61-70	47
81-90	50

54.5-60.5	37
61.5-66.5	39
67.5-72.5	42

9. The percentage of water, lipid, protein and other materials are 66.35%, 6.66%, 5.2%, 21.79% respectively in the body of a species of fish. Draw a pie chart with the help of the given data.

10. Following data were obtained in a hospital of Patna in respect of age and frequency of cancer. Make a frequency polygon :

Age	39-49	50-59	60-69	70-79	80-89
No. of cancer patients	2	3	15	21	

11. Draw a cumulative frequency diagram with the help of following frequency table.

Height of group in cm	**Frequency of each group**	**Cumulative class frequency**
160-162	10	10
162-164	15	25
164-166	17	42
166-168	19	61
168-170	20	81
170-172	26	107
172-174	29	136
174-176	30	166

176-178	22	188
178-180	12Σf=200	200

12. The body weight (g) and haemoglobin percentage (hb%-mg/100 ml) of 50 fishes of a species are given below : Make simple frequency table and non-overlapping cumulative frequency table :

Body weight (g) :	20.3	20.4	20.5	20.6	20.7	20.6	20.7	20.8	20.9	21	21
Haemoglobin (%) :	8.3	8.4	8.5	8.6	8.7	8.4	8.5	8.6	8.7	8.8	8.6
	21.1	21.2	21.3	21.4	22	22.1	22.2	22.3	22.4	23.7	23.8
	8.7	8.8	8.9	9	8.9	9	9.1	9.2	9.3	11.5	11.6
	23.9	24	24.1	24.4	24.5	24.6	24.7	24.8	14.6	14.7	14.8
	11.7	11.8	11.9	12.4	12.5	12.6	12.7	12.8	11.6	11.7	11.8
	25.9	26	27.1	27.2	27.3	27.1	27.5	27.9	28	28.1	28.2
	11.9	12	9.5	9.6	9.7	9.8	9.9	9.4	9.5	9.6	9.7
	28.3	28.4	28.3	28.4	28.5	28.6					
	9.8	12	12.1	12.2	12.3	12.4					

7.28 References

- Amble, V.N. — Statistical Methods in Animal Sciences, Indian Society of Agricultural Statistics, New Delhi.
- Bailey, N.T.J. — Statistical Methods in Biology, English University Press, London.
- Denenberg, V.H. — Statistical and Experimental Design for Behavioral and Biological Researchers, Hemisphere Publication Corp., Washington D.C.

- Snedecor, G.W. and Cochran W.G. — Statistical Methods, Oxford and IBH Pub. Co., New Delhi
- Sokal, R.R. and Rohlf, F.J. — Biometry : The principles and practice of statistics in biological research, W.H. Freeman and Co., San Francisco.
- S.P. Gupta — Statistical method, Published by Sultan Chand & Sons, Thirty Fourth Edition, 2005.
- T.K. Saha — Biostatistics in Theory and Practice, Emkay Publication, Delhi

Unit - 8

Frequency distributions & centering constants (Mean, Median and Mode). Measures of variation (standard deviation, variance, standard error of the Mean)

Structure of the Unit

8.1 Objectives

After going through this unit you will be able to understand:

- Collection, classification of statistical data.
- Construction of discrete and continuous frequency distribution tables.
- Types of measures of central tendency.
- Mathematical & Positional Averages.
- Measures of Dispersions.

8.2 Collection of Data

The study of techniques of collection of data and its presentation enables one to draw reliable conclusions from the collected data, which is obtained through various experiments and help in analysis and interpretation.

8.2.1 Specific Aspects of Statistical Data

There are four specific aspects of statistical data :

1. **Collection of data** : The first step in a statistical investigation is collection of data. A data collected in the original form is called **raw data** or ungrouped data. The data obtained by personal investigation is called

primary data. There are several methods of collection of primary data. In scientific research, data is obtained from experimental results.

2. **Presentation of data**: Collected data are presented in an orderly condensed manner to facilitate statistical analysis. There are different methods of presentation of data such as tables, diagrams, graphs etc.
3. **Analysis** : Data represented in tables, diagrams or graphs are analysed carefully. There are numerous methods of analysis of presented data. Measures of central tendency, measures of dispersion, correlation, regression etc. are a few examples of methods of analysis of presented data.
4. **Interpretation** : Drawing conclusion from analysis of data is called interpretation. Correct interpretation leads to valid conclusion.

8.2.2 Statistical units :

The units which the measurements are made in any statistical investigation are called the statistical units. In life science, survey of the unit may be a species or an individual.

8.3 Classification of Data

Data obtained from a experiment are classified i.e. converted into frequency distribution to make things simple and compact.

8.3.1 Meaning of classification

Classification is a process of condensation of raw data into systematized data that can be put up to a more systematic and proper use.

8.3.2 Need of classification

Statistical calculations from raw data are not advisable because it will require too much of time, space and labour. It is better to summarize the raw data into a **frequency distribution** table and then to give statistical treatment. The purpose of classification of data is to organise the data into a more compact form without obscuring the essential information contained in the values. The frequency distribution presents data very concisely indicating the number of repetition of values of variables. It records how frequently a variable occurs in a group study.

8.3.3 Objectives of classification

Process of classification is carried out with the following objectives :

(1) To bring out the unity of attributes out of the diversified things in the collected data.

(2) To condense the universe and to make things easily intelligible.,

(3) To make the study and comparison easier.

(4) To give prominence to the important information gathered while dropping out unnecessary elements.
(5) To put up the collected material to statistical treatment.
(6) To help the drafting of the final report.
(7) To simplify the complexities of the raw data and make it possible to draw statistical inferences.
(8) To make proper use of the collected data.

8.4 Construction of frequency distribution

If there are repetitions in individual values or items of investigation suitable frequency table can framed. These frequency tables may be discrete or continuous in nature, but they must maintain the frequency concerned in their respective classes.

8.4.1 Discrete frequency distribution

All the observations are listed in ascending o descending orders.

Following raw data were obtained in a biological experiment. Rate of reproduction (Fecundity) of 45 fishes was recorded as follows :

Raw data (A)

80	70	70	70	16	50	20	20	20
45	16	50	30	65	40	30	50	50
70	45	20	70	2	79	16	20	19
40	50	30	2	45	30	50	45	30
40	45	80	50	39	50	50	20	30

A frequency distribution table is framed on the basis of above raw data. Following steps are taken while framing frequency distribution table.

Converting raw data in arrayed data : The primary duty of a biostatistician is to convert **raw data** in **arrayed data.** This can be done by arranging the raw data into ascending or descending orders. For biostatistics data are usually arranged in an ascending order. The above raw data arranged in **ascending order** to make **arrayed data.**

Arrayed data (B)

2	2	16	16	16	19	20	20	20	20	20	20	30	30
30	30	30	30	39	40	40	40	45	45	45	45	45	50
50	50	50	50	50	50	50	50	65	70	70	70	70	70
79	80	80											

Presentation of data in this form is quite a tedious and time consuming job particularly when the number of observations in an experiment is large. To make it easily understandable, we can tabulate data in a simple frequency table.

Framing a **simple frequency** table (grouped series in discrete condition):

(1) A table of two columns is framed. First column contains **variables** and the second column for **repetition number** i.e. frequency or variables.

(2) On perusal of the above arrayed data (B), we find that variable 2 is repeated twice. Therefore, frequency 2 is mentioned in second column i.e. in the frequency column. Variable 16 is obtained three times and hence 3 is mentioned. In the same fashion frequencies of all variable are mentioned and following discrete frequency distribution table is framed (Table 1).

***Table* 1.**

Variable	2	16	19	20	30	39	40	45	50	65	70	79	80
Frequency	2	3	1	6	6	1	3	5	9	1	5	1	2

8.4.2 Continuous distribution table :

Arranging data in ascending or descending order is a tedious job. Moreover making discrete frequency distribution table consumes a lot space and time. Therefore, to bring out certain salient features of data, we further simplify presentation of data and condense them into classes or groups.

Class interval or Class

When a large number of observations varying in a wide range are available, these are classified in several groups according to the size of values. Each of these groups defined by an interval is called class interval or class. In practice statistician usually take a minimum of 3 and maximum of 20 classes.

See **arrayed data (B)**. It shows the highest value of observation is 80 and lowest value is 2 and number of classes i.e. *k* is 5.

$$i = \frac{80-2}{5} = \frac{78}{5} = 15$$

Thus, a table may be framed having width of class interval 10 or 15. But for convenience, an investigator can take suitable number of classes and width of class interval. Frequency distribution table in class interval may be prepared in two ways:

8.4.3 Overlapping frequency table or exclusive method

Values of variables are grouped in such a fashion that the upper limit of one class interval is the lower limit of succeeding class interval. An overlapping class interval frequency distribution table 2 can be prepared using data of table 1.

Data of table 1 tells that the rate of reproduction of the given specie of fish range between 2 and 80. We can keep the width of class interval 10. Then the range of first class interval will be 0-10, 2nd between 10-20, 3rd between 20-30 and so on. Here on thing is **remarkable** - fishes having rate of reproduction upto 9 are taken into consideration in the first class interval. Therefore, the frequency of class interval 0-10 is 2. Fishes having 10 rate of reproduction have to be included in the succeeding class interval. Four fishes come under second class interval. Hence, frequency of 2nd class interval is 4 (Table 2).

8.4.4 Non-overlapping class interval or inclusive method

Values of variables are grouped in such a fashion that the upper limit of one class interval do not overlap the succeeding class interval. A non-overlapping frequency table 3 can be prepared using the data of table 1.

Here class interval may be 1-10, 11-20, 21-30 and so on keeping the width of class interval 10. Here upper limit of one class interval is not overlapped by lower limit of preceeding class interval (Table 3.)

Table 2. ***Table 3.***

Overlapping frequency table		Non-overlapping frequency table	
Class interval	**Frequency**	**Class interval**	**Frequency**
0-10	2	1-10	2
10-20	4	11-20	10
20-30	6	21-30	6
30-40	7	31-40	4
40-50	8	41-50	14
50-60	9	51-60	0
60-70	1	61-70	6
70-80	6	71-80	3
80-90	2	81-90	0
	$\Sigma f = 45$		$\Sigma f = 45$

Note : For biostatistics usually non-overlapping frequency distribution table is used.

8.5 Measures of Central Tendency

Generally it is found that values of the variable tend to concentrate around some central value of observations of an investigation, which can be taken as a representative for the whole data. This tendency of the distribution is known as central tendency and the measures devised to consider this tendency are known as measures of central tendency. Measures of central tendency provide a single figure called average which describes the entire series of observations.

8.5.1 Types of Measures of Central Tendency

There are usually three basic measures of central tendency. These are :

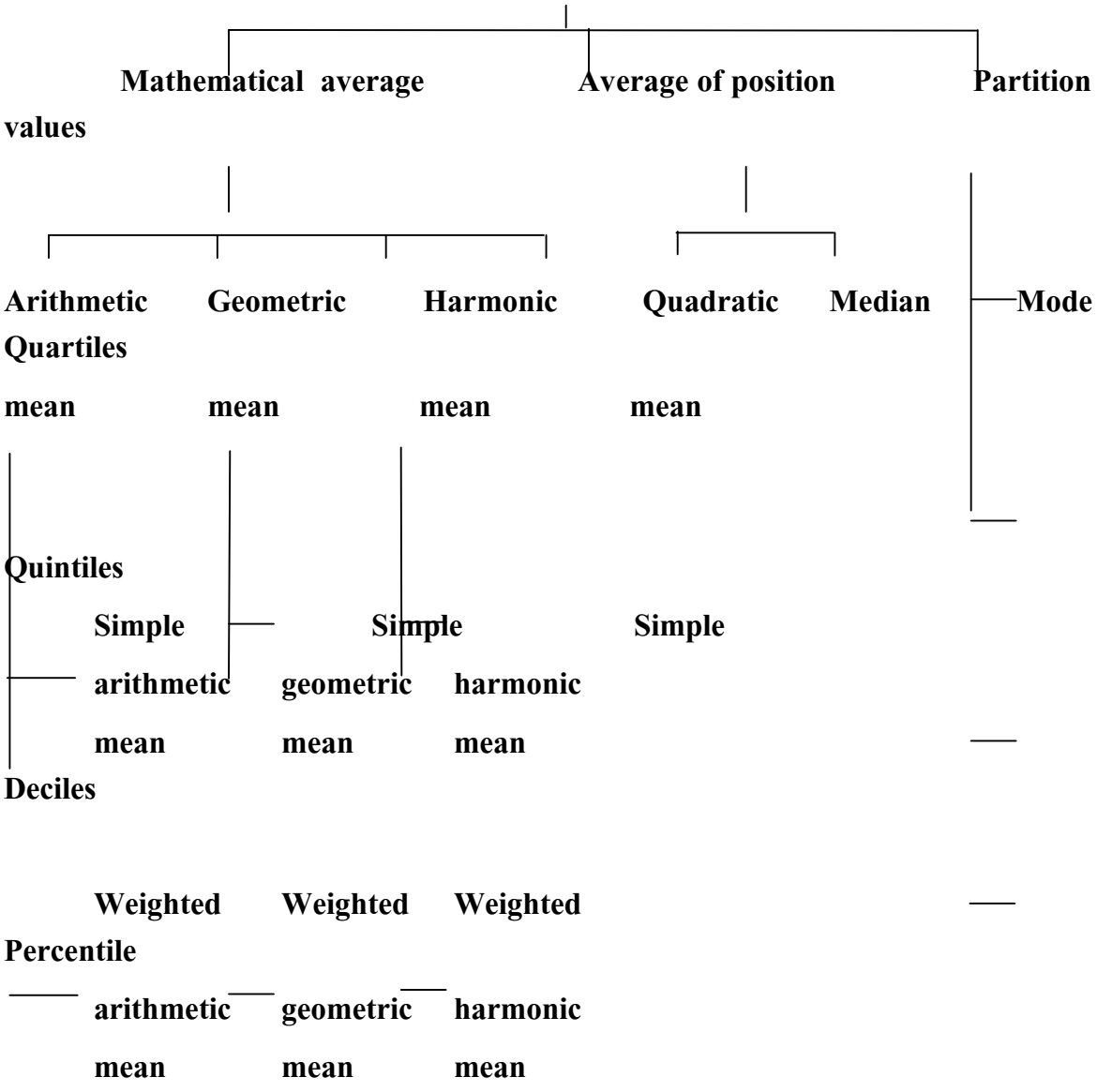
(1) Mathematical average
(2) Averages of position
(3) Measures of partition values

(1) **Mathematical average.** Averages represented purely in mathematical values are known as mathematical average. It is of three types : (i) Arithmetic mean, (ii) Geometric mean, (iii) Harmonic mean.

(2) **Averages of position.** Mean exhibited by position is called average of position. It is of two types : (i) Median and (ii) Mode.

(3) **Measures of partition value.** It is measures of location. It divides the total observations by an imaginary line into two or more parts expressed in percentage.

Types of measures of central tendency



Combined mean

8.6 Mathematical Average

8.6.1 Arithmetic mean

Central value or average, obtained arithmetically, is known as arithmetic mean. It is the most common average, used in our day today life. Depending upon whether all the items in the data are to be considered of equal or unequal importance we get three sub-types of arithmetic mean. Accordingly arithmetic mean are of following types :

(1) Simple arithmetic mean

(2) Combined mean

1. Simple arithmetic mean : It is most commonly used of all the averages. It is the value which we get by dividing the aggregate of various items of the same series by the total number of observations.

Arithmetic mean (Ungrouped data): If the values of *N* items are X_1, X_2, X_3, X_n..... be the value of variate *X*, then simple arithmetic mean *(X)* is obtained by dividing the sum of the values of all the items by the total number of observations. Symbolically :

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{N} \quad or \quad \overline{X} = \frac{\sum X}{N} \quad or \quad X = \frac{1}{N} \sum_{i=1}^{N} Xi$$

Where $\overline{X}$ = Arithmetic mean

$X_1, X_2, X_3, \ldots\ldots\ldots X_n$ are each individual observation. ΣX = Sum of individual observations of the variable. N = number of observations.

Worked example : Haemoglobin percentage (Hb%) of ten patients suffering from AIDS was recorded as 5.2 mg, 5.3 mg, 5.6 mg, 5.7 mg, 5.4 mg, 5.2 mg, 5.3 mg, 5.3 mg, 5.4 mg and 5.2 mg. Find out the mean Hb% of patients suffering from AIDS (Hypothetical data).

Solution : Following formula is applied to obtain arithmetic mean from above ungrouped data.

$$\overline{X} = \frac{\sum X}{N} = \frac{5.2+5.3+5.6+5.7+5.4+5.2+5.3+5.3+5.4+5.2}{10}$$

$$= \frac{53.6}{10} = 5.36\ mg/100\ ml.$$ **Ans**.

Computation of arithmetic mean for grouped data : ***(a) Discrete series.*** If data is in frequency distribution but not in class interval, then it is called as discrete series. For computation of arithmetic mean in a discrete series, each value of the variable in multiplied by their respective frequencies. Sum of all values is obtained which is divided by total number of frequencies.

Let the variable X take the values $X_1, X_2, X_3 .....X_n$ and let their frequencies be $f_1, f_2, f_3....f_n$. Then the arithmetic mean $(\overline{X})$ is computed by formula :

$$\overline{X} = \frac{f_1.X_1 + f_2.X_2 + f_3.X_{3,}.....f_n.X_n}{f_1 + f_2 + f_3.... + f_n} \ or \ \frac{\Sigma f.X}{\Sigma f}$$

Where, $\overline{X}$ arithmetic mean; $\Sigma f.X$ = sum of values of the variables and their corresponding frequencies and Σf= sum of frequencies.

Worked example : Rate of respiration of 43 fishes and their respective frequency was recorded as follows :

Find the arithmetic mean from data.

Rate of Resp.^{n}	2	16	20	30	39	40	45	49	50	65	70	79	80
Frequency	3	4	7	7	1	3	5	1	2	2	5	1	2

Solution : First step is to make a table of three columns. 1st column for variable (Rate of respiration), 2nd for corresponding frequencies and 3rd for variable × frequency.

Table 1.

Variable	**Frequency**	**Variable**
(X)	***(f)***	***(f.X)***
(X)		***(F.X)***

2	3	6
16	4	64
20	7	140
30	7	210
39	1	39
40	3	120
45	5	225
49	1	49
50	2	450
65	2	130
70	5	350
79	1	79
80	2	160
	$\Sigma f = 43$	$\Sigma f.X = 2022$

$\because \quad \overline{X} = \sum f.X / \sum f; \quad \therefore \quad \overline{X} = 2022/43 = 47.02$ Ans.

(b) Continuous Series : In the case of grouped continuous series, the arithmetic mean is calculated after taking into consideration the mid-points of various classes. Hence, formula given for discrete series is also applicable to grouped continuous series.

$$\overline{X} = \frac{\sum f.m}{\sum f}$$

Where $\overline{X} =$ Arithmetic mean

$\sum f.m =$ Sum of values of mid-points multiplied by their corresponding frequencies.

$\sum f =$ Sum of frequencies

m = mid point of various class intervals.

Mid point is obtained by following method

$$Mid - point\ (m) = \frac{Lower\ \ limit\ of\ \ class\ interval + Upper\ limit\ of\ \ CI}{2}$$

Worked example : Rate of respiration of 50 fishes of a species and their frequencies are given in continuous series. Find out mean of this experimental data.

Rate of Resp.^{n}	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	3	11	7	4	10	5	7	3

Solution : Make a table of 4 columns. 1st column for class-interval of variable, 2nd for mid-point, 3rd for frequency and 4th for mid value multiplied by their respective frequencies.

Table 2.

Class-interval ***(CI)***	**Mid-Value** ***(m)***	**Frequency** ***(f)***	**Mid-value**×Freq. $m \times f$
1-10	(1+10)/2=5.5	3	5.5×3=16.5
11-20	(11+20)/2=15.5	3	15.5×11=170.5
21-30	(21+30)/2=25.5	7	25.5×7=178.5
31-40	(31+40)/2=35.5	4	35.5×4=142.0
41-50	(41+50)/2=45.5	10	45.5×10=455.0
51-60	(51+60)/2=55.5	5	55.5×5=277.5
61-70	(61+70)/2=65.5	7	65.5×7=458.5
71-80	(71+80)/2=75.5	3	75.5×3=226.5

		$\Sigma f = 50$	$\Sigma m \times f = 1925$

$\overline{X} = \sum f.m / \sum f$ = 1925/50=38.5 **Ans.**

[**Note** : Here in place of (X) mid value (m) is applied].

2. Combined mean : If $\overline{X_1}$ and $\overline{X_2}$ are the arithmetic mean of two series whose number of terms are N_1 and N_2 respectively, then the combined mean $\overline{X}$ is given by

$$\overline{X} = \frac{N_1 \overline{X}_1 + N_2 \overline{X}_2}{N_1 + N_2}$$

Worked example : If the mean length of a group of 40 earthworms is 65 cm and the mean length of other group of 50 earthworms is 60 cm. find the combined mean length of the two groups.

Solution : $\overline{X} = \frac{40 \times 65 + 50 \times 60}{40 + 50} = \frac{2600 + 3000}{90} = \frac{5600}{90}$ = 62.2 cm **Ans.**

Calculation of arithmetic mean is series having open-end classes : In a few cases, the lower limit of the initial class interval and the upper limit of the last class interval are not known. Such class intervals are called open-end classes. Arithmetic mean cannot be calculated in open-end classes unless one knows the lower and upper limits of the classes. In this circumstance, an assumption is made for unknown class-limit and it depends upon the class interval.

Worked example : Class interval in uniform

No. of Seeds	Below 10	10-20	20-30	30-40	40-50	50-60	Above 60
No. of pods	7	9	17	11	10	6	3

Solution : In above example the class interval in uniform. Therefore, the lower limit of the initial class would be zero and the upper limit of last class would be 70. Thus the first class is 0-10 and last class 60-70.

Worked example : Class interval is not uniform

No. of Seeds	Below 10	10-30	30-60	60-100	100-150	Above 150
No. of pods	7	26	27	19	5	7

Solution : In above case the class interval in increasing by 10. Therefore, the appropriate assumption would be that the first class in 0-10 and the last class 150-210.

Merit and demerits of arithmetic mean : It is mot commonly used for measures of central tendency though it has got both merits and demerits.

Merits :

(1) It is rightly defined and is an easy and ideal measures of central tendency.
(2) It covers all the observations and is easy to calculate and understand.
(3) It is affected least by fluctuation of sampling. In other words arithmetic mean is a stable average.
(4) Arithmetic mean provides base of many other methods of statistics.

Demerits :

(1) Obtained mean in a series may not be represented by any observation.
(2) It is very much affected by extreme observation.
(3) By eliminating even a single series, calculation becomes unreal.
(4) It cannot be determined by inspection nor can be represented graphically.
(5) In extremely skewed distribution arithmetic mean is not representative of the distribution.

8.6.2 Geometric mean

The **geometric mean** is defined as the Nth root of the product of n observations.

Computation of geometric mean for ungrouped data (individual series). If X_1, X_2, X_3 .....X_n are the numbers of the data then their geometric mean (GM) = $\sqrt[n]{X_1, X_2, X_3.....X_n}$ or GM = $(X_1, X_2, X_3....X_n)^{1/n,}$ where X_1, X_2, X_3 .....X_n are the N values of the variate X.

Note : The above method can be applied if there are two or three items. But if n is very large number, the problem of finding the nth root is very difficult and as such logarithms are used to facilitate the computation of geometric mean. For this reason geometric mean is also referred as logarithmic mean. When logarithms are used, geometric mean (GM) can be calculated as the arithmetic mean. The logarithm of **GM** is equal to the Arithmetic mean of logarithms of individuals values.

$$\text{GM} = \text{Antilog}\ \frac{log\ X_1 + log\ X_2 + log\, X_3.....log\, X_n}{N} = Anti\, log\ \frac{log\, X}{N}$$

Worked example : On 1st January weight of a pig was recorded as 14 kg. on 1st March weight of the same pig was 20 kg. What was the approximate weight of the pig on 1st February.

Solution :

$\because$ **(GM)** = $2\sqrt{X_1 \,.\, X_2}$ *or* $(X_1 \,.\, X_2)^{1/2}$ for two values

$\therefore$ **(GM)** = $(14 \times 20)^{1/2} = \sqrt{280} =$ 16.73 kg.

The weight of the pig was 16.73 kg approximately on 1st February.

Computation of GM for grouped data.

(a) Discrete series.

(GM) = $\sqrt[N]{X_1^{f_1}, X_2^{f_2} \ldots\ldots X_n^{f_n}}$

Where $N = f_1 + f_2 + \ldots\ldots f_n = \sum_{i=1}^{n} fi$

or (GM) = Antilog (Σf . log x / Σf)

Where X = values of the variate , Σf = sum of all frequencies.

i.e. total no. of observation.

Worked example : The number of Basophils (a kind of WBC) of 30 patients was recorded in frequencies. Calculate the GM.

No. of Basophils	11	14	17	19	22
Frequency	5	6	8	7	4

Solution : A table of four columns is prepared to find the goal. 1st column for scores, second for frequency, 3rd for log x and 4th for f log x.

Table 3.

Variable *(X)*	**Freq. *(f)***	**Log (*x*)**	**Freq. log *x***
11	5	1.0414	5.2070

14	6	1.1461	6.8766
17	8	1.2304	9.8432
19	7	1.2788	8.9516
22	4	1.3424	0.3696
	$\Sigma f = 30$		$\Sigma f.\log x = 36.2480$

GM= antilog GM=Σf . log $x/\Sigma f$ = antilog (36.2480/30) = antilog (1.20826)

GM = antilog of 1.20826= 16.15 **Ans.**

(b) Continuous series. When we have to find the GM of a frequency distribution for continuous series, the frequencies of the various classes are taken as weights (W) and the mid-point of each class *(m)* is taken as the value of the class interval. Symbolically :

GM= antilog ($\Sigma f.\log m/\Sigma f$)

where , m = mid-point of class interval.

Worked example : Calculate the GM from the data obtained in an experiment related to the number of panicles and number of grains in wheat.

No. of grains	51-100	101-150	151-200	201-250	251-300
No. of panicles	7	9	10	8	5

Solution : First step is to frame a table of 5 columns for those components which are required for formula.

1st column for no. of grains, given in class interval.

2nd column for mind points (m) of each class interval

3rd column for number of panicles i.e. frequency

4th column for log m (or log X)

5th column for frequency multiplied with m.

Table 4.

No. of grains in	**Mid-point**	**No. of**	**log**	***f.* log**

class-interval ***(x)***	***(m)***	**panicles** ***(f)***	***(m)***	***m.***
51-100	75	7	1.88	13.16
101-150	125	9	210	18.90
151-200	175	10	2.24	22.40
201-250	225	8	2.35	18.8
251-300	275	5	2.44	12.2
		$\Sigma f = 39$		$\Sigma f.\log m = 85.46$

GM = antilog ($\Sigma f.\log m/\Sigma f$)

= antilog (85.46/39)

= antilog 2.19 (155.34) **Ans .**

Note : The Geometric mean is always less than the Arithmetic mean unless all the quantities $X_1, X_2, ...X_n$ are equal.

Merits and demerits of Geometric mean

Merits :

(1) It is based on all the observations.
(2) It is rigidly defined.
(3) It is capable of further algebric manipulation.
(4) It is not much affected by fluctuation of sampling.
(5) It is particularly useful in dealing with ratios, rates and percentages.

Demerits :

(1) It cannot be used when any of the quantities are zero or negative.
(2) It is difficult to calculate and interpret.
(3) It may come out to be a value which is not existing in the series.

8.6.3 Harmonic mean

The Harmonic mean is defined as the "reciprocal of the arithmetic mean of the reciprocals of the given values." For example, reciprocal of 5 in 1/5, reciprocal of 9

is 1/9 and so on. If variables are expressed in ratios or rates, the proper average to be used is Harmonic mean.

Computation of Harmonic mean for Ungrouped data.

For observations $X_1, X_2, X_3 \ldots\ldots X_n$

$$HM = \frac{N}{1/X_1 + 1.X_2 + 1/X_3 + \ldots.. 1/X_n} = \text{or} \frac{N}{\Sigma(1/X)}$$

Worked example : Find the Harmonic mean of the following data relating to the weight of ovary of 8 fishes in g : 20.1, 22.0, 18.1, 30.2, 18.1, 24.0, 32.0, 30.0 g. respectively.

Solution : Make a table of 2 columns. 1st for variable and 2nd for reciprocal of value of variable.

Table 5.

Values of X	20.1	22.0	18.1	30.2	18.1	24.0	32.0	30.00	$\Sigma 1/X = 0.344$
Frequency	0.049	0.045	0.055	0.033	0.055	0.041	0.031	0.033	

$$HM = \frac{N}{\Sigma 1/X} = \frac{8}{0.344} = 23.95$$ **Ans.**

Worked example : A migratory bird migrates from feeding place to breeding place at a speed of 70 km/hour and returns at the speed of 50 km/hour. Calculate HM.

Solution :

$\therefore$ $HM = \frac{N}{\Sigma(N/X)}$ Here, N=2; X_1=70 and X_2 =50

$\therefore$ $HM = \frac{2}{1/70 + 1/50} = \frac{2}{0.0143 + 0.02} = \frac{2}{0.0343} = 58.33$ km/hour.

Harmonic mean for grouped data. ***(a) Discrete series.*** Following formula is used to calculate HM in discrete series : Suppose frequencies of a score $X_1, X_2, X_3 \ldots\ldots X_n$ are $f_1, f_2, f_3 \ldots\ldots f_n$ respectively,

Then, $HM = \frac{f_1 + f_2 + f_3 + \ldots\ldots + f_n}{f_1/X_1 + f_2/X_2 + f_3/X_3 + \ldots\ldots + f_n/X_n}$ or,

$$HM = \frac{\Sigma f}{\Sigma(f/X)}$$

Worked example : Hb% of 10 persons of a family was recorded and placed in frequency distribution discrete series (Hypothetical data).

Hb%	12 mg	13 mg	14 mg	15 mg	16 mg
Frequency	3	3	1	2	1

Solution : The reciprocals of the various should be taken first and then the reciprocal multiplied by the respective frequency and total $\Sigma f / X)$ is obtained.
A table of 4 columns is prepared. 1st for values of variables, 2nd for frequency, 3rd for reciprocal of X and fourth for *f/X*.

Table 6.

Hb%(mg/100ml) value (*X*)	**Freq. *(f)***	**1/*(X)***	**F/*X***
12	3	0.083	0.25
13	3	0.076	0.23
14	1	0.071	0.07
15	2	0.066	0.13
16	1	0.062	0.06
	$\Sigma f = 10$		$\Sigma f / X = 0.746$

$$\text{HM} = \frac{\Sigma f}{\Sigma(f / X)} = \frac{10}{0.746} = 13.39$$ **Ans.**

(b). Continuous series : In a continuous series, we take the mid-points of class intervals. Here we take reciprocal of the mid-points. Following formula is used :

$$\text{HM} = \frac{\Sigma f}{\Sigma(f / m)}$$

Worked example : Length of 68 plants was recorded and presented in following frequency distribution continuous series.

Classes	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	6	12	14	16	8	8	4

Solution : A table of 4 columns is prepared. 1st for class interval, 2nd for mid-points, 3rd for frequency and 4th for f/m.

Table 7.

Class interval ***(X)***	**Mid-points** ***(m)***	**Frequency** ***(f)***	***(f/m)***
10-20	15	6	0.40
20-30	25	12	0.48
30-40	35	14	0.40
40-50	45	16	0.36
50-60	55	8	0.15
60-70	65	8	0.12
70-80	75	4	0.05
		$\sum f = 68$	$\sum f/x = 1.96$

$$\text{HM} = \frac{\Sigma f}{\Sigma(f/m)} = \frac{68}{1.96} = 34.69 \textbf{ Ans.}$$

Merits and demerits of Harmonic mean

Merits :

(1) It is rightly defined.

(2) It is based on all the observations of a series.

(3) It gives greater weight age to the smaller items.

(4) It is useful to study the rate of respiration, rate of pulse, heart beat etc. in unit time.

(5) It is not much affected by sampling fluctuations.

Demerits :

(1) It is not easy to calculate and understand.

(2) It cannot be calculated if one value is zero.

(3) It cannot be calculated if negative and positive values are given in a series.

Relationship between Arithmetic mean,

Geometric mean and Harmonic mean

The Arithmetic mean (AM), and Geometric mean (GM) are greater than the Harmonic mean (HM). The sequence even for the same value -AM > GM > HM.

Worked example : The length of 4 larvae of an insect was recorded as 2, 3, 4 and 5 cm. Now calculate AM, GM and HM and establish the greater value in terms of relationship.

$$\text{HM} = \frac{\Sigma X}{N} = \frac{2+3+4+5}{4} = \frac{14}{4} = 3.50$$

$$\text{GM} = (X_1 \,.\, X_2 \,.\, X \,.\, X_4)^{1/4} = (2\times 3\times 4\times 5)^{1/4} = (120)^{1/4} = 3.31$$

$$\text{HM} = \frac{N}{\Sigma(1/X)} = \frac{4}{(1/2+1/3+1/4+1/5} = \frac{4}{1.28} = 3.13$$

We find that even for the same values of variables of a series

AM > GM > HM.

8.7 Averages of Position

As the name suggests averages of position indicate the position of an average in a series of observation arranged in increasing order of magnitude. Averages of position is of two types.

8.7.1 Median

A median of a distribution is defined as the value of that variable which divides the total frequency into two equal parts when the series is arranged in ascending or descending order of magnitude.

Median for ungrouped data. Median value is the value of the $(N + 1/2)^{th}$ item. Median is computed differently if number of observation (n) is **odd** and when **even.**

Computation of Median when N is odd.

Merits and demerits of mode

Worked example : RBCs count in lac/mm^3 was recorded as 6, 7, 4, 5, 5, 3, 4 in different blood samples of an animal. Compute Median.

Solution : Our first step in to arrange the value of series in an ascending order e.g. 3, 4, 4, 5, 5, 6, 7.

Here total number of observation is 7 i.e. **odd** number.

Median = $\left(\frac{N+1}{2}\right)^{th}$ item = $\left(\frac{7+1}{2}\right)^{th}$ item = 4^{th} item.

In above example, 4^{th} item = 5 ∴ Median = 5 Ans.

Computation of Median when *N* in even. In case of even series computation of median is slightly different than **odd** series. Median is given by arithmetic mean of the middle terms, $N/2^{th}$ and $\left(\frac{N+1}{2}\right)^{th}$ item

∴ Median = $\frac{(N/2)^{th}+(N/2+1)^{th}}{2}$ item.

Worked example *:* Oxygen consumption of 8 fishes was recorded as : 35, 44, 38, 36, 39, 40, 42 and 41 cc/100ml/hour. Find Median from the given data.

Solution : Arrange the raw data in ascending order : 35, 36, 38, 39, 40, 41, 42 and 44.

Median = $\left(\frac{n+1}{2}\right)^{th}$ item = $\left(\frac{8+1}{2}\right)^{th}$ item $\frac{9}{2}=4.5^{th}$ item.

It tells that Median lies in between 4^{th} and 5^{th} items. Average of the 4^{th} and the 5^{th} items is calculated as below :

4^{th} item = 39 and 5^{th} item = 40 ; ∴ Median = $\frac{4^{th}\text{ item}+5^{th}\text{ item}}{2}$

Median = $\frac{39+40}{2}=\frac{79}{2}=39.5$ **Ans.**

Computation of median for grouped data : ***(a) Discrete series*** : In a discrete series the item are first arranged according to the ascending order of magnitude and respective frequencies are written against them. The frequencies are then cumulated and the position of the Median is located by the same formula :

Median = $\frac{(n+1)^{th}}{2}$ item where $N=f_1+f_2+.....f_n=\Sigma f_i$

Worked example : Protein content of 100 fishes in (g) of a species was obtained as below. Find the median.

Protein content	5	5.5	6	6.5	7	7.5	8	8.5
Frequency	2	4	8	10	15	25	22	14

Solution : Make a table of 3 columns. 1^{st} for variable, 2^{nd} for frequency and 3^{rd} for cumulative frequency.

Table 8.

Protein	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5
Frequency	2	4	8	10	15	25	22	14
Cum. Freq.	2	6	14	24	39	64	86	100

Median = Size of $\frac{(\text{Cumulative Freq.}+1)^{\text{th}}}{2}$ item = Size of $\frac{(100+1)^{\text{th}}}{2}$ item

$= \frac{101}{2} = 50.5^{\text{th}}$ item.

Now items from 39 to 64 have protein 7.5 as shown by cumulative frequency $\therefore$ Median = 7.5 **Ans.**

(b) Continuous series : In this case the median cannot by found directly without recourse to the original data. It may however be estimated with sufficient accuracy by interpolation. To do this we first find the position of the median item i.e., $\frac{(n+1)}{2}$ and from the cumulative frequencies of the class, we determine the class in which median item occupying this position lies. Then the Median is given by the formula.

$$\text{Median} = L_1 + \frac{N/2 - c}{f_m} \times h$$

Where, L_1 = lower limit of the class in which median lies. f_m = frequency of the class in which the median lies. c = cumulative frequency of the class preceding the median class. N = total frequency $= \Sigma f$ and h= width of the class interval of the median class.

Worked example : Following is the fecundity of 50 fishes of a species of fish n frequency distribution in continuous series. Find the median fecundity.

Fecundity in C.I.	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	3	15	2	8	11	4	1	6

Solution : Frame a table of 3 columns. 1st column for class interval, 2nd for frequency and third for cumulative frequency.

Table 9.

Fecundity	X	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	Y	3	15	2	8	11	4	1	6
Cum. Freq.	cf	3	18	20	28	39	43	44	50

Here, $N = 50$; Median = $\frac{(N+1`)^{th}}{2}$ item $\frac{(50+1)^{th}}{2}$ item $\frac{51^{th}}{2}$ item $= 25.5^{th}$ item

From the cumulative frequency column, we find that the 25.5^{th} item falls between 31-40 group.

Here, L_1 -= 31 ; $\Sigma f = 50$; $c = 20$; $f_m = 8$; $h = 10$

$$\text{Median} = L_1 + \frac{(N/2 - c}{f_m} \times h = 31 + \frac{(50/2 - 20)}{8} \times 10$$

$$= 31 + \frac{(25-20)}{8} \times 10 \text{ item}$$

$$= 31 + \left(\frac{5}{8}\right) \times 10 = 31 + (0.625) \times 10$$

$= 31 + 6.25 = 37.25$ **Ans.**

Merits and demerits of median

Merits

(1) If found directly, it represents an actual item.

(2) It eliminates the effects of extreme items, since they are not taken into account in its calculations, except of arranging the data in increasing or decreasing order.

(3) The values of only the middle items are required to be known.

(4) It can be found even for data which cannot be measured quantitatively. For example, if Hb% of 10 frogs are arrange according to their numerical value, the median Hb% is the Hb% of the middle frog.

(5) It is easy to calculate.

(6) The median is most suitable for expressing qualitative data such as colour, health, intelligence etc.

Demerits.

(1) It may not be representative when the distribution is irregular. For example, heights of different plants of selected area of a forest are : 1, 5, 9, 12, 6, 60, 61, 72, 78.

Arrange the data in ascending order of magnitude – 1, 5, 6, 9, 12, 60, 61, 72, 78. The median height is $\frac{(9+1)^{th}}{2}$ item = $\frac{10}{2}$ = 5^{th} item is 12 which is not at all a proper representative of the average height of plants.

(2) It cannot be located with precision when the items are grouped. Then it can only be estimated and the estimated value may not be found in the series.

(3) The data must be kept in ascending or descending order. This involves considerable work if the number of items is large.

(4) The aggregate value of the items cannot be obtained when the median and the number of items are known.

(5) It is very useful is further analysis, because it is difficult to handle mathematically.

8.7.2 Mode

Mode of a frequency distribution is defined as **"that value of the variable for which the frequency is maximum".**

The definition of mode indicates that mode cannot be determined from series of individual observations as it depends on the frequency of occurrence of the items. Hence to get the value of mode the series must be converted into a frequency distribution.

Mode. ***Discrete series.*** If the distribution is regular and only one maximum frequency is there (data is unimodal) then the mode value can be obtained by mere inspection.

Inspection method.

Worked example : Water percentage of 15 fishes of a species of fish was recorded as 60, 64, 62, 76, 70, 74, 70, 84, 82, 72, 76, 84, 78, 84 and 86. Find the mode of this series.

Solution : As stated above the series must be converted into a frequency distribution and get a grouped data. First of all data is arranged in ascending order. Not even single value is spared. It comes as 60, 62, 64, 70, 70, 72, 74, 76, 76, 78, 82, 84, 84, 84 and 86. Make a grouped frequency distribution table.

Water %	60	62	64	70	72	74	76	78	82	84	86
Frequency	1	1	1	2	1	1	2	1	1	3	1

On a perusal of above table we find that repetition of 84 is maximum times i.e. 3 times. None of other items appear so frequently. Therefore, 84 is the mode of the series.

Sometimes a series have more than one mode (bimodal or multimodal). Then mode is obtained by grouping method.

Grouping method : When the discrete series is bimodal or multimodal then grouping of frequencies of series is done to ascertain mode. Grouping has been done as follows :

First of all table of eight vertical columns is framed.

(1) Values of variable is mentioned is first column.
(2) Frequencies of all variables is noted in second column.
(3) Sum of 2-2 frequencies is noted in 3rd column.
(4) Sum of 2-2 frequencies ignoring the first frequency is noted in fourth column .
(5) Sum of 3-3 frequencies is noted is fifth column.
(6) Sum of 3-3 frequencies ignoring the first frequency is noted in sixth column.
(7) Sum of 3-3 frequencies excluding the first two frequencies is noted in seventh column.
(8) According to need, grouping of 4-4 or 5-5 or even more can be done is the same fashion.

After completing the table, detection of maximum frequency is done in each column of group. The variables of maximum frequency are bracketed in analysis table. After completion of table for each group, these bracketed figures are counted. Variables having maximum bracketed items are noted as Mode.

Worked example : Weight is (g) of 50 Grasshoppers and their frequency was recorded as :

Weight of grasshoppers	21	22	23	24	25	26	27	28	29	30
Frequency	4	2	6	4	9	9	7	5	1	3

Find out the Mode of the series.

Solution : Usually the item having maximum frequency is the Mode of the series. But here item 25 and 26 both have maximum and same frequency i.e. 9. Now the problems is that which item, 25 or 26, should be considered as Mode ?

In this circumstance method of grouping is used. With the help of grouping method, following table 10 has been prepared.

Table 10.

Wt. of grasshopper	**Frequency**						Items having maximum frequency
Variable	**f**	**In group of two**		**In groups of three**			
	I	**II**	**III**	**IV**	**V**	**VI**	
21	4	6	8	12			0
22	2						0
23	6	10	13		12		0
24	4						1
25	9	18		22		19	4
26	9		16		25		6
27	7	12					3
28	5			13		21	1
29	1	5	6		10		0
30	4						0

Method adopted to frame grouping table. Firstly all values of variables are mentioned in a column. Then frequencies are given in column I. In column II and

III the frequencies of column I are grouped in two's. In column II beginning from the first item. $1^{st} + 2^{nd}$, $3^{rd} + 4^{th}$, $5^{th} + 6^{th}$, $7^{th} + 8^{th}$ and $9^{th} + 10^{th}$. In column III frequencies are shown in two's but with a difference. The pairing is done from 2^{nd} item, 1^{st} item is left deliberately. Her beginning from the 2^{nd}, 3^{rd}, $4^{th} + 5^{th}$, $6^{th} + 7^{th}$, $8^{th} + 9^{th.}$. The last item could not be paired and was left untouched. In column IV the grouping is done in three's. Starting from the first item. $1^{st} + 2^{nd} + 3^{rd}$; $4^{th} + 5^{th} + 6^{th}$; $7^{th} + 8^{th} + 9^{th}$. In this case last item could not be considered. In column V, the grouping is done is **three's but 1^{st} item is ignored** i.e., $2^{nd} + 3^{rd} + 4^{th}$, $5^{th} + 6^{th}$, 7^{th}, $8^{th} + 9^{th}$, 10^{th}. In column VI the grouping is done is **three's but first two items are ignored** i.e., $3^{rd} + 4^{th}$, $5^{th} + 6^{th}$, 7^{th}, 8^{th} . Here last two items are left unused because third partner is not there. Thus a table 11 was prepared.
Total frequencies obtained from table 10 is analysed in table 11.

Table 11.

Columns	**Items having maximum frequencies item**
I	25, 26,
II	25, 26,
III	26, 27,
IV	24 25, 26,
V	25, 26, 27,
VI	26, 27, 28
Total	1, 4, 6, 3, 1

In column I maximum frequency is 9, 9 is repeated two times. One 9 represents 25 and 2^{nd} 9 represents 26. In column II maximum frequency is 18 which represents 9 of column I. First 9 representing 25 and 2^{nd} 9 representing 26. In same fashion modal value for III, IV, V and VI are mentioned.

On perusal of above inspection table is appears that item no. 26 is repeated maximum times i.e. 6 times and hence Mode = 26.

Mode. ***Continuous series***. In the case of bimodal or trimodal condition we prepare grouping and analysis table and find out the modal class. Then apply the formula.

$$\text{Mode} = \left(\frac{f_1 - f_o}{f_1 - f_o - f_2} \right) \times i \quad or \quad L_1 + \left[\frac{\Delta_1 +}{\Delta_1 + \Delta_2} \right] \times i$$

Where,

L_1 = Lower limit of modal class.

L_2 = Upper limit of modal class.

f_0 = Frequency of the preceding modal class.

f_1 = Frequency of modal class.

f_2 = Frequency of the succeeding modal class.

Δ_1 = The difference between the frequency of the modal class and the frequency of the premodal class (f_1-f_o)

Δ_2 = The difference between the frequency of the modal class and the frequency of the succeeding modal class (f_1-$f2$).

i = Width of the class interval.

Worked example : Length of 15 Earthworms were recorded as 30, 32, 31, 38, 35, 37, 35, 42, 41, 36, 38, 42, 39, 40 and 44 cm. Calculate Mode of above observations.

Solution : Make a table of two columns. One **for variable** in class interval and the other **for frequency** and then use above formula.

Table 12.

Class interval	30-34	35-39	40-45
Frequency	$3f_o$	$7f_1$	$5f_2$

$\Delta_1 = f_1 - f_o = 7\text{-}3 = 4,$
$\Delta_2 = f_1 - f_o = 7\text{-}5 = 2$

Here modal class is considered as 35-39 because this class has got maximum frequency.

$L_1 = 35, L_2 = 39, f_o = 3, f1 = 7, f_2 = 5, i = 5$

$$\text{Mode} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times i \quad 35 + \left(\frac{4}{4+2} \right) \times 5$$

$$= 35 + \frac{20}{6} = 35 + 3.33 + 38.33 \quad \textbf{Ans.}$$

Empirical formula for mode. Mode is also computed by the empirical relation.

Mode = 2 Median – 2 Mean

Thus, a knowledge of Mean and Median enables us to calculate the Mode roughly. Note that the above relation is only approximately true and should rather be used as a check on the values of mean, median and mode and not for finding any one of them when the other two are known.

Problems *:* Find the Mode of the following distribution

C.I.	0.6	6-12	12-18	18-24	24-30	30-36	36-42	42-48	48-54	54-60	60-66	66-72	72-78
Cum. Freq.	2	6	13	18	26	29	33	38	46	50	58	59	60

Solution : For finding mode by applying the formula : **Mode = 2 Md^{n}-2 mean.** Here we have to make a table of three columns. 1st for variable. 2nd for frequency. (Find frequency of each variable from given cumulative frequency). 3rd column for cumulative frequency.

Table. 13

C.I.	0.6	6-12	12-18	18-24	24-30	30-36	36-42	42-48	48-54	54-60	60-66	66-72	72-78
Freq.	2	4	7	5	8	3	4	5	8	4	8	1	1
Cum. Freq.	2	6	13	18	26	29	33	38	46	50	58	59	60

Median number = value of $\left(\frac{N+1}{2}\right)^{\text{th item}}$ value of $\left(\frac{60+1}{2}\right)^{\text{th item}}$

$= 30.5^{\text{th}}$ item

As 30.5th item lies under 33, hence median class 36-42.

Here, Median $= L_1 + \left(\frac{\Sigma f/2 - F}{fm}\right) \times i = 36 + \left[\frac{60-29}{2}\right] \times 6 = 37.5$

For mean we have to make a table for four columns.

Table 14.

Variable	0.6	6-12	12-18	18-24	24-30	30-36	36-42	42-48	48-54	54-60	60-66	66-72	72-78
Mid Point	3	9	15	21	27	33	39	45	51	57	63	69	75
Freq.	2	4	7	5	8	3	4	5	8	4	8	1	1
F.X.	6	36	85	105	116	99	156	225	408	228	504	69	75

$\Sigma f.X = 2212$; Mean $= \frac{\Sigma f.X}{\Sigma f} = \frac{2212}{60} = 36.8$

Mode = 3 Median – 2 Mean = 3×37.5 - 2×36.8 = 1125.5- 73.6 = 38.9 **Ans.**

Merits and demerits of mode

Merits :

(1) It avoids the effects of extreme (and hence abnormal) items.

(2) Often it can be ascertained by mere inspection.

(3) Only the values occurring with high frequencies are required to be known for its determination. All values need not be known.

(4) It refers to a measurement which is the most usual and hence the most likely variate.

(5) Bi-modal distribution may give a good indication of the heterogeneity of a population.

Demerits :

(1) It is not well defined and is rarely used for higher life science researches.

(2) Arithmetic explanation of mode is not possible.

(3) Sometimes it is indefinite.

(4) It becomes difficult in multi-modal distribution.

(5) It is not based on all the observations of a series.

Relationship between Mean, Median, and Mode. There is empirical relationship between **Mean, Median and Mode** of a series of items. If distribution of item values be symmetrical then the Mean, Mode and Median coincides, otherwise the distance between the Mode and the Median is usually twice the distance between the Median and the Mean Thus :

Mode-Median =2 (Median-Mean) or, Mode-Mean

= 3 (Median-Mean)

Note : Symmetrical distribution : A distribution in which Mean, Median and Mode coincide is called a symmetrical distribution.

8.8 Measures of Dispersion

8.8.1 Meaning of Dispersion

In Statistics, dispersion is used commonly to mean scatter, deviation, fluctuation, spread or variability of data. Dispersion is used to denote a lack of uniformity in item values of a given variable.

Definition

The degree to which the individual values of the variate scatter away from the average or the central value, is called is **dispersion.**

The measure of variation is one of the important tools of statistics for biologists because biological phenomena are more variable than that of physical and chemical sciences. Length, weight, Hb%, RBCs number, rate of oxygen consumption etc. of two individuals of same species, of same age and sex will differ definitely. Same plant of sweet pea produces different number of seeds in different pods. Same person exhibit different pulse rate and heart beat in different physiological conditions. Cure rate with the same drug varies in different patients, even of the same age and sex.

8.8.2 Standard Deviation

Standard Deviation may be defined as **"the square root oft he arithmetic mean of the squares of deviations from the arithmetic mean".** Karl Person 1893 gave the concept of Standard Deviation. It is widely used measures of studying dispersion.

Computation of Standard deviation S or (σ)for ungrouped data

Following six steps are observed before the computation of Standard deviation : (i) Calculate the mean (ii) Find the difference of each observation from the mean

(iii) Square the differences of observations from the mean (iv) Add the squared values to get the sum of squares. (v) Divide this sum by the number of observation. (vi) Find the square root of this variance.

Apply following formula to calculate Standard Deviation.

$$S = \sigma = \sqrt{\frac{\Sigma x^2}{N}} \text{ or } \sigma = \sqrt{\frac{\Sigma x^2}{N-1}}$$

Where S= Standard Deviation ; x = deviation obtained from mean i.e. X- $\overline{X}$

N = total number of observations.

[Note : S is computed by using N-1 in the denominator of above formula instead of N, if the size of sample i.e. total number of observations are less than 30. If the size of sample is more than 30 then first mentioned formula :

$$S = \sqrt{\frac{\Sigma x^2}{N}} \text{ is applied]}$$

Worked example : Haemoglobin percent g/100 ml of liver fed *Wallago attu* was recorded as 23, 22, 20, 24, 16, 17, 18, 19 and 21. Calculate the Standard Deviation.

Solution : As suggested above prepare a table of 4 columns. First column for values of variables i.e. X. 2nd column for observation-mean ($X - \overline{X}$). 3rd for deviation (x) and 4th for deviation square (x^2).

Calculate the Mean $\overline{X} = \frac{180}{9} = 20$

Table 1.

Variable X	**Observation-Mean $X - \overline{X}$**	**Deviation x**	**Deviation square x^2**
16	16-20	-4	16
17	17-20	-3	9
18	18-20	-2	4
19	19-20	-1	1
20	20-20	0	0

21	21-20	+1	1
22	22-20	+2	4
23	23-20	+3	9
24	24+20	+4	16
$\Sigma X = 180$			$\Sigma X^2 = 60$

Given formula is applied : $S = \sqrt{\frac{\Sigma X^2}{N-1}}$ (Sample size is less than 30).

Put the values, $S = \sqrt{\frac{60}{9-1}} = \sqrt{\frac{60}{8}} = \sqrt{7.5} = 2.75$ Ans.

Computation of Standard deviation

for grouped data ***(Discrete series)***

$$S = \sqrt{\frac{\Sigma f x^2}{\Sigma f}} \qquad \text{or} \qquad S = \sqrt{\frac{\Sigma f x^2}{\Sigma f - 1}}$$

Worked example : Weight of ovary of 50 fishes is given below in a simple frequency distribution table. Find the standard deviations.

Ovary weight	2	2.5	2.7	2.9	3	3.1	3.3	3.7	3.9	4	4.6	4.8	4.9	5	5.5	5.9	6	6.1	6.7	6.9
Freq.	2	1	1	2	3	1	3	2	4	3	2	3	3	3	2	3	3	3	3	3

Solution : A table of six columns is prepared :

First column for variable i.e., ovary weight. 2nd column for frequency. 3rd column for frequency multiplied by variable. 4th column for deviation obtained by (X - $\overline{X}$) formula. 5th column for deviation square. 6th column for frequency multiplied by deviation square.

Table 2.

X	***f***	***f.X***	***x***	x^2	$f.x^2$

2	2	4	-2.62	6.864	13.736
2.5	1	2.5	-2.12	4.494	4.494
2.7	1	2.7	-1.92	3.686	3.686
2.9	2	5.8	-1.72	2.958	5.916
3	3	9.0	-1.62	2.624	7.872
3.1	1	3.1	-1.52	2.310	2.310
3.3	3	9.9	-1.32	1.742	5.226
3.7	2	7.4	-0.92	0.846	1.692
3.9	4	15.6	-0.72	0.518	2.072
4.0	3	12.0	-0.62	0.384	1.152
4.6	2	9.2	0.02	0.0004	0.0008
4.8	3	14.4	0.17	0.028	0.084
4.9	3	14.7	0.28	0.078	0.234
5.0	3	15.0	0.37	0.136	0.408
5.5	2	11.0	0.87	0.756	1.512
5.9	3	17.7	1.27	1.612	4.836
6.0	3	18.0	1.37	1.876	5.628
6.1	3	18.3	1.47	2.160	6.48
6.7	3	20.1	2.07	4.284	12.852
6.9	3	20.7	2.68	7.182	21.546
		$\Sigma f.X = 231.1$			$\Sigma f.x^2 = 101.74$

Solution :

$$\text{Mean} = \frac{\Sigma fX}{\Sigma f} = \frac{231.1}{50} = 4.622$$

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f}} = \sqrt{\frac{101.74}{50}} = \sqrt{2.0348} = 1.427 \text{ Ans.}$$

Computation of Standard deviation by direct method

Standard deviation (S) can also be calculated by following formula where we do not need to obtain deviation : $S = \sqrt{\frac{\Sigma fX^2 - \overline{X}^2}{\Sigma f}}$

Worked example : A sample of ten common limpet shells (Patella vulgaris) from a rocky shore having the following maximum basal diameters in millimeters : 36, 34, 41, 39, 37, 43, 36, 37, 41, 39. Find out basal diameter and St. deviation of the shell.

Solution : In order to calculate the mean maximum basal diameters and the Standard deviation it is necessary to calculate $\Sigma X / f / X^2$ and $\overline{X}^2$ as mentioned in the table 3.

Table 3.

X	F	fX	$f.x^2$
34	1	34	1156
36	2	72	2592
37	2	74	2738
39	2	78	3042
41	2	82	3362
43	1	43	1849
	$\Sigma f = 10$	$\Sigma fx = 383$	$\Sigma fx^2 = 14739$

There, $\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{383}{10} = 38.3$

$\therefore \quad \overline{X}^2 = 1466.9$

Since $S = \sqrt{\frac{\Sigma f X^2}{\Sigma f} - \overline{X}^2}$

$$= \sqrt{\left(\frac{14739}{10} - 1466.9\right)}$$

$$= \sqrt{1473.9 - 1466.9}$$

$$= \sqrt{7}$$

= 2.65 or 2.7 **Ans .**

In this population of the Patella vulgaris the mean maximum basal diameter of the shell is 38.3 mm, with a standard deviation of 2.7 mm (correct to one decimal place). If these values are applied to a larger population of the Patella then it may be assumed, on statistical grounds, that approximately 68% of the population will have a basal diameter of the shell of 38.3 mm plus and minus 1 standard deviation (2.7 mm), that is they will lie within a range 35.6-41.0 mm; approximately 95% of the population will have a basal diameter of the shell of 38.3 plus and minus 2 Standard deviation (5.4 mm), that is they will lie within the range 32.9-43.7 mm, and practically 100% will lie within plus and minus 3 Standard deviations.

Computation of Standard deviation for grouped data

(Continuous series)

Worked example : Ovary wt. of 50 fishes and their frequency is given in class interval. Find Standard deviation.

Wt. of ovary	2-2.9	3-3.9	4-4.9	5-5.9	6-6.9
Frequency	6	13	11	8	12

Solution : It is computed by following six steps :

Step 1 : Find mind point (m) of each class interval.

Step 2 : Find mean of the series applying formula $\overline{X} = \frac{\Sigma f.X}{\Sigma f}$

Step 3 : Find deviation of each observation (X- $\overline{X}$).

Step 4 : Square each deviation.

Step 5 : Multiply each squared deviation with their frequency.

Step 6 : Sum all the multiplied value of f and X^2.

No a table of 4 columns is framed.

Table 4.

Class interval	Mid point X	Frequency f	$f.m$	Deviation X- $\overline{X}=x$	Deviation Square x^2	$f.x^2$
2-2.9	2.45	6	14.7	-2.14	4.579	27.47
3-3.9	3.45	13	44.85	-1.14	1.299	16.88
4-4.9	4.45	11	48.95	-0.14	0.019	0.21
5-5.9	5.45	8	43.6	+0.86	0.739	5.91
6-6.9	6.45	12	77.4	+1.86	3.459	41.5
		$\Sigma f = 50$	$\Sigma f.m = 229.5$		$\Sigma x^2 = 10.088$	$\Sigma fx^2 = 91.97$

$$\overline{X} = \frac{\Sigma f.X}{\Sigma f} = \frac{229.5}{50} = 4.59$$

$$S = \frac{\Sigma f.x^2}{\Sigma f} = \sqrt{\frac{91.97}{50}} = 1.35 \textbf{ Ans.}$$

Coefficient of Standard deviation

It is ratio of the Standard Deviation to its arithmetic mean.

Coefficient S $= S/\overline{X}$. Here coefficient $S = \frac{1.35}{4.59} = 0.29$ **Ans.**

Merits and demerits of Standard deviation

Merits :

(1) Standard deviation summarises the deviation of a large distribution from mean in one figure used as a unit of variation.

(2) It indicates whether the variation of difference of an individual from the mean is real or by chance.

(3) Standard Deviation helps in finding the suitable size of sample for valid conclusions.

(4) It helps in calculating the Standard error.

Demerits : Standard deviation gives weight age to only extreme values. The process of squaring deviations and then taking square root involves lengthy calculations.

8.8.3 Standard Error

If we calculate mean for a large number of samples of same size from a population and prepare frequency polygons, we will notice that the standard deviation of the distribution of these means $(\bar{x})$ will be different from the standard deviation of the original population (μ).

So standard error is the measure of reliability of the mea of a data and obtained by the following formula :

$$\text{Standard error } \left(S_{\bar{m}}\right) = \frac{\sqrt{s^2}}{n} = \frac{\text{Varianc}}{\text{Sampl No.}} = \frac{S}{\sqrt{n}}$$

The determination of standard error indicates that if values are distributed about the mean ($\bar{x}$) of the sample will be distributed about the true mean of population (μ) by one standard error two-thirds of the time. It indicates that if in a large population several means are determine, then nearly two-thirds of the mean will deviate about true mean (μ) by one $S_{\bar{m}}$. The lesser the value of S.E., the reliable the mean (See Fig.).

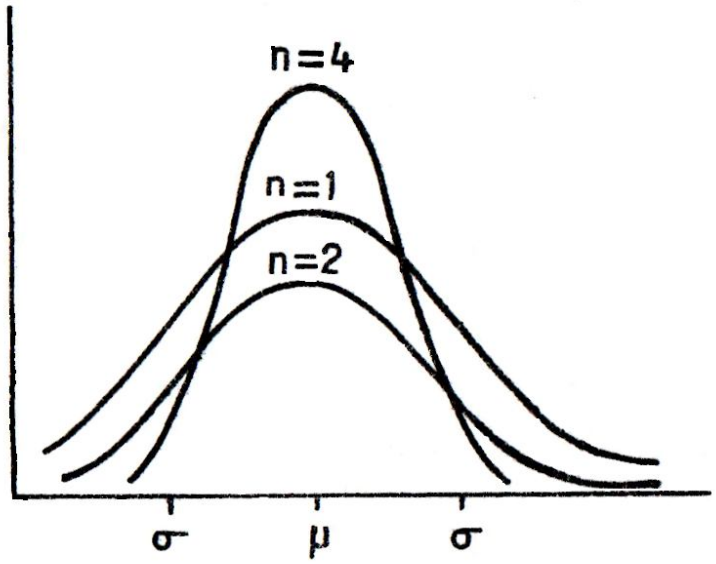


Fig. : Distribution of means of samples of three different size drawn from the same population showing a decrease in a standard error of means with an increase in sample size.

Standard Error of the Difference Between Means

When we compare mean from two samples and determine if they are significantly different, that is whether they came from separate populations or whether there was significant treatment effect.

The standard error of the difference of the means is computed according to the following formula :

$$\text{S.E. } \bar{x}_1 - \bar{x}_2 = \sqrt{\frac{s_1^2}{N_1} - \frac{s_2^2}{N_2}}$$

If the difference between the two means is larger than two times the standard error of the difference, $\text{S.E. } \bar{x}_1 - \bar{x}_2$, they are significantly different.

DEGREE OF FREEDOM

The degrees of freedom are the number of values in a set of data, which are unrestricted, independent and free to vary. For example, if there are 3 values, A, B, and C having mean of 9,

$$\frac{A + B + C}{3} = 9$$

In this case values of A and B may vary, but once they are fixed, they value of C is also fixed. If A =8, B = 9, then the value of C has to be 10 to satisfy the mean. So we can say that three numbers have 2 degrees of freedom. Degrees of freedom are normally expressed as *n*-1, unless specified.

Most of the statistical works in biology in carried out by taking a small sample from a large population. The inferences are then extrapolated to the whole population and generalised. However in actual the statistical measures of the sample may differ from their population. The population statistics are in most cases unknown and unknowable (generally speaking), therefore we have to rely only on sample statistics. It is possible, however to some extent to overcome this situation by using *n*-1 degrees of freedom rather than using total number (*n).*

Confidence Limits

It is possible that the mean obtained for a sample may not be the true value of the actual mean of the population. To express this uncertainty, confidence limits are assigned to the observed mean $(\bar{x})$. Most customarily in Zoology, and Botany, to get satisfactory results, the value of confidence limits in chosen as 95%. It means that the observed mean will enclose the true mean with the frequency of this confidence limit or we can say, that there are 95% chances of the true mean being

present some where in the range of 95% confidence limit values. Confidence limit are obtained by the following formula :

95% confidence limits = $\bar{x} \pm (t \times$ standard error) value of *'t'* can be obtained from the *'t'* distribution table by entering *n*-1 degrees of freedom for a probability of 0.05 (5%).

It is conventional is statistics to use Greek letters for population parameters and Roman letters for sample statistics.

Statistics	***Population***	***Sample***
Mean	μ	$\bar{x}$
Standard deviation	σ	S
Variance	σ^2	S^2

8.8.4 Variance

Definition

Variance of a distribution is defined as the square of the Standard Deviation Thus variance $V = (S)^2$

Interpretation of variance

Suppose that we have a sample of only one case, with only one score. Therefore there is no variance or variability. Consider a second individual with his score in the same test or experiment. We now have on difference. Consider a third case and we then have two additional differences, three altogether. There are as many differences as there are possible pairs of individuals. We could compute all these interpair differences and could average them to get a single, representative value. We could also square them and then average them. It is most easy to find a mean of all scores and to use that value as a common reference point.

Each difference then becomes a deviation from that reference point, and there are only as many deviations as there are individuals. Either the variance or the SD is a single representative value for all the individual difference when taken from a common reference point.

Worked example : Haemoglobin content in g/100 ml of 10 persons of a locality was recorded as 7, 8, 9, 10, 11, 12, 13, 14, 15, and 15.5. Find out the variance of the India.

Variance or $(S^2) = \left(\frac{\Sigma f . X^2}{\Sigma f} - \overline{X}^2 \right)^2$

Variance is useful in ecological investigations including nutrition, reproduction and behaviour since it gives an indication how organisms are dispersed within the population.

Population dispersion = $\frac{\text{Variance}}{\text{Mean}}$ or $\frac{S^2}{\overline{X}}$

Solution : Following table is prepared to calculate variance using formula.

$$V = \Sigma x^2 / N$$

Table 5.

Hb%	7	8	9	10	11	12	13	14	15	15.5
Deviation	7-11.5	8-11.5	9-11.5	10-11.5	11-11.5	12-11.5	13-11.5	14.11.5	15.11.5	15.5-11.5
X	=-4.5	=-3.5	=-2.5	=-1.5	=-0.5	=0.5	=1.5	=3.0	=3.5	= 4
x^2	20.25	12.25	6.25	2.25	0.25	0.25	0.25	9.0	12.25	16.0

$\Sigma X = 115$ & $\overline{X}$ = 115/10=11.5; $\Sigma X^2 = 80.75$

Variance = $\frac{\Sigma x^2}{N}$ = 80.75/10 = 8.075 **Ans.**

Coefficient of variance

The Standard deviation (S) is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variance. It is used in such problems where we want to compare the variability of two or more than two series. If, for an example, an analysis of seed number per fruit in two batches 10 fruits in a garden, batch I has a mean score $\overline{X_2}$ = 80 with SD (S_2) = 2.4, then it is

clear that batch I having a lesser value of SD (S) are more consistent in producing seeds than the batch II.

We also meet situations when two or more distributions having unequal means or different units or measurements are to be compared in respect of their variability. For making such comparisons we use a method of statistics called coefficient of variation. coefficient of variation is denoted by C.V. and is calculated as C.V. = $S / \overline{X} \times 100$.

Worked example : Following are the weights of two rats in 10 (Ten) months. Both were fed same normal diet. If the consistency performance is the criterion calculate which one maintains consistency ?

Rat X	50	45	55	40	47	50	45	40	48	50	$\Sigma X = 470$
Ray Y	55	46	51	45	52	45	50	45	40	47	$\Sigma Y = 476$

Solution : Frame a table of 8 columns to obtain various values to calculate C.V.

Table 6

	X	$\left(X - \overline{X}\right)$ **=*X*=47=*x***	x^2	***Y***	$\left(Y - \overline{Y}\right)$ **=*Y*=47.6=*y***	y^2
1	50	+3	9	55	+7.4	54.76
2	45	-2	4	46	-1.6	2.56
3	55	+8	64	51	3.4	11.56
4	40	-7	49	45	-2.6	6.76
5	47	0	0	52	+4.4	19.36
6	50	+3	9	45	-2.6	6.76
7	45	-2	4	50	+2.4	5.76
8	40	-7	49	45	-2.6	6.76

9	48	+1	1	40	-7.6	57.76
10	50	+3	9	47	-0.6	0.36
N=10	$\Sigma X = 470$	-18+18 $\Sigma x^2 = 0$	$\Sigma x^2 = 198$	$\Sigma Y = 476$	17.0+17.6 $\Sigma y = 0.6$	$\Sigma y^2 = 172.40$

C.V. for the *X* Rat

$$\overline{X} = \frac{\Sigma X}{N} = \frac{470}{10} = 47$$

$$\text{S.D.} = \frac{\Sigma x^2}{N} = \sqrt{\frac{198}{10}} = \sqrt{19.8} = 4.44$$

$$\text{C.V.} = \frac{S.D.}{\overline{X}} \times 100 = \frac{4.44}{47} \times 100 = 9.44$$

Coefficient of variation of the *X* rat = 9.44

C.V. for the Y Rat

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{470}{10} = 47.6$$

$$\text{S.D.} = \frac{\Sigma y^2}{N} = \sqrt{\frac{172.40}{10}} = \sqrt{17.24} = 4.15$$

$$\text{C.V.} = \frac{S.D.}{\overline{X}} \times 100 = \frac{4.15}{47.6} \times 100 = 8.71$$

Coefficient of variation of the *Y* rat = 8.71

Conclusion : Since coefficient of variation is less (8.71) in the case of Y Rat as compared to X Rat (9.44). Therefore, Y is more consistent.

Worked example : An analysis of seed number per fruit in 10 fruits each of two batches is given below.

Find C.V. of both batches and mention which one of the two groups has lower range or variance.

Fruits No.	1	2	3	4	5	6	7	8	9	10	
No. of seeds Batch I X_1	7	9	6	8	6	5	7	8	6	8	$\Sigma X_1 = 70$
No. of seeds Batch II X_2	10	8	9	10	11	10	5	6	4	7	$\Sigma X_2 = 80$

Solution : Here N_1=10 and N_2=10; $\Sigma X_1 = 70$ and $\Sigma X_2 = 80$; $\overline{X_1} = 70/10 = 7$; $\overline{X_2} = 80/10 = 8$ $S_1 = 1.25$ and $S_2 = 2.4$; C.V. $= S/\overline{X_1} \times 100$; $C.V_1 = 1.25/7 \times 100 = 17.8\%$; $C.V_2 = 2.4/8 \times 100 = 30\%$;

Conclusion : Since coefficient of variation is less for fruits of batch I and hence fruit of batch I are more consistent in seed production than batch III.

8.9 Summary

The study of technique of collection of data and its representation enables one to draw reliable conclusions from the collected data, which is obtained through various experiments and help in analysis and interpretation.

Measures of central tendency provide a single figure called average which describes the entire series of observations. Central tendency can be measures mathematically or positionaly. Mathematical average can be calculated by Arithematic mean, Geometric mean and Harmonic mean, and Harmonic mean. Average of position exhibited by Median and Mode. The degree to which the individual values of the variate scatter away from due average in called dispersion. Dispersion is one of one important tool of statistics to calculate variability. Dispersion exhibited through standard deviation, variance and Standard error.

8.10 Self Learning Exercise

1. Define and explain Arithmetic mean, Geometric mean Harmonic mean. Mention merits and demerits of each measure.
2. Mention formula applied to calculate Arithmetic mean for individual data, discrete series and continuous series.
3. Define and explain Median and Mode, Mention formula for individual data and grouped data (discrete and continuous series) to compute Median and Mode.
4. Calculate, Mean and Median from the given ungrouped data :
   (i) 10, 8, 20, 22, 39 and 18. **[Ans.** $\overline{X}$ =19.5, Median= 19]
   (ii) 19, 21,17,16,19,21,23 and 23 **[Ans.** $\overline{X}$ =19.8, Median= 20]
   (iii) 17,19,13,17,13,11 and 21 **[Ans.** $\overline{X}$ =15.8, Median= 17]
   (iv) 15.8, 13.3, 15.2, 13.3, 17.8, 81 and 18.9
   [Ans. $\overline{X}$ =16.5, Median= 15.8]
   (v) 1060, 1060, 1070, 1130, 1370, 1190, 1270 and 1170.
   [Ans. $\overline{X}$ =1165, Median= 1150]
   (vi) No. of animals per cage 3, 7, 8, 11, 13, 15, 16, 17, 18 and 20.
   [Ans. $\overline{X}$ =12.8, Median= 14]

(vii) Haemoglobin percent in g/100 ml 6.0, 6.5, 7.5, 8.2, 8.5, 8.7, 8.8, 8.9, 9 and 9.5 **[Ans.** $\overline{X}$ =8.16, Median= 8.06]

5. Compute N (no. of observation) when $\overline{X}$ =(mean) = 5 and $\Sigma\overline{X}$ =30. (Ans. 6)

6. The following data was obtained in a grassland community. Calculate Mean, Median and Mode in ungrouped and grouped series.
Numbers of seeds per plant (*Indigofera* species) : 39, 55, 35, 45, 49, 52, 48, 33, 48, 47, 50, 51, 53, 50, 55, 54, 53, 50, 48, 49, 31, 33, 50, 55, 51, 50, 53 55, 52, 49, 51, 50, 50, 44, 51, 50, 58, 59, 57, 59, 60, 58, 51. **[Ans.** $\overline{X}$ **=48.87, Median= 52.40]**

7. Calculate Mean, Median and Mode from the data given in following three tables :

Table A.

Class interval	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65
Frequency	4	4	9	7	13	3	3	2	2	3

Table B.

Class interval	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100-109	110-119
Frequency	5	4	9	8	9	10	4	3	2

Table C.

Class interval	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
Frequency	2	1	3	1	7	31	6	7	1	3

Ans.

Table A			**Table B**			**Table C**		
Mean	=	77.24	Mean	=	66.3	Mean	=	77.09
Median	=	36.38	Median	=	75.5	Median	=	77.74
Mode	=	38	Mode	=	84.5	Mode	=	77

8. The frequency distribution according to the age group of persons is as follows. Calculate the Arithmetic mean of their ages.

Groups	5-15	15-25	25-35	35-45	45-55	55-65
No. of persons	4	7	12	9	5	3

[**Ans.** 33.25]

9. In an experiment frogs were observed to prey on insects in unit time. Calculate the average number of insects consumed by a frog.

No. of insects	70-90	90-110	110-130	130-150	150-170	170-190
No. of Frogs	40	55	60	70	100	65

[**Ans.** 136.92]

10. Define mode. Describe various methods of its calculation with suitable example. Find mode of 7, 8, 9, 4, 6, 7, 9, 3, 6, 9, 8, 9, 7, 8, 9 and 11.
(Hint : Prepare frequency distribution table. Mode is that value where maximum frequency falls).

[Ans . 9]

11. (a) Compute the Geometric mean from the following data :

X	13	14	15	16	18
f	4	5	7	6	4

The increase in sales of a poultry farm for five years, each compared to the previous year are given below. Calculate the average percentage of increase.

Increase % : 15, 20, 23, 22 and 10 **[Ans.** 11.79%]

12. Calculate the Harmonic mean of the following frequency distribution :

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	12	18	27	20	17	6

[HM = 17.38]

13. Explain where and how the following calculations are used-(a) Mode (b) Median (c) Arithmetic mean.
14. Define mode, median and mean. Calculate the mean from the given table and arrange it to show the frequency of number.

28, 32, 45, 54, 60, 61, 70, 63, 70, 72, 76, 54, 63, 76, 32, 54, 60, 45, 72, 98. **[Ans.** 58.75]

15. Calculate the Arithmetic mean of the number of leaves and no. of plants.

No. of leaves (X)	5	10	15	20	25
No. of plants (f)	3	5	7	3	2

[HM = 14]

16. Number of tentacles was recorded as 4, 5, 6, 7, 8 in different individuals of Hydra oligacits. Calculate the Geometric mean.

[Hint : Make a table of two columns, one for variable and the other for log x value. Now apply formula GM=Antilog ($\Sigma log\ X/N$)]

[Ans. 5.69]

17. Calculate the Geometric mean of the following data recorded on the length of earthworm.

Length of earthworm-80, 82, 84, 85, 90, 90, 100, 105, 110, 115

[Ans. 92.90]

18. In a segregating F_2 of five mutants of sweets pea ; data obtained on dwarf pants are given below. Calculate Geometric mean.

Dwarf plants	8	12	15	21	35
No. of plants	201	206	3140	390	400

[Hint : Frame a table of 4 columns-1st for variable, 2nd for frequencies, 3rd for log x and 4th for freq. multiplied by log x. Now apply formula G=Antilog ($\Sigma f \log x/\Sigma f$)]

Ans. (18.23]

19. Calculate the Harmonic mean of series – 1, 5, 10, 15, 25

$$HM=\left[\frac{N}{1/X_1+1.X_2+1.X_3+.....+\frac{1}{x_n}}\right]$$ **Ans.** 2.55]

20. Find the Harmonic mean of the following data relating to the weight of ovary of 8 fishes in gm 20.1, 22.0, 18.1, 30.2, 18.1, 24.0, 32.0, 30.0, am respectively.

[Hint :Frame a table of two columns 1st fro variable and the other for reciprocal.

Apply formula $HM=\frac{N}{\Sigma(1/X)}$

Ans. 23.95]

21. Find Median from the following data recorded on number of seeds per pod : 5, 19, 42, 11, 50, 30, 21, 0, 52, 36, 27.

[Ans. 27]

22. Find the Median of the following discrete series.

Variate	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Freq.	2	5	8	9	12	14	14	15	11	13	9	7	4	3

[Hint : Make a table of 3 columns, variable, freq. and cumulative freq.

Median = Size of $\left(\frac{\text{Cum freq.}+1}{2}\right)^{\text{th}}$ item and find the Median]

[Ans. 10]

23. Prepare a grouping table for the following discrete series.

Length of earthworms in cm	50	52	54	56	58	60	62	64	66	68
No. of Earthworms	5	16	20	15	19	25	15	20	10	5

[Hint : If you to make it consult following table and try to understand].

No.	**Height in cms.**	**Columns**					
		1	**2**	**3**	**4**	**5**	**6**
		Maximum Frequency	1+2 3+4 5+6 7+8 9+10	2+3 4+5 6+7 8+9	1+2+3 4+5+6 7+8+9	2+3+4 5+6+7 8+9+10	3+4+5 6+7+8
1	50	5					
2	52	16	21	36	41		
3	54	20	35			51	
4	56	15		34	59		54
5	58	19	44			59	
6	60	25		40			
7	62	15	35				60
8	64	20		30	45	35	
9	66	10	15				

10	68	5

24. In a biological experiment following series was obtained. Find the mode of the series.

Class inter.	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Freq.	5	16	20	15	19	25	15	20	10	5

[Apply formula, mode = $L_1 + \left[\frac{f_1 + f_0}{f_1 - f_0 - f_2} \times i \right]$ **[Ans.** 53.75]

25. Find the Geometric mean of 2 and 8. **[Ans.** 4]

26. Find out the relation between AM, GM and HM.
(Calcutta Univ. Zoology Part II (H) 2003)

27. Find out the median of the following numbers.

(a) 21, 12, 49, 37, 88, 46, 55, 74, 63.

(b) 88, 72, 33, 29, 70, 86, 54, 91, 61,57.
[Ans. (1) 49, (b) 65.2]

28. What is Standard Deviation ? Why S.D. *(S)* is more popular than M.D. (δ) in Biological statistical analysis ?

29. While measuring the length the length of sugarcane plants an agriculturist measured 50 plants with a maximum length of 50 cm. The mean score was 24 and S=10. Let the maximum length of plant raise upto 100 cm, and the scores of each measurement of plant also doubled. Find the Standard deviation of these new scores.

[Ans. 20]

30. Define Standard deviation *(S)* and give its mathematical expression.

31. Define the term Standard deviation with the help of suitable example. Show the method of calculating it. What are its merits and demerits ?

32. Determine the Standard deviation of the first ten natural numbers. **[Ans.** 2.87]

33. The number of eggs laid per year by two birds during last ten years is as follows :

First bird - 12. 115, 6, 73, 7, 19, 119, 36, 84, 29

Second bird - 47, 12, 76, 42, 4, 51, 37, 48, 13, 0.

Which birds is better egg layer and more consistent ?

[Hint : Find Σx^2 of 1st and 2nd bird. Now compute S of both birds. Calculate C.V. of both birds. Lastly conclude. Conclusion-2nd birds lesser value of C.V. and hence she is more constant in laying eggs]

34. Standard deviation was first suggested by .....in 1883.

(1) Garret (2) Guilford

(3) Spearman (4) Karl Person.

[**Ans. 4**]

35. Standard deviation may be defined as :

(1) Difference between the highest value and lowest value in a set of data.

(2) The range of variable between 25th percentile and 75th percentile divided by 2.

(3) The positive square root of the arithmetic mean of the squares of deviations of the observation from the Arithmetic mean.

(4) Ratio of standard deviation (S) of the sample divided by the square root of the total number of observations.

[Ans. 3]

36. Which one of the following formula is used for computation of Standard deviation from ungrouped data (Individual series).

(1) $S = \sqrt{\frac{\Sigma X^2}{N}}$ (2) $S = \sqrt{\frac{\Sigma x^2}{N}}$

(3) $S = \sqrt{\frac{\Sigma f.dx^2}{\Sigma f}}$ (4) None

[Ans. 2]

37. Which one of the following formula is applied for computation of Standard deviation for (discrete series).

(1) $S = \sqrt{\frac{\Sigma f.dx^2}{\Sigma f}}$ (2) $S = \sqrt{\frac{\Sigma x^2}{N}}$

(3) $S = \sqrt{\frac{\Sigma x^2}{N}}$ (4) None

[Ans. 1]

38. Which one of the following formula is used for computation of Standard deviation for grouped (continuous series).

(1) $S = \sqrt{\frac{\Sigma f.dx^2}{\sum f}}$ (2) $S = \sqrt{\frac{\Sigma x^2}{N}}$

(3) $S = \sqrt{\frac{\Sigma x^2}{N}}$ (4) None

[Ans. 1]

39. Which of the following formula is used to obtain coefficient of Standard deviation:

(1) $\frac{S}{X}$ (2) $\frac{S.N}{\overline{x.N}}$

(3) $\frac{\delta}{\sqrt{N}}$ (4) $\frac{\delta.N}{\sqrt{N}}$

[Ans. 1]

40. Standard deviation expressed as a percentage of mean is called :

(1) Coefficient of variation (2) Mean deviation

(3) Standard error (4) None

[Ans. 1]

41. The Standard deviation of the given data 10, 10, 10, 10, 10, is :

(1) 1 (2) 10

(3) 0.1 (4) 0

[Ans. 4]

42. Standard deviation of hypothetical population is denoted by :

(1) δ (2) ρ

(3) X (4) r

[Ans. 1]

43. State whether the following statement is true or false :

The variance in a sample remains unchanged if the value of each observation was decreased.

[Ans. True]

44. Calculate the variance of the data – 7, 3, 4,6; 1, 6, 7, 6, 5

[Ans. 4]

8.11 References

- Amble, V.N. Statistical Methods in Animal Sciences, Indian Society of Agricultural Statistics, New Delhi.
- Bailey, N.T.J. Statistical Methods in Biology, English University Press, London.
- Denenberg, V.H. Statistical and Experimental Design for Behavioral and Biological Researchers, Hemisphere Publication Corp., Washington D.C.
- Snedecor, G.W. and Cochran W.G. Statistical Methods, Oxford and IBH Pub. Co., New Delhi
- Sokal, R.R. and Rohlf, F.J. Biometry : The principles and practice of statistics in biological research, W.H. Freeman and Co., San Francisco.
- S.P. Gupta Statistical method, Published by Sultan Chand & Sons, Thirty Fourth Edition, 2005.
- T.K. Saha Biostatistics in Theory and Practice, Emkay Publication, Delhi

Unit - 9

Sampling Variation of Proportion, Significance difference of proportion & Analysis of Variance

Structure of the Unit

9.0 Objectives

After going through this unit we will be able to understand

- Hypothesis testing, Level of significance, Power of the test, Critical region, One tailed and two tailed tests, Sampling distribution , Standard error
- To learn how to apply test procedure for test of hypotheses concerning Test for Single Proportion , Testing of difference between two Proportions, Test Statistic the difference between two proportions, One-Sample t test, Paired-Samples t test, Independent Samples t test
- To learn how to apply Chi-square tests, Test Statistic for Goodness-of-Fit Tests, Characteristics of the Chi-Square Distribution , Finding Critical Values of the Chi- Square Distribution, Chi-Square Test for Independence, Chi-Square Test for Homogeneity of Proportions
- To learn how to apply Analysis of Variance

9.1 Introduction

In general, we do not know the true value of population parameters - they must be estimated. However, we do have hypotheses about what the true values are. B. The major purpose of hypothesis testing is to choose between two competing hypotheses about the value of a population parameter. For example, one hypothesis might claim that the wages of men and women are equal, while the alternative might claim that men make more than women.

9.2.1 Testing of Hypothesis:

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

Definition

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

9.2.2 Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students"*. Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

9.2.3 Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$
$$(or)\, H_1 : \mu > 100$$
$$(or)\, H_1 : \mu < 100$$

9.2.4 Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error

2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

9.2.5 Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

α = P (Committing Type-I error)

= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

9.2.6 Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$.

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

9.2.7 Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

9.2.8 One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

9.2.9 Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_{n}$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_{n}$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

9.2.10 Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

$$\text{i.e.} \qquad \text{S.E (t)} = \sqrt{Var(t)}$$

9.2.11 Utility of standard error:

It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.

1. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.

2. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
3. It is used to determine the size of the sample.

9.2.12 Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

$$\text{i.e. test statistic } Z = \frac{t - E(t)}{S.E(t)}$$

9.2.13 Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α.
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α.
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

9.2.14 Test for Single Proportion (Large Sample)

Suppose in a sample of size n, x be the number of persons possessing the given attribute.

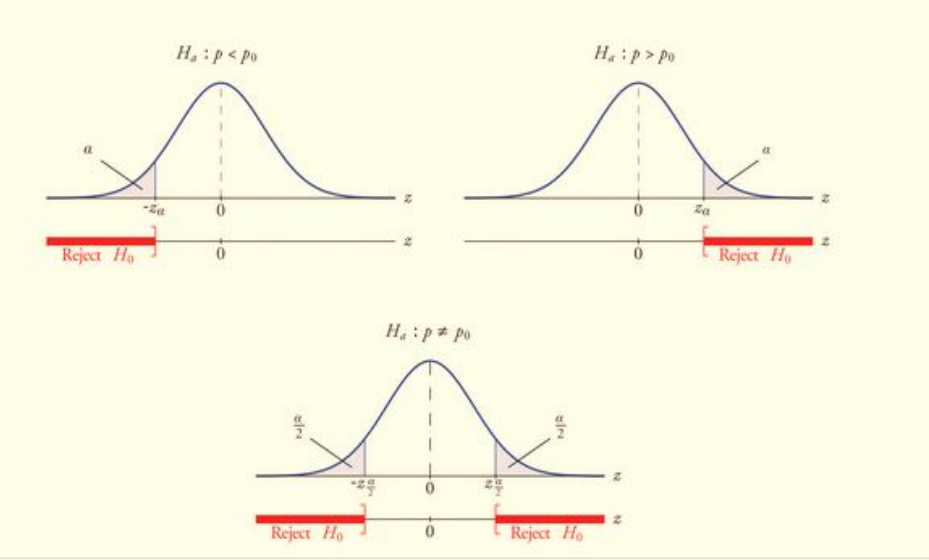
Then observed proportion of successes $= \frac{x}{n}$ =p

E(p) = E $(\frac{x}{n}) = \frac{1}{n} E(x) =$ P (population proportion)

and V(p) = $\frac{PQ}{n}$, Q = 1 - P

The normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{S.E(p)} = \frac{p - P}{PQ/\sqrt{n}} \sim N(0,1)$$



Distribution of the Standardized Test Statistic and the Rejection Region

9.2.15 Example: In a sample of 900 people, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular at 1% level of significance?

Solution: It is given that n = 1000, x = No. of rice eaters = 540,

p=sample proportion of rice eaters $= \frac{540}{1000} = 0.54$

P = Population proportion of rice eaters $= \frac{1}{2} = 0.5$

H_0 : Both rice and wheat are equally popular

$H_1 : P \neq 0.5$

$$Z = \frac{p - P)}{PQ/n} = \frac{0.54 - 0.5}{\sqrt{0.5X0.5/1000}} = 2.532 \sim N(0,1)$$

Since computed Z < 2.58 at 1% level of significance, therefore H_o is not rejected and we conclude that rice and wheat are equally popular.

9.2.16 Test of Significance for the difference between two Proportions (Large Sample)

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

1.2.17 Mean, Variance And Standard Error for the difference between two proportions:

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}} \text{ and } S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

9.2.18 Confidence interval for the difference between two proportions:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

9.2.19 Test of Significance for the difference between two proportions

The null hypothesis is $H_0 : p_1 = p_2$ against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

9.2.20 Critical Values

Critical Values Z_α	level of significance		
	1%	5%	10%
1. Two sided test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
2. Right sided test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
1. Left sided test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Conclusions:

If $\lvert Z \rvert > Z_\alpha$, reject the null hypothesis H_0

If $\lvert Z \rvert < Z_\alpha$, accept the null hypothesis H_0

9.2.21 Example : A sample of 50 randomly selected men with high triglyceride levels consumed 2 tablespoons of oat bran daily for 6 weeks. After 6 weeks, 60% of the men had lowered their triglyceride level. A sample of 80 men consumed 2 tablespoons of wheat bran for 6 weeks. After 6 weeks, 25% had lower triglyceride levels. Is there a significant difference in the two proportions, at the 0.01 significance level?

Solution: Since the statistics are given in percentages, $\hat{p}_1 = 60\%,\ or\ 0.60$ and $\hat{p}_2 = 25\%,\ or\ 0.25$.

To compute , we first find x_1 and x_2

Since $\hat{p}_1 = \frac{x_1}{n_1}$ $x_1 = \hat{p}_1 n_1$ And $x_2 = \hat{p}_2 n_2$

$x_1 = (0.60)(50)=30$ And $x_2 = (0.25)(80) = 20$

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{30+20}{50+80} = \frac{50}{130} = 0.385$$

q = 1-p = 1-0.385=0.615

STEP 1- State the hypotheses and identify the claim. H_0: $p_{1=}$ p2 and H_1: $p_1 = p_2$ (claim)

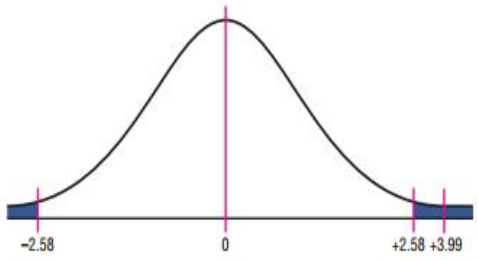
STEP 2- Find the critical values. Since $\alpha = 0.01$, the critical values are +2.58 and -2.58.

STEP 3 Compute the test value.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z=\frac{(0.60-0.25)}{\sqrt{(0.385)(0.615)\left(\frac{1}{50}+\frac{1}{80}\right)}} = 3.99$$

STEP 4 Make the decision. Reject the null hypothesis, since 3.99 > 2.58



STEP 5

Summarize the results. There is enough evidence to support the claim that there is a difference in proportions.

9.2.22 One-Sample t test:

- It is a test to evaluate whether the mean of a test variable is significantly different from a test value.
- For instance, we draw a sample from your class and test whether your class quiz score mean is different from 50, presuming that 100 is the full score.
- Thus, the hypotheses involved are:

  H_0: μ - 50 = 0

  H_A: μ - 50 > 0 or < 0 (one-tailed)

  H_A: μ - 50 $\neq$ 0 (two-tailed)
- In here, only one variable is being tested. Even when the mean score is statistically significant, i.e. significantly different from 50, how can we access the effect, a measure similar to the measures of association?
- We can use d to evaluate the degree that the mean scores on the test variable differ from the test value in standard deviation unit. The formula is:

$d = t/\sqrt{N}$ where t is the t value and N is the sample size.

- t is calculated from the formula:

$t = \frac{\bar{x} - a}{s/\sqrt{N}}$ where $\bar{x}$ is the sample mean, s is the sample standard deviation, and a is the tested value with sample size N

- by substituting t into d, it is found that d also equals to "mean difference divided by standard deviation" or $d = \frac{\bar{x} - a}{s}$
- d values of .2, .5 and .8, regardless of sign, are by convention interpreted as small, medium, and large effect sizes respectively.
- Note that the shape of the t distribution depends on degree of freedom. df = N - 1

9.2.23 Paired-Samples t test:

- Each case must have scores on two variables for a paired-sample t test. The paired-sample t test evaluates whether the mean of the difference between these two variables is significantly different from zero.
- Usually, it is used for a repeated-measures design in which a participant is assessed on two occasions or under two different conditions on one measure.
- Or in a match-subjects design, participants are paired and each participant in a pair is assessed once on a measure.
- A commonly acceptable sample size to yield accurate p values is 30 pairs of scores. Larger sample sizes may be required to produce relatively valid p values if the population distribution is substantially non-normal.
- An example: after the first quiz in Social Research, I deliberately scare you about the "poor" results so as to motivate you to do the second quiz better. I want to know whether there is a scaring effect by testing the first and second quiz scores for each of you.
- The hypotheses are:

  H_0: $\mu_{difference} = 0$

  H_A: $\mu_{difference} > 0$ or < 0 (one-tailed)

H_A: $\mu_{difference} \neq 0$ (two-tailed)

- To calculate sample mean difference ($\bar{x}_{difference}$), one needs to calculate the difference between a pair of scores, sum the differences up, and then take the average.
- t is calculated as follow:

$$t = \frac{\bar{x}_{difference} - 0}{s_{difference} / \sqrt{N}}$$

where $\bar{x}_{difference}$ is the sample mean of score differences, $s_{difference}$ is the sample standard deviation of score differences, and N is the sample size.

- Note that the shape of the t distribution depends on degree of freedom. df = number of pairs – 1
- Again, d is used to assess the effect, which has formula $d = t/\sqrt{N}$

9.2.24 Independent Samples t test:

- The independent-samples t test evaluates the difference between the means of two independent groups.
- The statistical null hypothesis states that the two population means μ_1 and μ_2 are equal. Or put it differently:

  H_0: $\mu_1 - \mu_2 = 0$

  H_A: $\mu_1 - \mu_2 > 0$ or < 0 (one-tailed)

  H_A: $\mu_1 - \mu_2 \neq 0$ (two-tailed)
- Two assumptions of the t test are:

1. the populations sampled are normally distributed
2. the population variances are equal (even when the variances are not equal, SPSS will also give an independent-samples t test without the assumption of equal variances)

- The test statistic t is calculated as follow:
- The numerator of the t is simply the mean difference between the two groups.

- The denominator of it is quite complicated. It should be the common population variance. However, since this parameter is unknown to us.
- Hence, we use the weighted average of the two sample variances of the two groups to estimate the population variance.
- The weights we give to the two groups are simply their respective number of sample cases less one divided by the total number of cases less two. It is because we believe that the more cases a sample has, the larger variance of the sample will be.
- The t-value is computed by the following equation:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{1/n_1 + 1/n_2}} \quad \text{where} \quad s^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}s_2^2$$

- $n_1 + n_2 - 2$ is the degree of freedom of the t-value. The degrees of freedom for a particular sum of squares is equal to the number of terms in the sum we need to know in order to find the remaining terms and thereby complete the sum.
- In SPSS the test of equal variances is known as Levene's test. If the Levene's test is significant, that means the variances are not equal. Then you should look at the row with "equal variance not assumed". Otherwise, look at the other row.

Again, d is used to assess the effect, which has formula $d = t\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$ where N_1 and N_2 are

the sample sizes of the two groups

9.2.25 C

hi-square tests:

Definition: A goodness-of-fit test is an inferential procedure used to determine whether a frequency distribution follows a claimed distribution. It is a test of the agreement or conformity between the observed frequencies (Oi) and the expected frequencies (Ei) for several classes or categories

10.2.26 Test Statistic for Goodness-of-Fit Tests:

$$\chi^2 = \sum \frac{(Observed - \exp ected)^2}{\exp ected}$$

9.2.27 Characteristics of the Chi-Square Distribution :

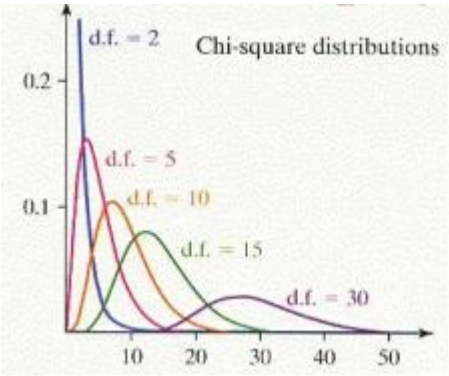
1. It is not symmetric.

2. The values of χ^2 are non-negative (i.e. $\chi^2 > 0$).

3. The chi-square distribution is asymptotic to the horizontal axis on the right-hand-side.

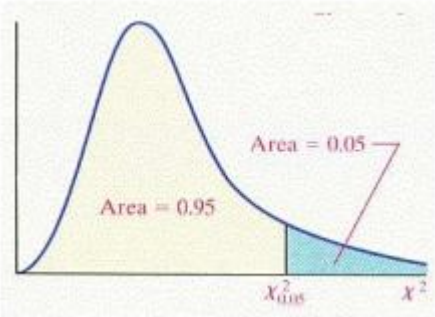
4. The shape of the chi-square distribution depends upon the degrees of freedom, just like Student's t-distribution and Fisher's F-distribution.

5. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetric, as illustrated in Figure 1 below.



6. Total area under the curve is equal to 1.0.

9.2.28 Finding Critical Values of the Chi-Square Distribution: Critical values of the chi-square distribution are found in Table IV in Appendix A. Find the critical value of chi-square for a one-tail (right-tail) test with $\alpha = 0.05$ and df=15.



Degrees of Freedom	Area to the Right of the Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.365	7.215	8.231	9.288	10.265	25.200	28.601	31.595	24.265	36.456

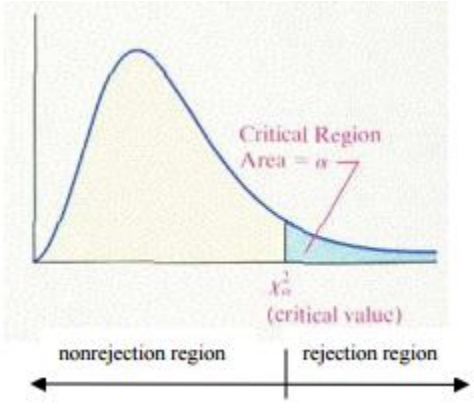
Steps for Chi-Square Goodness-of-Fit Test

Step 1: A claim is made regarding a distribution.

H_0: The random variable follows the claimed distribution.

H_1: The random variable does not follow the claimed distribution.

Step 2: Select a significance level, **α**, and find the critical value of chi-square, χ_2^{α} with df=k-1



Step 3: Calculate the χ^2 test-statistic:

$$\chi^2 = \sum \frac{(Observed - \exp ected)^2}{\exp ected}$$

Step 4: Draw a conclusion.

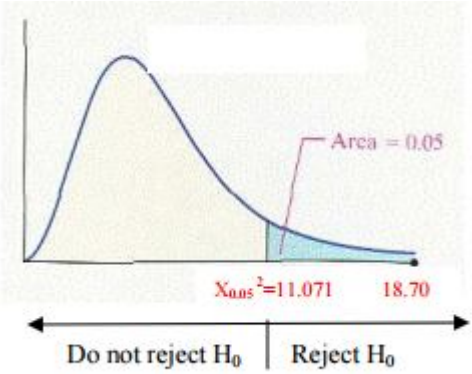
- Compare the test statistic with the critical value. If, $\chi^2 > \chi_2^{\alpha}$ reject H_0.
- Interpret the conclusion in the context of the problem.

Example: A die is tossed 120 times. Test the hypothesis that the die is "fair."

Step 1: Null and alternative hypotheses H_0: Die is "fair" (i.e., $p_1=p_2=\ldots=p_6=1/6$)

H1: Die is not fair (i.e., at least one pi ≠ 1/6)

Step 2: Select α=0.05 and find the critical value of chi-square, $\chi^2_{0.05}$ = 11.071 with df=(6-1)=5.



Step 3: Toss the die 120 times and record the number of 1's, 2's, ..., and 6's. Calculate the expected frequencies using $E_i=np_i$.

No. on Face of Die	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
1	13	20	$\frac{(13-20)^2}{20}=2.45$
2	28	20	$\frac{(28-20)^2}{20}=3.20$
3	16	20	$\frac{(16-20)^2}{20}=0.80$
4	10	20	$\frac{(10-20)^2}{20}=5.00$
5	32	20	$\frac{(32-20)^2}{20}=7.20$
6	21	20	$\frac{(21-20)^2}{20}=0.05$
	120	120	$\sum\frac{(O_i-E_i)^2}{E_i}=18.70$

$E_1=120(1/6)=20$

The H_0 is rejected largely due to the small number of observed 4's (O_4=10) and the large number of observed 5's (O_5=32).

Step 4: Conclusion—Because the χ^2-statistic=18.70 > $\chi_{0.05}^2$ = 11.071, reject H_0 at the 0.05 significance level. The sample data imply that the die is not fair.

9.2.29 Chi-Square Test for Independence

Definition: The chi-square independence test is used to find out whether there is an association between a row variable and column variable in a contingency table constructed from sample data. The null hypothesis is that the variables are not associated: in other words, they are independent. The alternative hypothesis is that the variables are associated, or dependent.

Assumptions

• The data are randomly selected.

• All expected frequencies are greater than or equal to 1 (i.e., Ei>1.)

• No more than 20% of the expected frequencies are less than 5.

Step 1: A claim is made regarding the independence (or dependence) of two variables.

H_0: The row variable and column variable are independent.

H_1: The row variable and column variable are dependent.

Step 2: Select a significance level, α, and find the critical value of chi-square. All chi-square independence tests are right-tailed tests, so the critical value is χ_2^{α} with (r-1)(c-1) degrees of freedom. The shaded region represents the critical region in the figure below

Step 3: Compute the test statistic: $\chi^2 = \sum \frac{(Observed - \exp ected)^2}{\exp ected}$

Step 4: Draw a conclusion.

- Compare the test statistic to the critical chi-square. If $\chi^2 > \chi_2^{\alpha}$, reject H_0.
- Interpret the conclusion in the context of the problem.

9.2.30 Example :Consider the data in the table below that represent the eye color and hair shade of a random sample of 50 individuals.

	Hair Shade	
Eye Color	Light Hair	Dark Hair
Blue Eyes	23	7
Brown Eyes	4	16

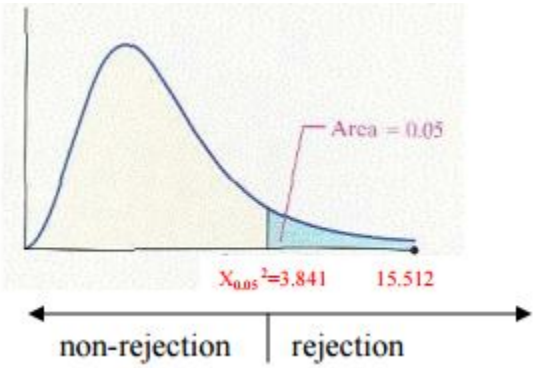
Is eye color and shade of hair related in individuals? Can we conclude from the data shown below that there is a significant connection between eye color and hair shade?

Solution:

Step 1: Null and alternative hypotheses H_0: Eye color and hair shade are independent.

H_1: Eye color and hair shade are dependent.

Step 2: Select **α**=0.05 and find the critical value of χ_2^{α} = 3.841 with df=(2-1)(2-1)=1.



Step 3:
A random sample of 50 individuals was selected and classified according to eye color and shade of hair (shown at the top of the table below). Expected frequencies were calculated using the formula: $Exp.\ freq = \frac{(row\ tot)(col\ tot)}{table\ total}$ (shown at the bottom of the table).

Observed Frequencies:

	Light Hair	Dark Hair	*Row Total*
Blue Eyes	23	7	30
Brown Eyes	4	16	20
Column Total	27	23	50

Expected Frequencies:

	Light Hair	Dark Hair	*Row Total*
Blue Eyes	$\frac{(30)(27)}{50} = 16.2$	$\frac{(30)(23)}{50} = 13.8$	30
Brown Eyes	$\frac{(20)(27)}{50} = 10.8$	$\frac{(20)(23)}{50} = 9.2$	20
Column Total	27	23	50

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(23-16.2)^2}{16.2} + \frac{(7-13.8)^2}{13.8} + \frac{(4-10.8)^2}{10.8} + \frac{(16-9.2)^2}{9.2} = 15.512$$

Step 4: Conclusion—Because the calculated χ^2=15.521 > the critical χ_α^2 = 3.841, reject H_0 at the 0.05 significance level. A significant relation exists between eye color and hair shade (i.e., eye color and hair shade are dependent).

9.2.31 Chi-Square Test for Homogeneity of Proportions

Definition: In a chi-square test for homogeneity of proportions, we test the claim that different populations have the same proportion of individuals with some characteristic.

9.3 Analysis of Variance (ANOVA)

Analysis of Variance which is usually abbreviated as ANOVA is a techniques of decomposing the total variation present in data in terms of variation due to identifiable and interesting sources that causes the variation. . It is a powerful statistical tool for test of significance. In a situation when we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population to find whether the effect of these samples is significantly different or in other words ,

whether the samples have come from the same normal population, then we use techniques of ANOVA.

One-way ANOVA is used to determine whether there are any significant differences between the means of two or more independent (unrelated). For example, we can use one-way ANOVA to understand whether exam performance differed based on test anxiety levels amongst students, dividing students into three independent groups (e.g., low, medium and high-stressed students).

- ANOVA is a basic statistical technique for analyzing experimental data. It subdivides the total variation of a data set into meaningful component parts associated with specific sources of variation in order to test a hypothesis on the parameters of the model or to estimate variance components.
- According to R.A. Fisher, ANOVA is the "Separation of variance ascribable to one group of causes from the variance ascribable to the other group."
- ANOVA consists in the estimation of the amount of variation due to each of the independent factor separately and then comparing these estimates due to assignable factors with the estimate due to chance factors, the latter being known as experimental error or simply error.
- ANOVA is used to test the hypothesis that several means are equal.

9.3.1 Assumptions For Anova Test

- The samples must be independent.
- The variances of the populations must be equal.
- Various treatment and environmental effects are additive in nature.
- The populations from which the samples is obtained must be normally or approximately normally distributed.

9.3.2 One-Way ANOVA:

- In one-way ANOVA, each case must have scores on two variables: a factor (or independent variable) and a dependent variable. The factor divides cases into two or more groups or levels, while the dependent variable differentiates cases

on some quantitative dimension. The ANOVA F test assesses whether the group means on the dependent variable differ significantly from each other.

- If the factor divides cases into two groups, the analysis is essential similar to an independent-samples t test. Thus, you can say that one-way ANOVA is to compare the means of more than two groups.
- One-way ANOVA is based upon the assumption that we can decompose each observation into three additive terms:

  observation = overall mean + deviation of group mean from overall mean + deviation of observation from group mean

- The above equation can be interpreted as follow:

  observed value of Y = constant + effect of being in a particular group (effect of the factor X) + effect of all other variables (residual)

- The summation of "the square of the deviation of each group mean from the overall mean times the number of cases in the corresponding group" is known as "sum of squares between group" (SSB).
- The summation of "the square of the deviation of each observation from the corresponding group mean" is known as "sum of squares within group" (SSW). It is also known as sum of squares residual (RSS).
- The degree of freedom of the SSB equals to the number of groups less one (df_{SSB}).
- The degree of freedom of the SSW equals to the number of cases less the number of groups (df_{SSW}).
- The statistical hypothesis of ANOVA is that all the group means are equal, i.e. there is no difference between groups:

  H_0: $\mu_1=\mu_2=\mu_3=\ldots=\mu_n$

  H_A: any two group means are unequal

- In order to test the above hypothesis, we use the F-test. The F-ratio is computed as:

$$F = \frac{SSB / df_{SSB}}{SSW / df_{SSW}}$$

- The SSB tells us how large the effect of the factor or explanatory variable (i.e. the groups) is on the dependent variable, while the SSW indicates the random variation of the dependent variable due to other uncontrolled variables.
- Put it differently, F-ratio = variation between the means/unexplained (error) variation.

9.3.3 Example: The three groups (6 people in Group 1, and 5 each in Groups 2 and 3).

In this example, we will compute a one-way ANOVA for data from three independent groups.

Here are the raw data from the three groups (6 people in Group 1, and 5 each in Groups 2 and 3).

Group 1	Group 2	Group 3
3	4	9
1	3	7
3	5	8
2	5	11
4	4	9
3		

The first step in the computation is to add the scores in each column and compute the sum of the squared scores for each column. We also count the number of scores in each column, compute the mean by dividing the sum by the number of scores, and compute the sum of squares.

We fined the sum of squares as the average squared deviation from the mean. However, there is an easier computational formula for the SS, which

$$SS = \sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

Summary Statistics

We have done the computations described above for each of the three groups and organized them in three columns. We have also included a fourth column to put the total scores, total sum of X2, and total sample size, all of which will also be needed for the computations. In this way, we have all of the values that we will need for the computation of a one-way ANOVA at our fingertips.

	Group 1	Group 2	Group 3	Totals
Sum of X	16	21	44	81
Sum of X2	48	91	396	535
n	6	5	5	16
Mean	2.67	4.20	8.80	
SS	5.33	2.80	8.80	

Compute the SSs for the ANOVA

The formulas for computing the three sums of squares (between, within, and Total) are shown below, with the numbers plugged in. The notation may look complicated, but all of the needed values can be found in the summary table that we just prepared. The only real new terminology uses summation notation in which we sum across the groups (i, which refers to the group number, goes from 1 to k, which is the number of groups). We use this notation because we can have any number of groups in a design like this and we want a formula that will describe what we should do regardless of how many groups we have.

Most students find it easier to understand the notation by looking at the formula and see where the numbers for the formula can be found in the summary table above. We have used more parentheses than actually needed algebraically to specify what needs to be done. The rule is that you always do things inside a parenthesis before you do things outside of the parenthesis. If you remember that simple rule, you will not have to remember the more complicated algebraic rules about what computations should be done first.

$$SS_b = \left\{ \sum_{i=1}^{k} \frac{\left(\left(\sum X\right)_i\right)^2}{n_i} \right\} - \frac{\left(\left(\sum X\right)_T\right)^2}{N}$$

$$SS_b = \left\{ \frac{16^2}{6} + \frac{21^2}{5} + \frac{44^2}{5} \right\} - \frac{81^2}{16} = 108.00$$

$$SS_w = \sum_{i=1}^{k} SS_i = 5.33 + 2.80 + 8.80 = 16.93$$

$$SS_T = \left(\sum X^2\right)_T - \frac{\left(\left(\sum X\right)_T\right)^2}{N} = 535 - \frac{81^2}{16} = 124.94$$

Double check the computation of the SSs by seeing if they add up. SST = SSb + SSw = 98.00 + 16.93 = 124.93 [There is a slight difference here due to rounding error.]

Fill in the Summary Table

The dfb is equal to the number of groups (k) minus 1. The dfw is equal to the total number of participants minus the number of groups (N -k). The dfT is equal to the total number of participants (N) minus 1. Note that the dfT is equal to the dfb plus the dfw in the same way that the SST is equal to the sum of the SSb and SSw.

The MSs are computed by dividing the SSs by their respective dfs, and the F is computed by dividing the MSb by the MSw. All of these values have been inserted into the standard summary table for ANOVA below.

Source	df	SS	MS	F
Between	2	108.00	54.00	41.46
Within	13	16.93	.30	
Total	15	124.94		

The final step is to compare the value of the F computed in this analysis with the critical value of F in the F Table. You look up the critical value by using the degrees of freedom. In our case, the dfb is 2 and the dfw is 13. The critical value of F for an alpha of .05 is 3.80. Since our obtained F exceeds this value, we reject the null hypothesis and conclude that there is a significant difference between the groups.

9.4 Glossary

- **Alternative hypothesis (H_1)** - The opposite of the null hypothesis, declaring a non-chance difference.
- **Alpha (α)** - The probability the researcher is willing to take of falsely rejecting an incorrect null hypothesis. In fixed-level testing, this serves as the cutoff point for making decisions about H0 .
- **Null hypothesis (H_0)** - A statement that declares that the observed difference is due to unexplained "chance." It is the hypothesis the researcher hopes to reject
- **Test statistic** - A statistic used to test the null hypothesis.
- **p value -** A probability statement that answers the question "If the null hypothesis were true, what is the probability of observing the current

data or data that is more extreme than the current data?." It is the probability of the data conditional on the truth of H0 . It is NOT the probability that the null hypothesis is true.

- **Type I error** - a rejection of a true null hypothesis; a "false alarm."
- **Type II error** - a retention of an incorrect null hypothesis; "failure to sound the alarm."
- **Confidence (1 - α)** - the complement of alpha; the probability of correctly retaining a true null hypothesis.
- **Beta (β)** - the probability of a type II error; probability of a retaining a false null hypothesis.
- **Power** (1 - β) - the complement of b; the probability of avoiding a type II error; the probability of rejecting a false null hypothesis.

9.5 Self-Learning Exercise

Section -A (Very Short Answer Type):

1. The hypothesis under verification is known as.....

2. Null hypothesis and is denoted by.....

3. level of significance is denoted by.....

4. Power of test is denoted by .....

5. Standard error for sample mean $\bar{x}$ is .....

6. Standard error for difference of two sample means is.....

7 In Various treatment and environmental effects are …….in nature

8. In ANOVA we use …… test

9. …… test is used for goodness of fit.

10. Test statistic for chi square is……

Section -B (Short Answer Type) :

1. What is Hypothesis?

2. What is type-I error& type-II error?

3. What is test statistic & level of significance ?

4. Define One tailed and two tailed tests with examples.

5. What is Critical region ?

6. What is Confidence interval for the difference between two proportions. ?

7. Define Independent Samples t test.

8. Define One-Sample t test.

9. Define Paired-Samples t test

10. Define Independent Samples t test

12. Define ANOVA?

Section -C (Long Answer Type)

1. What is meant by a statistical hypothesis ?Explain the concept of type-I error& type- II error , level of significance & critical region.

2. Define Test of Significance for the difference between two Proportions.. Also find Mean, Variance , Standard Error and Confidence Interval

3. Define Chi-square tests with its characteristics . And also Find its Critical Values

6. Explain the meaning of "Analysis of Variance" and give its uses. State the basic assumptions in the Analysis of Variance

Answer Key of Section-A

1. Null Hypothesis

2. H_0

3. α

4. $(1 - \beta)$

5. $\frac{\sigma}{\sqrt{n}}$

6. $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

7. additive

8. F

9. Chi-Square

10. $\chi^2 = \sum \frac{(Observed - \exp ected)^2}{\exp ected}$

11.Rare

12.Poisson

13. $p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,3.....\infty$

14. $e^{-\lambda(e^{it}-1)}$

15. $e^{-\lambda(e^{t}-1)}$

16. $x = \mu$

17.Coincide

18. $\beta_1 = 0$

19. $\beta_2 = 3$

20. $-\infty \; to \; \infty$

9.6 References

- Gupta, S.C. and Kapoor, V.K. (2007): Fundamentals of Applied Statistics, 4th Revised Edn., (Reprint), Sultan Chand and Sons.
- Joshi D.D. Linear Estimation and design of Experiments, Wily Eastern Ltd.
- Das M.N, Giri N.C, Design and Analysis of Experiments, Wily Eastern Ltd.
- V.Rajgopalan Selected l Statistical Tests, New Age International (P) Ltd.

- B.L. Agarwal, Programmed Statistics , New Age International Publishers Ltd.
- http://www.aaec.ttu.edu
- Mukhopadhyay, P : Mathmatical Statistics, new central book agency.

Unit - 10

Student's t test, Chi-square test. Correlation and regression

Structure of Unit:

10.1. Objectives

10.2. Basic Terms

10.3. Student's t test

10.3.1. Introduction

10.3.2. Uses

10.3.3. Unpaired and paired two-sample *t*-tests

a) Independent (unpaired) samples

b) Paired samples

10.3.4. Calculations

10.3.4.1. Independent two-sample *t*-test

a) Equal sample sizes, equal variance

b) Equal or unequal sample sizes, equal variance

c) Equal or unequal sample sizes, unequal variances

10.3.5. Explanations

a) One sample t test

b) Paired sample t test

c) Independent samples t test

d) Students t Test (Independent samples)

10.4. Chi-square test

10.4.1. Introduction

10.4.2. Explanation

10.5. Correlation and regression

10.5.1. Introduction

10.5.2. Types of Correlation

a) Positive and Negative Correlation

b) Linear and Non – Linear Correlation

10.1. Objectives

After completing this unit we will be able to understand:

- Some basic statistical terms
- Various types and applications of Student's t test.
- Chi-square test and its uses.
- correlation analysis and its types
- regression analysis and its types
- interpretation and conduction of basic regression analysis.

In any study, statistics play a major role in understanding the significance of the matter being investigated or studied. In any aspect there are variables and hypotheses that we need to find out their significance by using statistics and this is very important in biological as well as in other branches of science. In Zoology statistic help in analyzing the behaviour of the animal, it's population study, its physiological change or adaptation as well as any other aspects regarding the environment or the ecology which has an impact (positive or negative) on the individual.

10.2. Basic Terms

- **Population:**

  The group of individuals, under study is called is called population.

- **Sample**:

  A finite subset of statistical individuals in a population is called Sample.

- **Sample size**:

  The number of individuals in a sample is called the Sample size.

- **Parameters and Statistics:**

  The statistical constants of the population are referred as Parameters and the statistical constants of the Sample are referred as Statistics.

- **Standard Error** :

  The standard deviation of sampling distribution of a statistic is known as its standard error and is denoted by (S.E)

- **Test of Significance** :

  It enable us to decide on the basis of the sample results if the deviation between the observed sample statistic and the hypothetical parameter value is significant or the deviation between two sample statistics is significant.

- **Null Hypothesis**:

  A definite statement about the population parameter which is usually a hypothesis of no-difference and is denoted by H_o.

- **Alternative Hypothesis**:

  Any hypothesis which is complementary to the null hypothesis is called an Alternative Hypothesis and is denoted by H_1.

- **Level of Significance** :

  The probability α that a random value of the statistic "t" belongs to the critical region is known as the level of significance. In otherwords the level of significance is the size of the type I error. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

- **Types of samples :**

  Small sample and Large sample

  Small sample ($n \leq 30$) : "Students t test, F test ,
  Chi Square test Large sample ($n > 30$) : Z test.

- **Application of t – distribution**

  When the size of the sample is less than 30, ' t ' test is used in (a)

single mean and (b) difference of two means.

- **Distinguish between parameters and statistics.**

  Statistical constant of the population are usually referred to as parameters. Statistical measures computed from sample observations alone are usually referred to as statistic.

  In practice, parameter values are not known and their estimates based

- **Critical value.**

  The critical or rejection region is the region which corresponds to a predetermined level of significance a. Whenever the sample statistic falls in the critical region we reject the null hypothesis as it will be considered to be probably false. The value that separates the rejection region from the acceptance region is called the critical value.

- **Level of significance**

  The probability a that a random value of the statistic 't' belongs to the critical region is known as the level of significance. In other words level of significance is the size of type I error. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

10.3. Student's t test

A *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t* distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's *t* distribution. The t-test, is a test of significance that can be used to determine whether a significant difference exists or does not exist between two groups. A t-test helps us to compare whether two groups have different average values (for example, whether men and women have different average heights).

10.3.1. Introduction

The *t*-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). Gosset devised the *t*-test as a cheap way to monitor the quality of stout. The Student's *t*-test work was submitted to and accepted in the journal *Biometrika* and published in 1908.

10.3.2. Uses

Among the most frequently used *t*-tests are:

- A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
- A two-sample location test of the null hypothesis such that the means of two populations are equal. All such tests are usually called Student's *t*-tests, (name should only be used if the variances of the two populations are also assumed to be equal) the form of the test used when this assumption is dropped is sometimes called Welch's *t*-test. These tests are often referred to as "unpaired" or "independent samples" *t*-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.
- A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired" or "repeated measures" *t*-test. A test of whether the slope of a regression line differs significantly from 0.

10.3.3. Unpaired and paired two-sample *t*-tests

Two-sample *t*-tests for a difference in mean involve independent samples or paired samples. Paired *t*-tests are a form of blocking, and have greater power than unpaired tests when the paired units are similar with respect to "noise factors" that are independent of membership in the two groups being compared. In a different context, paired *t*-tests can be used to reduce the effects of confounding factors in an observational study.

a) Independent (unpaired) samples

The independent samples *t*-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomly assign 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the *t*-test. The randomization is not essential here – if we contacted 100 people by phone and obtained each person's age and gender, and then used a two-sample *t*-test to see whether the mean ages differ by gender, this would also be an independent samples *t*-test, even though the data are observational.

b) Paired samples

Paired samples *t*-tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" *t*-test). A typical example of the repeated measures *t*-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control. That way the correct rejection of the null hypothesis can become much more likely, with statistical power increasing simply because the random between-patient variation has now been eliminated. Note however that an increase of statistical power comes at a price: more tests are required, each subject having to be tested twice. Because half of the sample now depends on the other half, the paired version of Student's *t*-test has only "n/2–1" degrees of freedom (with *n* being the total number of observations). Pairs become individual test units, and the sample has to be doubled to achieve the same number of degrees of freedom.

A paired samples *t*-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors.

Paired samples *t*-tests are often referred to as "dependent samples *t*-tests".

10.3.4. Calculations

General expressions that can be used to carry out various *t*-tests (Independent two-sample *t*-test) are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a *t*-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out t test.

Once a *t* value is determined, a *p*-value can be found using a table of values from Student's *t*-distribution. If the calculated *p*-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis is rejected in favor of the alternative hypothesis.

10.3.4.1. Independent two-sample *t*-test

a) Equal sample sizes, equal variance

This test is only used when both:

- The two sample sizes (that is, the number, *n*, of participants of each group) are equal;
- It can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{1}{2}(s_{X_1}^2 + s_{X_2}^2)}$$

Here $s_{X_1X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of *t* is the standard error of the difference between two means.

For significance testing, the degree of freedom for this test is $2n - 2$ where *n* is the number of participants in each group.

b) Equal or unequal sample sizes, equal variance

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated) The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}.$$

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute n for n_1 and n_2).

$s_{X_1X_2}$ is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, n = number of participants, 1 = group one, 2 = group two. $n - 1$ is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.

c) Equal or unequal sample sizes, unequal variances

This test, also known as Welch's *t*-test, is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal) and hence must be estimated separately. The *t* statistic to test whether the population means are different is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Here s^2 is the unbiased estimator of the variance of the two samples, n_i = number of participants in group i, i=1 or 2. Note that in this case $s_{\overline{X}_1 - \overline{X}_2}^2$ is not a pooled

variance. For use in significance testing, the distribution of the test statistic is approximated as an ordinary Student's *t* distribution with the degrees of freedom calculated using

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

This is known as the Welch–Satterthwaite equation. The true distribution of the test statistic actually depends (slightly) on the two unknown population variances

10.3.5. Explanations

a) One-Sample t test

- It is a test to evaluate whether the mean of a test variable is significantly different from a test value.
- For instance, we draw a sample from our class and test whether our class quiz score mean is different from 50, presuming that 100 is the full score.
- Thus, the hypotheses involved are:

  H_0: μ - 50 = 0

  H_A: μ - 50 > 0 or < 0 (one-tailed)

  H_A: μ - 50 $\neq$ 0 (two-tailed)

- In here, only one variable is being tested. Even when the mean score is statistically significant, i.e. significantly different from 50, how can we access the effect, a measure similar to the measures of association?
- We can use d to evaluate the degree that the mean scores on the test variable differ from the test value in standard deviation unit. The formula is:

  $d = t/\sqrt{N}$ where t is the t value and N is the sample size.

- t is calculated from the formula:

  $t = \frac{\bar{x} - a}{s/\sqrt{N}}$ where $\bar{x}$ is the sample mean, s is the sample standard deviation, and a is the tested value with sample size N

- by substituting t into d, it is found that d also equals to "mean difference divided by standard deviation" or $d = \frac{\bar{x} - a}{s}$

- d values of .2, .5 and .8, regardless of sign, are by convention interpreted as small, medium, and large effect sizes respectively.
- Note that the shape of the t distribution depends on degree of freedom.

df = N - 1

b) Paired-Samples t test

- Each case must have scores on two variables for a paired-sample t test. The paired-sample t test evaluates whether the mean of the difference between these two variables is significantly different from zero.
- Usually, it is used for a repeated-measures design in which a participant is assessed on two occasions or under two different conditions on one measure.
- Or in a match-subjects design, participants are paired and each participant in a pair is assessed once on a measure.
- A commonly acceptable sample size to yield accurate p values is 30 pairs of scores. Larger sample sizes may be required to produce relatively valid p values if the population distribution is substantially non-normal.
- An example: after the first quiz in Social Research, teacher deliberately scare you about the "poor" results so as to motivate you to do the second quiz better. Teacher wants to know whether there is a scaring effect by testing the first and second quiz scores for each of you.
- The hypotheses are:

  H_0: $\mu_{difference} = 0$

  H_A: $\mu_{difference} > 0$ or < 0 (one-tailed)

  H_A: $\mu_{difference} \neq 0$ (two-tailed)
- To calculate sample mean difference ($\bar{x}_{difference}$), one needs to calculate the difference between a pair of scores, sum the differences up, and then take the average.
- t is calculated as follow:

$$t = \frac{\bar{x}_{difference} - 0}{s_{difference} / \sqrt{N}}$$

where $\bar{x}_{difference}$ is the sample mean of score differences, $s_{difference}$ is the sample standard deviation of score

differences, and N is the sample size.

- Note that the shape of the t distribution depends on degree of freedom. df = number of pairs – 1
- Again, d is used to assess the effect, which has formula $d = t/\sqrt{N}$

c) Independent Samples t test

- The independent-samples t test evaluates the difference between the means of two independent groups.
- The statistical null hypothesis states that the two population means μ_1 and μ_2 are equal. Or put it differently:

  H_0: $\mu_1 - \mu_2 = 0$

  H_A: $\mu_1 - \mu_2 > 0$ or < 0 (one-tailed)

  H_A: $\mu_1 - \mu_2 \neq 0$ (two-tailed)
- Two assumptions of the t test are:

3. the populations sampled are normally distributed
- the population variances are equal
- The test statistic t is calculated as follow:
- The numerator of the t is simply the mean difference between the two groups.
- The denominator of it is quite complicated. It should be the common population variance. However, since this parameter is unknown to us.
- Hence, we use the weighted average of the two sample variances of the two groups to estimate the population variance.
- The weights we give to the two groups are simply their respective number of sample cases less one divided by the total number of cases less two. It is because we believe that the more cases a sample has, the larger variance of the sample will be.
- The t-value is computed by the following equation:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{1/n_1 + 1/n_2}} \quad \text{where} \quad s^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}s_2^2$$

- $n_1 + n_2 - 2$ is the degree of freedom of the t-value. The degrees of freedom for a particular sum of squares is equal to the number of terms in the sum we need to know in order to find the remaining terms and thereby complete the sum.
- Again, d is used to assess the effect, which has formula $d = t\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$ where N_1 and N_2 are the sample sizes of the two groups.

d) 'Student's' t Test (For Independent Samples)

We can use this test to compare two small sets of quantitative data when samples are collected independently of one another. When one randomly takes replicate measurements from a population he/she is collecting an independent sample. Use of a paired t test, to which some statistics programs unfortunately default, requires nonrandom sampling.

Criteria

- *Only* if there is a direct relationship between each specific data point in the first set and one and only one specific data point in the second set, such as measurements on the same subject 'before and after,' then the paired t test may be appropriate.
- If samples are collected from two different populations or from randomly selected individuals from the same population at different times, use the test for independent samples (unpaired).
- Here's a simple check to determine if the paired t test can apply - if one sample can have a different number of data points from the other, then the paired t test *cannot* apply.

'Student's' t Test is one of the most commonly used techniques for testing a hypothesis on the basis of a difference between sample means. Explained in layman's terms, the t test determines a probability that two populations are the same with respect to the variable tested.

For example, suppose we collected data on the heights of male basketball and football players, and compared the sample means using the t test. A probability of 0.4 would mean that there is a 40% liklihood that we cannot distinguish a group of basketball players from a group of football players by height alone. That's about as far as the t test or any statistical test, for that matter, can take us. If we calculate a

probability of 0.05 or less, then we can reject the null hypothesis (that is, we can conclude that the two groups of athletes can be distinguished by height.

Leaves were collected from wax-leaf ligustrum grown in shade and in full sun. The thickness in micrometers of the palisade layer was recorded for each type of leaf. Thicknesses of 7 sun leaves were reported as: 150, 100, 210, 300, 200, 210, and 300, respectively. Thicknesses of 7 shade leaves were reported as 120, 125, 160, 130, 200, 170, and 200, respectively. The mean ± standard deviation for sun leaves was 210 ± 73 micrometers and for shade leaves it was158 ± 34 micrometers. Note that since all data were rounded to the nearest micrometer, it is inappropriate to include decimal places in either the mean or standard deviation.

For the t test for independent samples we do not have to have the same number of data points in each group. We have to assume that the population follows a normal distribution (small samples have more scatter and follow what is called a t distribution). Corrections can be made for groups that do not show a normal distribution (skewed samples).

The t test can be performed knowing just the means, standard deviation, and number of data points. Note that the raw data must be used for the t test or any statistical test, for that matter. The two sample t test yields a statistic t, in which

$$t = |\bar{x}_1 - \bar{x}_2| \div \sqrt{A * B}$$

Where,

$$A = (n_1 + n_2) \div n_1 n_2 ,$$

and

$$B = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] \div [n_1 + n_2 - 2]$$

X-bar, of course, is the sample mean, and s is the sample standard deviation. Note that the numerator of the formula is the difference between means. The denominator is a measurement of experimental error in the two groups combined. The wider the difference between means, the more confident you are in the data. The more experimental error you have, the less confident you are in the data. Thus the higher the value of t, the greater the confidence that there is a difference.

To understand how a precise probability value can be attached to that confidence we need to study the mathematics behind the t distribution in a formal statistics course. The value t is just an intermediate statistic. Probability tables have been prepared based on the t distribution originally worked out by W.S. Gossett. To use the table provided, find the critical value that corrresponds to the number of degrees of freedom we have (degrees of freedom = number of data points in the two groups combined, minus 2). If t exceeds the tabled value, the means are significantly different at the probability level that is listed. When using tables report the lowest probability value for which t exceeds the critical value. Report as 'p < (probability value).'

In the example, the difference between means is 52, A = 14/49, and B = 3242.5. Then t = 1.71 (rounding up). There are (7 + 7 -2) = 12 degrees of freedom, so the critical value for p = 0.05 is 2.18. 1.71 is less than 2.18, so we cannot reject the null hypothesis that the two populations have the same palisade layer thickness. So now we might collect more data. With a well designed experiment, sufficient data can overcome the uncertainty contributed by experimental error, and yield a significant difference between samples, if one exists.

10.4. Chi-square test

10.4.1. Introduction

A **chi-square test**, also referred to as χ^2 **test** , is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. Also considered a chi-square test is a test in which this is *asymptotically* true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the number of individuals or objects that fall in each category differ significantly from the number we would expect? Is this difference between the expected and observed due to sampling variation, or is it a real difference?

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if,

according to Mendel's laws, we expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then we might want to know about the "goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before us, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what scientists call the **null hypothesis,** which states that there is no significant difference between the expected and observed result.

The formula for calculating chi-square (χ^2) is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

That is, chi-square is the sum of the squared difference between observed (*O*) and the expected (*E*) data (or the deviation, *d*), divided by the expected data in all possible categories.

10.4.2. Explanation

Suppose that a cross between two pea plants yields a population of 880 plants, 639 with green seeds and 241 with yellow seeds. We are asked to propose the genotypes of the parents. Our *hypothesis* is that the allele for green is dominant to the allele for yellow and that the parent plants were both heterozygous for this trait. If our hypothesis is true, then the predicted ratio of offspring from this cross would be 3:1 (based on Mendel's laws) as predicted from the results of the Punnett square .

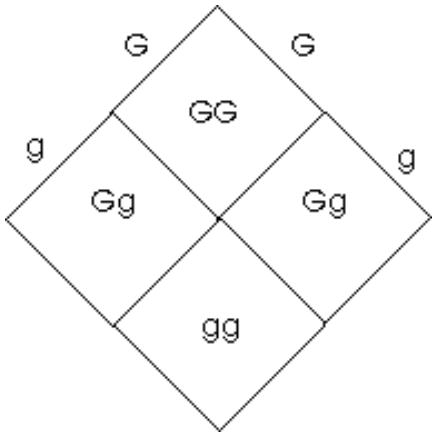


Figure - Punnett Square. Predicted offspring from cross between green and yellow-seeded plants. Green (G) is dominant (3/4 green; 1/4 yellow). To calculate χ^2, first determine the number *expected* in each category. If the ratio is 3:1 and the total number of observed individuals is 880, then the *expected numerical values* should be 660 green and 220 yellow.

Then calculate χ^2 using this formula, as shown in Table B. Note that we get a value of 2.668 for χ^2. But what does this number mean? Here's how to interpret the χ^2 value:

1. Determine degrees of freedom (df). Degrees of freedom can be calculated as the number of categories in the problem minus 1. In our example, there are two categories (green and yellow); therefore, there is I degree of freedom.

2. Determine a relative standard to serve as the basis for accepting or rejecting the hypothesis. The relative standard commonly used in biological research is $p > 0.05$. The p value is the *probability* that the deviation of the observed from that expected is due to chance alone (no other forces acting). In this case, using $p > 0.05$, we would expect any deviation to be due to chance alone 5% of the time or less.

3. Refer to a chi-square distribution table (Table B). Using the appropriate degrees of 'freedom, locate the value closest to our calculated chi-square in the table. Determine the closest *p* (probability) value associated with our chi-square and degrees of freedom. In this case (χ^2=2.668), the p value is about 0.10, which means that there is a 10% probability that any deviation from expected results is due to chance only. Based on our standard $p > 0.05$, this is within the range of acceptable deviation. In terms of our hypothesis for this example, the observed chi-square is not significantly different from expected. The observed numbers are consistent with those expected under Mendel's law.

Step-by-Step Procedure for Testing our Hypothesis and Calculating Chi-Square

1. State the hypothesis being tested and the predicted results. Gather the data by conducting the proper experiment (or, if working genetics problems, use the data provided in the problem).

2. Determine the expected numbers for each observational class. Remember to use numbers, not percentages.
3. Calculate χ^2 using the formula. Complete all calculations to three significant digits. Round off our answer to two significant digits.
4. Use the chi-square distribution table to determine significance of the value.
   a. Determine degrees of freedom and locate the value in the appropriate column.
   b. Locate the value closest to our calculated χ^2 on that degrees of freedom *df* row.
   c. Move up the column to determine the p value.
5. State our conclusion in terms of your hypothesis.
   a. If the *p* value for the calculated χ^2 is $p > 0.05$, accept our hypothesis. 'The deviation is small enough that chance alone accounts for it. A *p* value of 0.6, for example, means that there is a 60% probability that any deviation from expected is due to chance only. This is within the range of acceptable deviation.
   b. If the p value for the calculated χ^2 is $p < 0.05$, reject our hypothesis, and conclude that some factor other than chance is operating for the deviation to be so great. For example, a p value of 0.01 means that there is only a 1% chance that this deviation is due to chance alone. Therefore, other factors must be involved.

The chi-square test will be used to test for the "goodness to fit" between observed and expected data from several laboratory investigations in this lab manual.

Table A

Calculating Chi-Square

	Green	**Yellow**
Observed (O)	639	241
Expected (E)	660	220

Deviation (O - *E)*	-21	21
Deviation^2 (d2)	441	441
d^2/e	0.668	2
$\chi^2 = \Sigma d^2/e = 2.668$	.	.

- Chi-square requires that we use numerical values, not percentages or ratios.
- Chi-square should not be calculated if the expected value in any category is less than 5.

Table B

Chi-Square Distribution

Degrees of Freedom *(df)*	**Probability** (*p*)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32

8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

10.5. Correlation and regression

10.5.1. Introduction

Generally the distributions involves only one variable but sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

Suppose we have a set of 30 students in a class and we want to measure the heights and weights of all the students. We observe that each individual (unit) of the set assumes two values – one relating to the height and the other to the weight. Such a distribution in which each individual or unit of the set is made up of two values is called a bivariate distribution. The following examples will illustrate clearly the meaning of bivariate distribution.

In a class of 60 students the series of marks obtained in two subjects by all of them.

The series of sales revenue and advertising expenditure of two companies in a particular year.

The series of ages of husbands and wives in a sample of selected married couples.

Thus in a bivariate distribution, we are given a set of pairs of observations, wherein each pair represents the values of two variables.

In a bivariate distribution, we are interested in finding a relationship (if it exists) between the two variables under study.

The concept of ‘correlation’ is a statistical tool which studies the relationship between two variables and Correlation Analysis involves various methods and

techniques used for studying and measuring the extent of the relationship between the two variables.

"Two variables are said to be in correlation if the change in one of the variables results in a change in the other variable".

10.5.2. Types of Correlation

There are two important types of correlation. They are (1) Positive and Negative correlation and (2) Linear and Non – Linear correlation.

a) Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

- Heights and weights;
- Household income and expenditure;
- Price and supply of commodities;
- Amount of rainfall and yield of crops.

For example, when we have a fever, the fever its self is not the root cause of the problem; it is a correlation or secondary symptom of the root problem. This is not to say that correlations can have direct cause and effect result of their own; for example over heating due to too high of a fever,

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.

Some examples of series of negative correlation are:

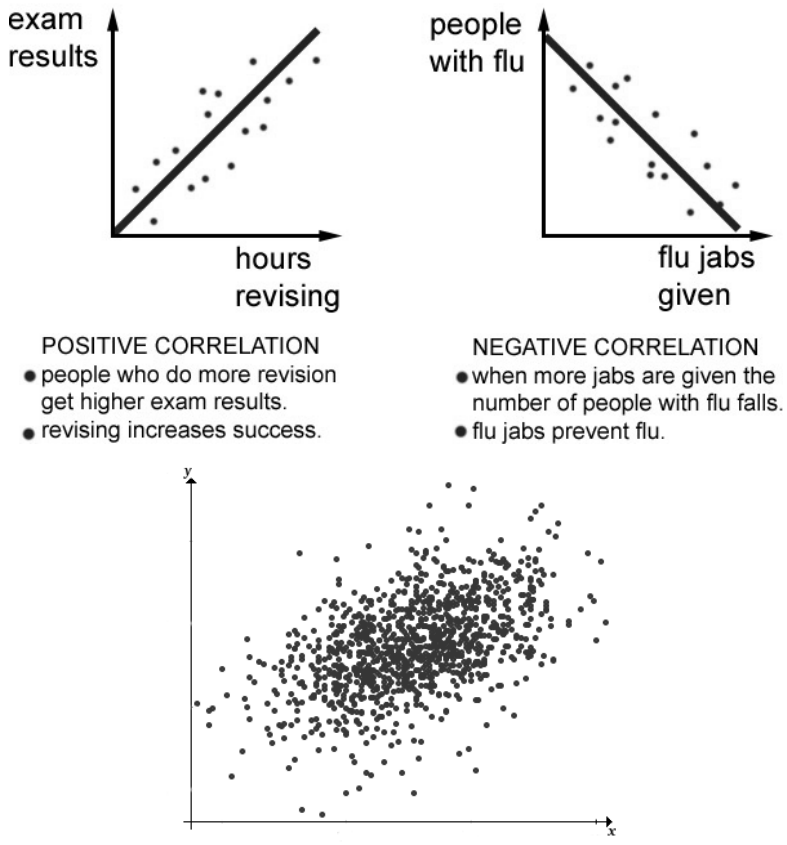
Volume and pressure of perfect gas;

Current and resistance [keeping the voltage constant] $(R=\frac{V}{I})$;

Price and demand of goods.

Graphs of Positive and Negative correlation:

Suppose we are given sets of data relating to heights and weights of students in a class. They can be plotted on the coordinate plane using x – axis to represent heights and y – axis to represent weights. The different graphs shown below illustrate the different types of correlations.



Zero Correlation

Note:

(i) If the points are very close to each other, a fairly good amount of correlation can be expected between the two variables. On the other hand if they are widely scattered a poor correlation can be expected between them.

(ii) If the points are scattered and they reveal no upward or downward trend as in the case of (d) then we say the variables are uncorrelated.

(iii) If there is an upward trend rising from the lower left hand corner and going upward to the upper right hand corner, the correlation obtained from the graph is said to be positive. Also, if there is a downward trend from the upper left hand corner the correlation obtained is said to be negative.

(iv) The graphs shown above are generally termed as scatter diagrams.

Example:1: The following are the heights and weights of 15 students of a class. Draw a graph to indicate whether the correlation is negative or positive.

Heights (cms)	Weights (kgs)
170	65
172	66
181	69
157	55
150	51
168	63
166	61
175	75
177	72
165	64
163	61
152	52
161	60
173	70
175	72

Since the points are dense (close to each other) we can expect a high degree of correlation between the series of heights and weights. Further, since the points reveal an upward trend, the correlation is positive. Arrange the data in increasing order of height and check that , as height increases, the weight also increases, except for some (stray) cases..

b) Linear and Non – Linear Correlation

The correlation between two variables is said to be **linear** if the change of one unit in one variable result in the corresponding change in the other

variable over the entire range of values.

For example consider the following data.

X	2	4	6	8	10
Y	7	13	19	25	31

Thus, for a unit change in the value of x, there is a constant change in the corresponding values of y and the above data can be expressed by the relation

$$y = 3x + 1$$

In general two variables x and y are said to be **linearly related**, if there exists a relationship of the form

$$y = a + bx$$

where 'a' and 'b' are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

The relationship between two variables is said to be **non – linear** if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form

$y = a + bx + cx^2$ or more general polynomial.

10.5.3. Explanation

a) The Coefficient of Correlation

The coefficient of correlation 'r' is given by the formula

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

The following example illustrates this idea.

Example: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of

correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

Solution:				
X	**Y**	x^2	y^2	xy
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372

796	1546	63,776	243036	1242069

Then

$$r = \frac{10(124206) - (796)(1546)}{\sqrt{(10)63776 - (796)^2}\sqrt{(10)(243036) - (1546)^2}}$$

$$= \frac{11444}{(1144)(40244)}$$

$$= 0.5966$$

b) Rank Correlation

Data which are arranged in numerical order, usually from largest to smallest and numbered 1,2,3 ---- are said to be in ranks or ranked data.. These ranks prove useful at certain times when two or more values of one variable are the same. The coefficient of correlation for such type of data is given by Spearman rank difference correlation coefficient and is denoted by R.

In order to calculate R, we arrange data in ranks computing the difference in rank 'd' for each pair. The following example will explain the usefulness of R. R is given by the formula

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

Example: The data given below are obtained from student records. Calculate the rank correlation coefficient 'R' for the data.

Subject	**Grade Point Average (x)**	**Graduate Record exam score**
1.	8.3	2300
2.	8.6	2250
3.	9.2	2380
4.	9.8	2400
5.	8.0	2000
6.	7.8	2100
7.	9.4	2360

8.	9.0	2350
9.	7.2	2000
10.	8.6	2260

Note that in the G. P. A. column we have two students having a grade point average of 8.6 also in G. R. E. score there is a tie for 2000.

Now we first arrange the data in descending order and then rank 1,2,3,----10 accordingly. In case of a tie, the rank of each tied value is the mean of all positions they occupy. In x, for instance, 8.6 occupy ranks 5 and 6

So each has a rank

$$\frac{5+6}{2} = 5.5$$

Similarly in 'y' 2000 occupies ranks 9 and 10, so each has rank

$$\frac{9+10}{2} = 9.5$$

Now we come back to our formula $R = 1 - 6\frac{(\sum d^2)}{n(n^2-1)}$

We compute 'd' , square it and substitute its value in the formula.

Subject	x	Y	Rank of x	Rank of y	d	d2
1.	8.3	2300	7	5	2	4
2.	8.6	2250	5.5	7	-1.5	2.25
3.	9.2	2380	3	2	1	1
4.	9.8	2400	1	1	0	0
5.	8.0	2000	8	9.5	-1.5	2.25
6.	7.8	2100	9	8	1	1
7.	9.4	2360	2	3	-1	1
8.	9.0	2350	4	4	0	0
9.	7.2	2000	10	9.5	0.5	0.25
10.	8.6	2260	5.5	6	-0.5	0.25

So here, n = 10, sum of d2 = 12. So

$$R = 1 - \frac{6(12)}{10\,(100 - 1)}$$

$$= 1 - 0.0727$$

Note: If we are provided with only ranks without giving the values of x and y we can still find Spearman rank difference correlation R by taking the difference of the ranks and proceeding in the above shown manner.

10.5.4. Regression

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy.

Regression analysis was explained by M. M. Blair as follows: "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

10.5.5. Explanation

a) Regression Equation

Suppose we have a sample of size 'n' and it has two sets of measures, denoted by x and y. We can predict the values of 'y' given the values of 'x' by using the equation, called the REGRESSION EQUATION.

$y^* = a + bx$

where the coefficients a and b are given by

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

The symbol y* refers to the predicted value of y from a given value of x from the regression equation.

Example: Scores made by students in a statistics class in the mid - term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

Student	Mid – Term	Final
1.	98	90
2.	66	74
3.	100	98
4.	96	88
5.	88	80
6.	45	62
7.	76	78
8.	60	74
9.	74	86
10.	82	80

Solution:

We want to predict the final exam scores from the midterm scores. So let us designate 'y' for the final exam scores and 'x' for the mid – term exam scores. We open the following table for the calculations.

Stud	X	y	X^2	xy
1	98	90	9604	8820
2	66	74	4356	4884
3	100	98	10,000	9800
4	96	88	9216	8448
5	88	80	7744	7040
6	45	62	2025	2790
7	76	78	5776	5928
8	60	74	3600	4440
9	74	86	5476	6364

10	82	80	6724	6560
Total	785	810	64,521	65,071

Numerator of b = 10 * 65,071 – 785 * 810 = 6,50,710 – 6,35,850 = 14,860

Denominator of b = 10 * 64, 521 – $(785)^2$ = 6,45,210 – 6,16,225 = 28,985

Therefore, b = 14,860 / 28,985 = 0.5127

Numerator of a = 810 – 785 * 0.5127 = 810 – 402.4695 = 407.5305

Denominator of a = 10

Therefore a = 40.7531

Thus , the regression equation is given by

$$y^* = 40.7531 + (0.5127)\, x$$

We can use this to find the projected or estimated final scores of the students.

For example, for the midterm score of 50 the projected final score is

$y^* = 40.7531 + (0.5127)\, 50 = 40.7531 + 25.635 = 66.3881$

which is a quite a good estimation.

To give another example, consider the midterm score of 70. Then the projected final score is

$y^* = 40.7531 + (0.5127)\, 70 = 40.7531 + 35.889 = 76.6421,$

which is again a very good estimation.

10. Summary

When we begin to study something, we often begin by looking at accompanying factors or correlations. Often, in modern terminology we use the words "risk factors" when identifying correlations that are statistically significant in the presence of say a medical condition, set of behavioural observations, or a social dynamic. In a slightly less than technical definition, we can think of correlations as being "symptoms" of the problem whether they are a direct result or not.

A *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t* distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most

commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic follows a Student's *t* distribution

Chi square is used most frequently to test the statistical significance of results reported in bivariate tables, and interpreting bivariate tables is integral to interpreting the results of a chi square test.

Regression analysis is one of the most frequently used tools in commercial research. In its simplest form, regression analysis allows researchers to analyze relationships between one independent and one dependent variable.

In marketing applications, the dependent variable is usually the outcome we care about (e.g., sales), while the independent variables are the instruments we have to achieve those outcomes with (e.g., pricing or advertising). Regression analysis can provide insights that few other techniques can.

10.6 Self-Learning Exercise

1. Write about importance of biostatistics in animal sciences.
2. Explain Student's t test and its various types.
3. Explain Chi-square test in detail with example.
4. Describe regression analysis and its types.
5. Write about regression analysis.
6. Write short notes on
    a) Positive and negative correlation
    b) Graphs of correlation
    c) Coefficient of correlation
    d) Regression equation

10.7 Reference Books

- Elements of biostatistics by S.Prasad, Rastogi publications
- Fundamentals of Biostatistics - by Veer Bala Rastogi, Ane Books India
- Fundamentals of biostatistics-Bernard Rosner

- Introduction to Biostatistics-Robert R Sokal and F James Rohlf, Dover Publication

Unit - 11

Probability Distributions : Binomial, Poisson & Normal

Structure of the Unit

11.0 Objectives

After going through this unit we will be able to understand:

- How to use Binomial , Poisson & Normal Distribution.
- How to find mean, variance & other properties of these distributions.
- How to Apply these distributions to a variety of problems.
- What is the importance of these distributions

1.1 Introduction

In probability and statistics, a **probability distribution** assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference.

When an experiment is conducted, such as tossing the coins, rolling a die, several outcomes or events occur with certain probabilities. These events or outcomes may be regarded as a variable which takes different values and each value is associated with a probability. The values of this variable depends on chance or probability. Such a variable is called a random variable. Random variables which take a finite number of values or to be more specific those which do not take all values in any particular range are called discrete random variables. For example, when 20 coins are tossed, the number of heads obtained is a discrete random variable and it takes values 0,1,...,20. These are finite number of values and in this range, the variable does not take values such as 2.8, 5.7 or any number other than a whole number. In contrast to discrete variable, a variable is continuous if it can assume all values of a continuous scale. Measurements of time, length and temperature are on a continuous scale and these may be regarded as examples of continuous variables.

1.2.Binomial Distribution: Binomial Distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was published posthumously in 1713, eight years after his death.

James Bernoulli

Let X is random variable represents total no. of successes in 'n' trails. Let the probability of success in each trail is p and the probability of failure is q=1-p and p remains constant from trail to trail.

Then Probability of getting Success is

- Outcome Probability

  S **p = 1/2**

  F **q = 1/2**

Now, we have to find out the probability of x successes in n trails.

Let us suppose that a particular order of outcomes of x successes in n repetitions be as follows

SSSSSFFFSSFS……FS(x number of successes and n-x failures)

Since, the trails are all independent the probability for the joint occurrence of the event is

pppppqqqppqp……..qp

= (pppppp…..x times)(qqqqqq…… (n-x) times)

$= p^x q^{n-x}$

Further in a series of n trails x successes and n-x failures can occur in ${}^nc_x$ ways. So, the required probability is

Probability of x successes in n trails is

$$P(X=x) = {}^{n}c_{x} \; p^{x} q^{n-x}, \quad x = 0,1,2,.....,n$$

Number of arrangements of x successes	Probability of x successes	Probability of $n-x$ failures

This is called probability distribution of Binomial random variable X or simply Binomial distribution. Symbolically this can be written as B(X; n, p). When n = 1 the Binomial Distribution is called Bernoulli Distribution.

Definition: A random variable X is said to be follow a binomial distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X=x) = {}^{n}c_{x} \; p^{x} q^{n-x}, \; x = 0,1,2,.....,n$$

And p + q =1

Where n and p are called parameters of the binomial distribution.

ASSUMPTIONS

- Number of trials "n" is finite .
- Outcome is dichotomous.
- Outcome is mutually exclusive & exhaustive.
- Outcome of the "n" trials are independent. This means one outcome does not effect any other outcome
- Probability of success "p" is constant for each trial .

11.2.1. Mean of the Binomial distribution

For a binomial distribution the probability function is given by

$$P(X=x) = {}^{n}c_{x} \; p^{x} q^{n-x}, \; x = 0,1,2,.....,n$$

Now, the mean of the Binomial distribution is

$$E(X) = \sum_{x=0}^{n} x\, P(X = x)$$

$$= \sum_{x=0}^{n} x\ {}^{n}c_{x} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x\ {}^{n}c_{x} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n(n-1)!}{x(x-1)!(n-x)!} p\, p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$= np \sum_{x=1}^{n} {}^{n-1}c_{x-1}\ p^{x-1} q^{n-x}$$

$$= np(q+p)^{n-1}$$

$$= np(1)^{n-1} \quad [\because q + p = 1]$$

$$= np$$

$\therefore$ The mean of the binomial distribution is np

11.2.2. Variance of the Binomial distribution:

The variance of the Binomial distribution is

$$V(X) = E(X^2) - [E(X)]^2$$

$$= E(X^2) - (np)^2 \ldots\ldots\ldots\ldots\ldots (1)\ [\because E(X) = np]$$

Now, $E(X^2) = \sum_{x=0}^{n} x^2\ {}^{n}c_{x} p^{x} q^{n-x}$

$$= \sum_{x=0}^{n} [x(x-1) + x]\ {}^{n}c_{x} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^{x} q^{n-x} + \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^{x} q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n(n-1)(n-2)!}{x(x-1)(x-2)!(n-x)!} p^2 \; p^{x-2} q^{n-x} + E(X)$$

$$= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 \sum_{x=2}^{n} {}^{n-2}c_{x-2} \; p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np$$

$$= n(n-1)p^2 (1)^{n-2} + np \quad [\because q + p = 1]$$

$$= n(n-1)p^2 + np \; \ldots\ldots\ldots\ldots (2)$$

Putting (2) in (1) we get

$$V(X) = n(n-1)p^2 + np - (np)^2$$

$$= np(np - p + 1 - np)$$

$$= np(1-p)$$

$$= npq$$

$\therefore$ The variance of the Binomial distribution is npq

Note: In B.D since mean = np and variance = npq and p + q = 1 therefore ***mean > variance***

11.2.3. Moments of Binomial distribution

The First four moment of Binomial distribution is as below

Non central moments (about zero):

$$\mu_1^{\,1} = E(X) = Mean = np$$

$$\mu_2^{\,1} = E(X^2) = n(n-1)p^2 + np$$

$$\mu_3^{\,1} = E(X^3) = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

$$\mu_4^{\,1} = E(X^4) = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

Central moments (about mean):

$$\mu_1 = 0$$

$$\mu_2 = Variance = V(X) = E(X - np)^2 = npq$$

$$\mu_3 = E(X - np)^3 = npq(q - p)$$

$$\mu_4 = E(X - np)^4 = npq[1 + 3(n-2)pq]$$

11.2.4. Measure of skew ness of Binomial distribution

$$\sqrt{\beta_1} = \sqrt{\frac{\mu_3^{\,2}}{\mu_2^{\,3}}}$$

$$= \sqrt{\frac{n^2 p^2 q^2 (q-p)^2}{n^3 p^3 q^3}}$$

$$= \sqrt{\frac{(q-p)^2}{npq}}$$

$$= \sqrt{\frac{(1-2p)^2}{npq}} \quad [\because q + p = 1]$$

$\Rightarrow$ The measure of skew ness of B.D is $\sqrt{\frac{(1-2p)^2}{npq}}$

Note: The Binomial distribution is called ***Symmetric Binomial Distribution*** if $p = \frac{1}{2}$

11.2.5. Measure of Kurtosis of Binomial Distribution:

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}}$$

$$= \frac{npq[1 + 3(n-2)pq]}{n^2 p^2 q^2}$$

$$= \frac{[1 + 3(n-2)pq]}{npq}$$

$$= 3 + \frac{1 - 6pq}{npq}$$

$\therefore$ Measure of Kurtosis of B.D is $3 + \frac{1 - 6pq}{npq}$

11.2.6. Recurrence Relation for the probabilities of Binomial Distribution:

For a B.D the probability mass function is given by

$$P(X = x) = {}^{n}c_{x}\ p^{x} q^{n-x}$$

and $$P(X = x-1) = {}^{n}c_{x-1}\ p^{x-1} q^{n-x+1}$$

Now,

$$\frac{P(X = x)}{P(X = x-1)} = \frac{{}^{n}c_{x} p^{x} q^{n-x}}{{}^{n}c_{x-1} p^{x-1} q^{n-x+1}}$$

$$= \frac{n - x + 1}{x} \frac{p}{q}$$

$$\Rightarrow \boxed{P(X = x) = \frac{n - x + 1}{x} \frac{p}{q} P(X = x-1)}$$

By using the above recursion formula, if we know $P(X = x-1)$, we can find $P(X = x)$

i.e. if we know $P(X = 0)$ we can find

successively $P(X = 1), P(X = 2), P(X = 3)$ …… so on

11.2.7. Mode of the Binomial Distribution:

Case-1: The Binomial Distribution has unique mode if (n+1) p is not an integer and the value of mode is m, the integral part of (n+1) p

Case-2: The Binomial Distribution has two modes i.e. bimodal if (n+1)p is an integer and the modes are m and m-1

11.2.8. Additive Property of binomial distribution :

Let $X \in B(n_1, p_1)$ and $Y \in B(n_2, p_2)$ be independent random variables. Then

$M_X(t) = (q_1 + p_1 e^t)^{n_1}$ $\qquad$ $M_X(t) = (q_1 + p_1 e^t)^{n_1}$

Since X And Y are independent

Than distribution of X+Y will be

$$M_{X+Y}(t) = M_X(t).M_Y(t) = (q_1 + p_1 e^t)^{n_1}(q_2 + p_2 e^t)^{n_2} \ldots\ldots\ldots\ldots\ldots\ldots(*)$$

Since (*) cannot be expressed in the form $(q + p\ e^t)^n$ from the uniqueness theorem of m.g.f it follows that X+Y is not a binomial variate , *Hence in general the sum of two independent binomial variate is not a binomial variate. In other words binomial distribution does not possess the additive or reproductive property* .

However if we take $p_1 = p_2 = p$ then from (*)

$$M_{X+Y}(t) = (q + pe^t)^{n_1+n_2}$$

Which is the m.g.f of a bionomial variate with parameters $(n_1 + n_2, p)$. Hence by uniqueness theorem of m.g.f's $X + Y \in B(n_1 + n_2, p)$. Thus the binomial distribution possesses the additive reproductive property if $p_1 = p_2$

11.2.9. Moment Generating Function of Binomial Distribution

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tX} p(x) q^{n-x}$$

$$= \sum_{x=0}^{n} (pe^t)^x q^{n-x}$$

$$= (q + pe^t)^n$$

11.2.10. Characteristic Function of Binomial Distribution

$$\phi_X(t) = E(e^{itX}) = \sum_{x=0}^{n} e^{itX} p(x) = \sum_{x=0}^{n} e^{itX}\ {}^n c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} {}^n c_x (pe^{it})^x q^{n-x}$$

$$= (q + pe^{it})^n$$

11.2.11Theorem: The sum of the probabilities of the binomial distribution is unity.

Proof:

For a binomial distribution the probability function is given by

$$P(X=x) = {}^{n}c_{x}\ p^{x}q^{n-x},\ x = 0,1,2,.....,n$$

Now, $\sum_{x=0}^{n} p(X=x) = \sum_{x=0}^{n} {}^{n}c_{x}p^{x}q^{n-x}$

$$= {}^{n}c_{0}\ p^{0}q^{n-0} + {}^{n}c_{1}\ p^{1}q^{n-1} + {}^{n}c_{2}\ p^{2}q^{n-2} + \ldots\ldots\ldots + {}^{n}c_{n}\ p^{n}q^{n-n}$$

$$= (q+p)^{n} = 1\ [\because q+p=1]$$

11.2.12. Properties of Binomial distribution:

1) Binomial distribution is a discrete probability distribution with two parameters n and p and finite range from 0 to n.
2) The mean and the variance of the B.D are np and npq respectively and mean > variance
3) The measure of skew ness of B.D is $\sqrt{\beta_1} = \sqrt{\frac{(1-2p)^2}{npq}}$

   If $p = \frac{1}{2}$, the distribution is symmetric

   $p < \frac{1}{2}$, the distribution is positively skewed

   $p > \frac{1}{2}$, the distribution is negatively skewed

4) The measure of Kurtosis of B.D is $\beta_2 = 3 + \frac{1-6pq}{npq}$

   If $pq = \frac{1}{6}$, the distribution is mesokurtic

$pq < \frac{1}{6}$, the distribution is leptokurtic

$pq > \frac{1}{6}$, the distribution is plattykurtic

and for a symmetric binomial distribution i.e. for $p = \frac{1}{2}$, the kurtosis of B.D is $3 - \frac{2}{n}$ 5)

For $p = \frac{1}{2}$, the binomial distribution has maximum probability at $x = \frac{n}{2}$, if n is even

and

$x = \frac{n-1}{2}$ and $x = \frac{n+1}{2}$, if n is odd

5) Under certain conditions the B.D approaches to Poisson and Normal distributions.

11.2.13. Prove that binomial distribution is a Probability Distribution.

Proof: In case you are wondering, this actually is a probability distribution. To show that this is the case, we can use a theorem from high school algebra: $(a+b)^n = \sum_{x=1}^{n} \binom{n}{x} a^x b^{n-x}$. This theorem is called the Binomial Formula, which is where the binomial distribution gets its name. Notice that the items in the sum look just like the probabilities in the binomial distribution.

Therefore, we have:

1. $\sum_x P(x) = \sum_{x=1}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1^n = 1$
2. Clearly, $0 \leq P(x)$ because all the terms are positive. Combining this fact with knowledge that the sum of the P(x) terms is 1 tells us that $P(x) \leq 1$ as well.

Therefore, we see that the binomial distribution is a probability distribution.

11.2.14. Uses of Binomial Distribution:

1) It has major application in the field of industrial quality control when items are classified as defective and non defective

2) This distribution is used when we like to know the opinion of the public when the voters may be in favor of or against a candidate.

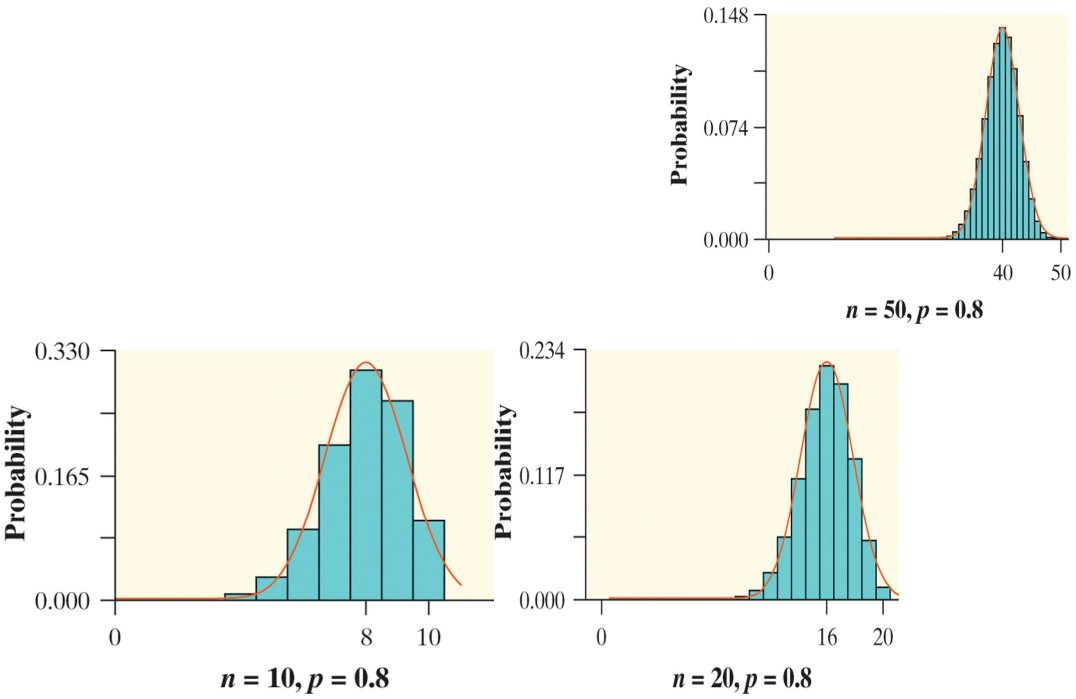
3) This distribution is also used in market researches where a consumer may prefer the product of brand A or brand B

4) This distribution is used in medical research where a particular drug might cure a person or not

5) This distribution also used in economic survey where respondents are in for or against a certain economic policy of the govt.

11.2.15. Normal Approximation for Binomial Distribution:

As n gets larger, something interesting happens to the shape of a binomial distribution.



Suppose that X has the binomial distribution with n trials and success probability p. When n is large, the distribution of X is approximately Normal with mean and standard deviation

$\mu_X = np$ $\sigma_X = \sqrt{np(1-p)}$

As a rule of thumb, we will use the Normal approximation when n is so large that $np \geq 10$ and $n(1-p) \geq 10$.

Example: Suppose a biased coin comes up heads with probability 0.3 when tossed. What is the probability of achieving 0, 1,..., 6 heads after six tosses?

$$\Pr(0 \text{ heads}) = f(0) = \Pr(X = 0) = \binom{6}{0} 0.3^0 (1-0.3)^{6-0} \approx 0.1176$$

$$\Pr(1 \text{ heads}) = f(1) = \Pr(X = 1) = \binom{6}{1} 0.3^1 (1-0.3)^{6-1} \approx 0.3025$$

$$\Pr(2 \text{ heads}) = f(2) = \Pr(X = 2) = \binom{6}{2} 0.3^2 (1-0.3)^{6-2} \approx 0.3241$$

$$\Pr(3 \text{ heads}) = f(3) = \Pr(X = 3) = \binom{6}{3} 0.3^3 (1-0.3)^{6-3} \approx 0.1852$$

$$\Pr(4 \text{ heads}) = f(4) = \Pr(X = 4) = \binom{6}{4} 0.3^4 (1-0.3)^{6-4} \approx 0.0595$$

$$\Pr(5 \text{ heads}) = f(5) = \Pr(X = 5) = \binom{6}{5} 0.3^5 (1-0.3)^{6-5} \approx 0.0102$$

$$\Pr(6 \text{ heads}) = f(6) = \Pr(X = 6) = \binom{6}{6} 0.3^6 (1-0.3)^{6-6} \approx 0.0007$$

Example:

Each child of a particular pair of parents has probability 0.25 of having blood type O. Suppose the parents have five children.

(a) Find the probability that exactly three of the children have type O blood.

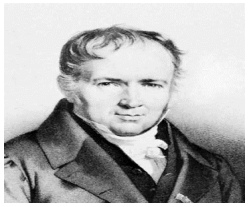
Let X = the number of children with type O blood. We know X has a binomial distribution with $n = 5$ and $p = 0.25$.

$$P(X = 3) = \binom{5}{3} (0.25)^3 (0.75)^2 = 10(0.25)^3 (0.75)^2 = 0.08789$$

b) Should the parents be surprised if more than three of their children have type O blood?

$$
\begin{aligned}
P(X>3) &= P(X=4)+P(X=5) \\
&= \binom{5}{4}(0.25)^4(0.75)^1 + \binom{5}{5}(0.25)^5(0.75)^0 \\
&= 5(0.25)^4(0.75)^1 + 1(0.25)^5(0.75)^0 \\
&= 0.01465 + 0.00098 = 0.01563
\end{aligned}
$$

Since there is only a 1.5% chance that more than three children out of five would have Type O blood, the parents should be surprised!



S. D. Poisson

11.3 Poisson distribution

Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781-1840) who published it in 1837. It is often used as a model for the number of events (such as the number of telephone calls at a business, Number of deaths from a disease such as cancer, Number of suicides reported in a particular city, number of customers in waiting lines, number of defects in a given surface area, airplane arrivals, or the number of accidents at an intersection) in a specific time period. The major difference between Poisson and Binomial distributions is that the Poisson does not have a fixed number of trials.

Instead, it uses the fixed interval of time or space in which the number of successes is recorded.

Poisson distribution is a discrete probability distribution, which is the limiting case of the binomial distribution under certain conditions.

1. When n is very indefinitely very large
2. Probability of success is very small.
3. np = λ is finite, $\lambda \in R^+$

Definition: A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,3.....\infty$$

Where e = 2.7183 and $\lambda > 0$

Here λ is called the ***parameter*** of the Poisson distribution.

ASSUMPTIONS

- At random (the occurrence of an event doesn't change the probability of it happening again)
- At a constant rate.

11.3.1. Mean of Poisson distribution:

For a Poisson distribution the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,3.....\infty$$

$$\text{Now, Mean} = E(X) = \sum_{x=0}^{\infty} x\,P(x;\lambda) = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda\lambda^{x-1}}{x(x-1)!}$$

$$= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}\lambda\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}\ldots\ldots\ldots+\infty\right]$$

$$= e^{-\lambda}\lambda e^{\lambda} \quad =\lambda$$

$\therefore$ Mean of the Poisson distribution is λ

11.3.2. Variance of Poisson distribution

Variance = V(X) = $E(X^2)-[E(X)]^2$

$= E(X^2)-[\lambda]^2$ …… (1) [$\because$ Mean of the Poisson distribution is λ]

From (1)

$$E(X^2) = \sum_{x=0}^{\infty}x^2\,P(x;\lambda) \;=\; \sum_{x=0}^{\infty}x^2\,\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}[x(x-1)+x]\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty}x(x-1)\frac{e^{-\lambda}\lambda^x}{x!}+\sum_{x=0}^{\infty}x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$=e^{-\lambda}\sum_{x=1}^{\infty}x(x-1)\frac{\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!}+\lambda$$ [$\because$ Mean of the Poisson distribution is λ]

$$=e^{-\lambda}\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}+\lambda$$

$$= e^{-\lambda}\lambda^2\left[\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}\ldots\ldots\ldots+\infty\right]+\lambda$$

$$= e^{-\lambda}\lambda^2 e^{\lambda}+\lambda$$

$\Rightarrow \quad E(X^2) = \lambda^2+\lambda$ …… (2)

Putting (2) in (1) we get

$$\text{V(X)} \;= E(X^2)-[\lambda]^2$$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda$$

$\therefore$ Variance of the Poisson distribution is λ

Note: In Poisson distribution the mean and variance are equal i.e. λ

11.3.3. Moments of the Poisson distribution:

Non central moments (about zero):

$$\mu_1^{\,1} = Mean = E(X) = \lambda$$

$$\mu_2^{\,1} = E(X^2) = \lambda^2 + \lambda$$

$$\mu_3^{\,1} = E(X^3) = \lambda^3 + 3\lambda^2 + \lambda$$

$$\mu_4^{\,1} = E(X^4) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

Central moments (about mean):

$$\mu_1 = 0$$

$$\mu_2 = Variance = \lambda$$

$$\mu_3 = \lambda$$

$$\mu_4 = 3\lambda^2 + \lambda$$

11.3.4. Skew ness of Poisson Distribution:

Measure of Skew ness of Poisson distribution is given by

$$\sqrt{\beta_1} = \sqrt{\frac{\mu_3^{\,2}}{\mu_2^{\,3}}} = \sqrt{\frac{\lambda^2}{\lambda^3}} = \sqrt{\frac{1}{\lambda}}$$

11.3.5. Kurtosis of Poisson Distribution:

Measure of Kurtosis of Poisson distribution is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{3\lambda^2 + \lambda}{\lambda^2} = 3 + \frac{1}{\lambda}$$

Note:

1) Since Skew ness $=\sqrt{\frac{1}{\lambda}}>0$, Poisson distribution is always positively skewed.

2) Since Kurtosis $=3+\frac{1}{\lambda}>3$, Poisson distribution is always Leptokurtic.

11.3.6. Recursion formula:

If X is a Poisson variate with p.m.f $P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}, x=0,1,2,3.....\infty$

And $P(X=x-1)=\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}, x=0,1,2,3.....\infty$

Now,

$$\frac{P(X=x)}{P(X=x-1)}=\frac{\frac{e^{-\lambda}\lambda^{x}}{x!}}{\frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}}=\frac{\lambda}{x}$$

$$\Rightarrow P(X=x)=\frac{\lambda}{x}P(X=x-1)$$

11.3.7. Mode of the Poisson distribution:

Case: 1) When λ is not an integer, the distribution is unimodal and the value of mode is the integral part of the λ.

Case: 2) When λ is an integer, the distribution is bimodal and the value of modes are λ and $\lambda-1$.

11.3.8. The sum of the probabilities of the Poisson distribution is unity i.e.

$\sum_{x=0}^{\infty}P(x;\lambda)=1$

Proof:

For a Poisson distribution the probability mass function is given by

$$p(X = x) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3.....\infty$$

Now, $\sum_{x=0}^{\infty} P(x;\lambda) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!}$

$$= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x}}{x!}$$

$$= e^{-\lambda}\left[\frac{\lambda^{0}}{0!}+\frac{\lambda^{1}}{1!}+\frac{\lambda^{2}}{2!}+\frac{\lambda^{3}}{3!}.....+\infty\right]$$

$$= e^{-\lambda}e^{\lambda}$$

$$=1$$

11.3.9. Moment Generating Function of Poisson Distribution.

$$M_{X}(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tX}\ \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \sum_{x=0}^{\infty}\frac{e^{-\lambda}(\lambda e^{t})^{x}}{x!}$$

$$= e^{-\lambda}\{1+\lambda e^{t}+\frac{(\lambda e^{t})^{2}}{2!}+.....\}$$

$$= e^{-\lambda(e^{t}-1)}$$

11.3.10. Characteristic Function of Poisson Distribution.

$$\varphi_{X}(t) = E(e^{itX}).\mathrm{p}(\mathrm{x},\lambda) = \sum_{x=0}^{\infty} e^{itX}\ \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{it})^{x}}{x!}$$

$$= e^{-\lambda}e^{\lambda e^{it}}$$

$$= e^{\lambda(e^{it}-1)}$$

11.3.11. Prove that the Poisson distribution is the limiting case of binomial distribution stating the required conditions.

Sol: The Poisson distribution can be limiting case of a binomial distribution under certain conditions.

1. Number of trails i.e. n is indefinitely large i.e. $n \to \infty$
2. p , the probability of success in each trail is indefinitely small i.e $p \to 0$
3. $np = \lambda$ is finite.

If X is a binomial variate then the probability mass function is given by

$$P(X=x) = {}^{n}c_{x}\, p^{x} q^{n-x},\ x = 0,1,2,.....,n$$

Under the above conditions

$$\lim_{n\to\infty} B(x;n,p) = \lim_{n\to\infty} {}^{n}c_{x}\, p^{x} q^{n-x}$$

$$= \lim_{n\to\infty} \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x} \quad [\because np = \lambda]$$

$$= \lim_{n\to\infty} \frac{n(n-1)(n-2).....(n-x+1)(n-x)!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^{x}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^{x}}{x!}\lim_{n\to\infty} \frac{n^{x}\left[1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left(1-\frac{x-1}{n}\right)\right]}{n^{x}} \frac{\left(1-\frac{\lambda}{n}\right)^{n}}{\left(1-\frac{\lambda}{n}\right)^{x}}$$

$$= \frac{\lambda^{x}}{x!}\lim_{n\to\infty} 1 \,.\, \frac{\left(1-\frac{\lambda}{n}\right)^{n}}{\left(1-\frac{\lambda}{n}\right)^{x}}$$

$$= \frac{\lambda^{x}}{x!} \frac{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}}{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{x}} = \frac{\lambda^{x}}{x!}\frac{e^{-\lambda}}{1^{x}} = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

11.3.12. Properties of Poisson distribution:

The following are the some of the properties of the Poisson distribution.

1. Poisson distribution is a discrete probability distribution with single parameter λ .
2. Both mean and variance of the Poisson distribution are equal to λ .
3. The distribution is positively skewed and leptokurtic.
4. It is asymptotic form of binomial distribution when p is small, n is large and np is finite.
5. the normal distribution is a limiting form of a Poisson distribution as $\lambda \to 0$
6. The distributio0n of rare events generally approximates to a Poisson distribution.
7. If X_1 and X_2 are two independent Poisson variates with mean λ_1 and λ_2 respectively, then $X = X_1 + X_2$ is also a Poisson variate with mean $\lambda_1 + \lambda_2$.

11.3.13. Applications of Poisson distribution:

This distribution is used to describe the behavior of the rare events like

- The number of phone calls arriving at a call center within a minute.
- The number of goals in sports involving two competing teams.
- The number of deaths per year in a given age group.
- No of printing mistakes per page in a large volume of a book.
- The number of accidents occurred annually at a busy crossing of city
- Under an assumption of homogeneity, the number of times a web server is accessed per minute.

The proportion of cells that will be infected at a given multiplicity of infection.

- In determining the number of deaths in a given period by a rare disease.
- The number of mutations in a given stretch of DNA after a certain amount of radiation.
- It has wide applications in industrial quality control.

11.3..14. Show that in a Poisson distribution with unit mean, mean deviation is 2/e times the standard deviation.

Sol: The Poisson distribution with unit mean i.e. $\lambda = 1$ is given by

$$P(x; 1) = \frac{e^{-1}1^x}{x!} = \frac{e^{-1}}{x!}$$

Now, we have to show that

$$\text{Mean deviation about mean} = \frac{2}{e}\text{Standard deviation}$$

$$\text{i.e. } E\left|X - E(X)\right| = \frac{2}{e}E(X^2) - [E(X)]^2$$

$$\text{i.e. } E\left|X - 1\right| = \frac{2}{e}\sqrt{1} \quad [\because \text{given mean} = 1]$$

$$\text{i.e. } E\left|X - 1\right| = \frac{2}{e}$$

$$\text{Now, } \quad E\left|X - 1\right| = \sum_{x=0}^{\infty}\left|x - 1\right|\frac{e^{-1}}{x!}$$

$$= \frac{1}{e}\left[1 + \sum_{x=1}^{\infty}\frac{x-1}{x!}\right]$$

$$= \frac{1}{e}\left[1 + \sum_{x=1}^{\infty}\frac{1}{(x-1)!} - \sum_{x=1}^{\infty}\frac{1}{x!}\right]$$

$$= \frac{1}{e}\left[1 + e - (e-1)\right]$$

$$= \frac{2}{e}$$ Hence in Poisson distribution with unit mean, mean deviation are 2/e times the standard deviation.

Example : Suppose the average number of calls by 104 in one minute is 2. What is the probability of 10 calls in 5 minutes?

Solution: Since the average number of calls by 104 in one minute is 2, thus the average number of calls in 5 minutes is 10. Let ***X*** represent the number of calls in 5 minutes. Then,

$$P(X = i) = f_x(i) = \frac{e^{-10}10^i}{i!},\ i = 0,1,2,\ldots$$

and

$$E(X) = 10 \quad .$$

Then,

$$P(10\,\text{calls in 5 minutes}) = P(X = 10) = f_x(10) = \frac{e^{-10}10^{10}}{10!} = 0.1251.$$

Example : Probability of an accident in a year is 0.00024. So in a town of 10,000, the rate.

Solution:

$$\lambda = np$$
$$= 10{,}000 \text{ x } 0.00024 = 2.4$$

$$P(X = 0) = \frac{e^{-2.4}(2.4)^o}{0!} = 0.0907$$

$$P(X = 1) = \frac{e^{-2.4}(2.4)^1}{1!} = 0.2177$$

11.4. Normal Distribution

The normal distribution was first discovered by the English mathematician De–-Moivre(1667-1754) in1733asalimitingcaseof thebinomialdistribution.Thenormal-distributionisalsoknownas Gaussian distribution in honor of Karl friedrich Gauss.

The normal distribution is the most important distribution in Statistics. It is a probability distribution of a continuous random variable and is often used to model the distribution of discrete random variable as well as the distribution of

other continuous random variables. The basic form of normal distribution is that of a bell, it has single mode and is symmetric about its central values

Definition: A continuous random variable X is said to have a normal distribution with parameters μ and σ^2 if its density function is given by the probability law

$$f(x/\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

Where $-\infty < X < \infty$

$-\infty < \mu < \infty,\ \sigma > 0$

$e = 2.7183,\ \pi = 3.1416$

$(\sqrt{2\pi} = 2.5066)$

Here μ and σ^2 are the mean and variance of the normal distribution respectively.

Note: A random variable X with mean μ and variance σ^2 and following the normal law can be expressed by $X \sim N(\mu,\sigma^2)$.

The following quotation due to Lipman rightly reveals the popularity and importance of the normal distribution: "Everybody believes in the law of errors (the normal curve), the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is experiment fact."

W.J.Youden of the National Bureau of Standards describes the importance of the normal distribution artistically in the following words:

THE NORMAL

LAW OF ERROR

STANDS OUT IN THE

EXPERIENCE OF MANKIND

AS ONE OF THE BROADEST

GENERALIZATIONS OF NATURAL

PHILOSOPHY. IT SERVES AS THE GUIDING INSTRUMENT IN RESEARCHES IN THE PHYSICAL AND SOCIAL SCIENCES AND IN MEDICINE, AGRICULTURE AND ENGINEERING. IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

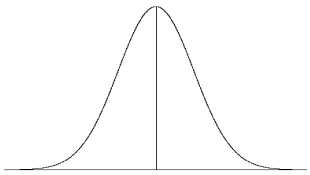
11.4.1 Importance of Normal Distribution

- Normal distribution is the most important and is widely used. The distribution of statistical measures such as mean or standard deviation tends to be normal when the sample size is large. Therefore, inferences are made about the nature of population from sample studies.
- The normal distribution may be used to approximate many kinds of natural phenomenon such as length of leaves, length of bones in mammals, height of adult males, intelligence quotient or tree diameters For example, in a large group of adult males belonging to the same race and living under same conditions, the distribution of heights closely resembles the normal distribution.
- For certain variables the nature of the distribution is not known. For the study of such variables, it is easy to wale the variables in such a way as to produce a normal distribution. It is indispensable in mental test study It is reasonable to assume that a selected group of children of a given age would show a normal distribution of intelligence test scores.

11.4.2.Properties of normal distribution:

The following points are important properties of normal distribution.

1. The normal curve is symmetrical and bell shaped. The range of the distribution is $-\infty \; to \; \infty$



Normal curve

2. The value of mean, median, mode will coincide as the distribution is symmetrical.

   i.e. mean = median = mode

3. The parameters μ and σ^2 represent the mean and variance of the distribution. It has only one mode i.e. the distribution is unimodal and it occurs at $x = \mu$.
4. The skew ness of the distribution is $\beta_1 = 0$ and kurtosis is $\beta_2 = 3$.
5. The odd ordered moments about mean vanishes i.e. $\mu_{2r+1} = 0$
6. The even ordered moments about mean $\mu_{2r} = (2r-1)\sigma^2 \mu_{2r-2}$.
7. The mean deviation from mean is $\sigma\sqrt{\frac{2}{\pi}} \cong \frac{4}{5}\sigma$.
8. The total area bounded by the curve and horizontal axis is equal to 1.
9. The maximum ordinate occurs at $x = \mu$ and its value is $\frac{1}{\sqrt{2\pi}\sigma}$.
10. The quartile deviation is $\frac{Q_3 - Q_1}{2} = 0.6745\sigma$
11. The first quartile $Q_1 = \mu - 0.6745\sigma$ and third quartile $Q_3 = \mu + 0.6745\sigma$
12. The co-efficient of quartile deviation $\frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.6745\frac{\sigma}{\mu}$.
13. A linear function $a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n$ of n independent normal variables $X_1, X_2, X_3, ..., X_n$ with means $\mu_1, \mu_2, \mu_3, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, ..., \sigma_n^2$ is also a normal variable with mean $a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + ...+ a_n\mu_n$ and variance $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_3^2\sigma_3^2 + ...+ a_n^2\sigma_n^2$.

Proof:

Let Z= $a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n$

Now, $E(Z) = E\left[a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n\right]$

$= a_1E[X_1] + a_2E[X_2] + a_3E[X_3] + ...+ a_nE[X_n]$

$= a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + ...+ a_n\mu_n$

And $V(Z) = V\left[a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n\right]$

$= a_1^2V(X_1) + a_2^2V(X_2) + a_3^2V(X_3) + ...+ a_n^2V(X_n)$

$= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_3^2\sigma_3^2 + ...+ a_n^2\sigma_n^2$

14. If X is a normal variate with mean μ and standard deviation σ, then the distribution of $Z = \frac{X-\mu}{\sigma}$ is also normal with mean 0 and variance 1. Here Z is called standard normal variable.

Symbolically if $X \sim N(\mu,\sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Proof:

$$E(Z) = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\left[E(X) - \mu\right] = \frac{1}{\sigma}\left[\mu - \mu\right] = 0$$

$$V(Z) = V\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2}\left[V(X)\right] = \frac{1}{\sigma^2}\left[\sigma^2\right] = 1.$$

Note: The probability density function of standard normal variable Z is

$$f(z/0,1) = N(0,1) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{1}{2}z^2}, -\infty < Z < \infty$$

15 Normal Distribution as a Limiting form of Binomial Distribution:

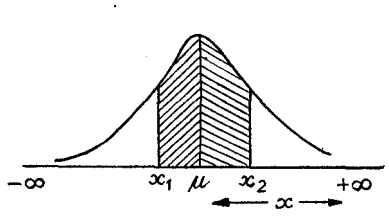
Normal distribution is another limiting form of the binomial distribution under the following conditions :

(i) n, the number of trials is indefinitely large, i.e., $n \to \infty$ and

(ii) neither p nor q is very small.

1.4.3.Area under normal curve

As the normal variable is a continuous random variable, the probability that the random variable X assumes a value $x = x_1$ and $x = x_2$ is represented by the area under the probability curve bounded by the values x_1 *and* x_2 can be defined as



$$\Pr ob(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Since the normal curve depends on two parameters μ and σ^2, the area represented by $\Pr ob(x_1 < x < x_2)$ is also dependent on μ and σ^2. Though theoretically this probability can be calculated by using the method of integral calculus, normal integral tables are available for the use of practicing statisticians. It is very voluminous work to compile tables for all possible values of μ and σ^2. In fact such tables would be infinitely many because $-\infty < \mu < \infty, \sigma > 0$.

To facilitate the preparation of tables, the normal variable is standardized or is transformed to a new variable which is also normal, but having mean 0 and variance 1. Thus if X is normal variable with mean μ and variance σ^2, then $Z = \frac{X - \mu}{\sigma}$ is a standardized normal variable having mean 0 and variance 1

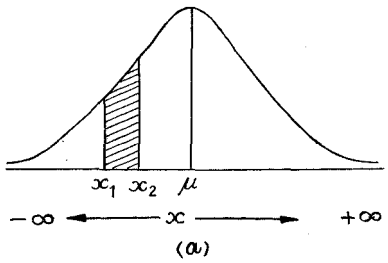
And thus $\Pr ob(x_1 < x < x_2) = \Pr ob\left(\frac{x_1 - \mu}{\sigma} < \frac{x - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}\right)$

$$= \Pr ob(z_1 < z < z_2)$$

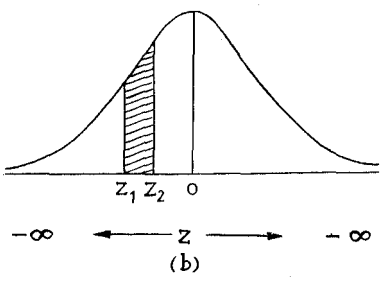
i.e.

$$\Pr ob(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(z)^2} dz = \Pr ob(z_1 < z < z_2)$$



Original normal curve



Standardised normal curve

11.4.4.Cumulative distribution functions of Z:

The cumulative distribution function of Z is defined as

$$\phi(t) = P(Z \le t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(z)^2}\, dz$$

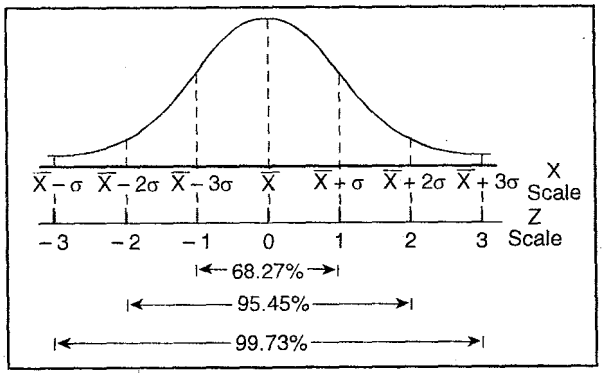
And

$$P(Z > t) = 1 - P(Z \le t)$$

Note: since the normal curve is symmetrical

$$P(Z > t) = P(Z < -t)$$

11.4.5. Areas of standardized normal curve:

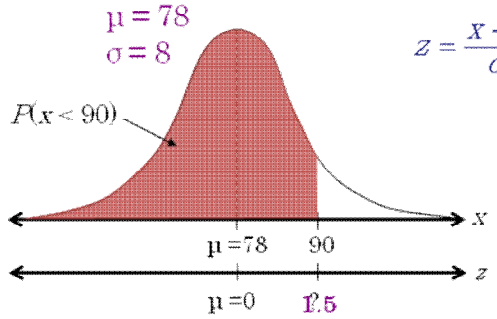


Distance from the mean ordinate	Area under curve

±0.745	50%
±1	68.27%
±1.96	95%
±2	95.45%
±2.58	99%
±3	99.73%

PROBABILITY AND NORMAL DISTRIBUTIONS

Example:
The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.



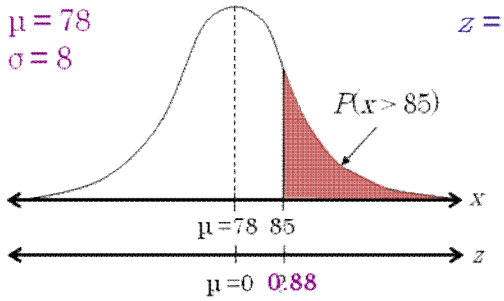
$$z = \frac{x - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

The probability that a student receives a test score less than 90 is 0.9332.

$$P(X < 90) = P(Z < 1.5) = 0.9332$$

PROBABILITY AND NORMAL DISTRIBUTIONS

Example:
The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than than 85.



$$z = \frac{x - \mu}{\sigma} = \frac{85 - 78}{8} = 0.875 \approx 0.88$$

The probability that a student receives a test score greater than 85 is 0.1894.

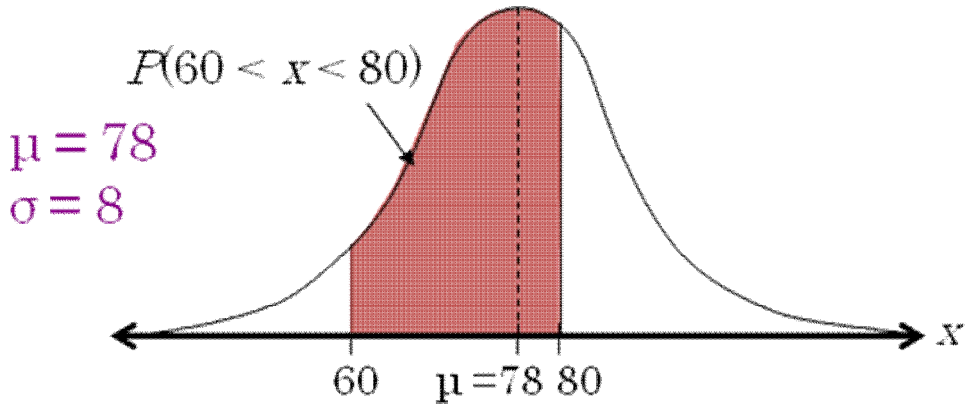
$$P(x > 85) = P(z > 0.88) = 1 - P(z < 0.88) = 1 - 0.8106 = 0.1894$$

Example: The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80**.**

Solution:

$$z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} = -2.25$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} = 0.25$$



The probability that a student receives a test score between 60 and 80 is 0.5865

$$P(60 < x < 80) = P(-2.25 < z < 0.25) = P(z < 0.25) - P(z < -2.25)$$

$$= 0.5987 - 0.0122 = 0.5865$$

11.5. Summary

In this unit, we learned fundamentals knowledge of probability distribution namely Binomial , Poisson & Normal Distribution. We find mean, variance & other properties of these distributions. We Apply these distributions to a variety of problems with their importance with specific example .

1.6 Glossary

- **Binary variable:** A variable whose only two possible values, usually zero and one.
- **Bimodal:** Having two modes
- **Bivariate:** Having two variables
- **Categorical Variable:** A variable whose value ranges over categories, such as Male & Female

- **Continuous variable:** A variable that can take on any number of possible values.
- **Gaussian distribution:** See normal distribution
- **Mean**: The sum of a list of numbers, divided by the number of numbers.
- **Median:** Middle value" of a list.
- **Mode:** mode is a most common (frequent) value.
- **Normal distribution:** A symmetric, bell-shaped distribution that is most useful for approximating the distribution of statistical estimators
- **Parameter:** An unknown quantity such as the population mean, population variance, difference in two means,
- **Random sample:** A sample selected by a random device that ensures that the sample (if large enough) is representative of the infinite group.
- **Variance:** A measure of the spread or variability of a distribution, equaling the average value of the squared difference between measurements and the population mean measurement.

11.7 Self-Learning Exercise

Section -A (Very Short Answer Type):

1. Binomial distribution applies to.....
2. Mean of Binomial distribution is .....
3. Variance of Binomial distribution is .....
4. Characteristic function of Binomial distribution is .....
5. MGF of Binomial distribution is .....
6. In Binomial distribution the probability of success p is..... for each trial.

7 In Binomial distribution trials are ..... of each other.

8. In Binomial distribution number of trials n is ……

9. In …… mean is greater than variance.
10. recurrence relation for the probabilities of binomial distribution is……
11. Poisson distribution applies to.....
12. Mean and Variance of ..... distribution is same.
13. In Poisson distribution probability mass function is .....
14. Characteristic function of Poisson distribution is .....
15. MGF of Poisson distribution is .....
16. Normal distribution is symmetrical about the line…….
17. Mean, Median and mode of Normal distribution…….
18. The skew ness of the Normal distribution is ……
19. The kurtosis of the Normal distribution is ……..
20. Range of Normal distribution is ……..

Section -B (Short Answer Type)

1. What is the Assumptions of binomial distribution?
2. What is the probability mass function of binomial distribution?
3. What is the recurrence relation for the probabilities of binomial distribution?
4. What is the Characteristic function of binomial distribution?
5. What is Moment Generating function of binomial distribution?
6. What is Kurtosis of Binomial Distribution. ?
7. What is Skewness of Binomial Distribution. ?
8. What is the probability generating function of binomial distribution?
9. What is the conditions at which Poisson distribution is a limiting case of Binomial distribution?

10. Prove that sum of the probabilities of the binomial distribution is unity

12. What is the probability mass function of Poisson distribution?

13. What is Moment Generating function of Poisson distribution??

14. What is the Characteristic function of Poisson distribution?

15. What is the Mode of Poisson distribution?

16. What is additive proper of Poisson distribution?

17. What is Kurtosis of Poisson Distribution?

18. What is the probability generating function of Poisson Distribution?

19. What is the conditions at which Normal distribution is a limiting case of Binomial distribution. ?

20. What is the mean deviation from mean in Normal distribution?

Section -C (Long Answer Type)

1. Define Binomial distribution with parameter p and n. Obtain the mean, variance, Characteristic function, of Binomial distribution. And give a situation in real life where the distribution is likely to be realized.

2. Obtain the Moment generating function of Binomial distribution .Derive from the result that sum of two binomial variates is a binomial variate if the variates are independent and have the same probability of success.

3. Define Poisson distribution with examples . Derive Poisson distribution as a limiting form of a binomial distribution.

4. State and prove the additive property of the Poisson distribution. Obtain the mean, variance, Moment generating function & Characteristic function of Poisson distribution .

5. What is the Applications & Properties of Poisson distribution .Show that in a Poisson distribution with unit mean, mean deviation is 2/e times the standard deviation

6. Define Normal distribution with its characteristics & Importence . Prove that thelinear function $a_1X_1 + a_2X_2 + a_3X_3 + ...+ a_nX_n$ of n independent

normal variables $X_1, X_2, X_3, \ldots\ldots\ldots, X_n$ with means $\mu_1, \mu_2, \mu_3, \ldots\ldots\ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots\ldots\ldots, \sigma_n^2$ is also a normal variable .

Answer Key of Section-A

1. Repeated two alternatives

2. np

3. npq

4. $(q + pe^{it})^n$

5. $(q + pe^{t})^n$

6. Constant

7. independent

8. finite

9.Binomial

10. $P(X = x) = \frac{n - x + 1}{x} \frac{p}{q} P(X = x - 1)$

11.Rare

12.Poisson

13. $p(X = x) = P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0,1,2,3\ldots\ldots\ldots\infty$

14. $e^{-\lambda(e^{it} - 1)}$

15. $e^{-\lambda(e^{t} - 1)}$

16. $x = \mu$

17.Coincide

18. $\beta_1 = 0$

19. $\beta_2 = 3$

20. $-\infty \; to \; \infty$

11.8 References

- Goon A.M., Gupta M.K. and Dasgupta B. (2005): Fundamentals of Statistics, Vol. I, 8th Edn. World Press, Kolkata.
- Goon, A.M., Gupta, M.K. and Dasgupta, B. (2003): An Outline of Statistical Theory, Vol. I, 4th Edn. World Press, Kolkata.
- Gupta, S.C. and Kapoor, V.K. (2007): Fundamentals of Mathematical Statistics, 11th Edn., (Reprint), Sultan Chand and Sons.
- Miller, Irwin and Miller, Marylees (2006): John E. Freund's Mathematical Statistics with Applications, (7th Edn.), Pearson Education, Asia.
- Mood, A.M. Graybill, F.A. and Boes, D.C. (2007): Introduction to the Theory of Statistics, 3rd Edn., (Reprint), Tata McGraw-Hill Pub. Co. Ltd.
- Rohatgi, V. K. and Saleh, A. K. Md. E. (2009): An Introduction to Probability and Statistics, 2nd Edn. (Reprint). John Wiley and Sons.
- Hogg R.V. and Craig R.G. - Introduction to mathematical statistics Ed 4 {1989} - Macmillan Pub. Co. New York
- B.L. Agarwal, Programmed Statistics , New Age International Publishers Ltd.
- Mukhopadhyay, P : Mathmatical Statistics, new central book agency.

Unit - 12

History and generation of computer, Fundamentals of computer

Structure of the Unit

12.0 Objectives

After completing this unit we will be able to understand:

- Introduction to Computer
- Classification of Computers
- Characteristics of a Computers
- History of computers
- Generations of computers
- Fundamentals of computers
- The working of computer system.
- Applications of computers

12.1 Introduction

When more than one thing is needed to work together before an action can take place, we say we have a system. The computer needs several things before it can be used to solve problems. Therefore, we say the computer is a system.

In relation to zoology this can be explained by our human body system. We may have systems around us and inside our body. How we eat or what happens to the food we eat is a system. This is called the Digestive System. We need mouth, tongue, teeth and stomach for us to make good use of the food we eat. In explanation of this; food is put in the mouth as INPUT; the food is digested and changed into useful things that our body needs; this is called PROCESSING, the

undigested food from the body in the form of waste passed out; this is called OUTPUT.

In the same way, relating this to computer, data is given to computer as INPUT, computer acts on the input by performing some operations on them; this called PROCESSING, the computer produces something after processing; this is called OUTPUT. These three stages of input, process and output are called COMPUTING and it gives the reason for referring to the computation as an I-P-O system. (Gbadeyan Et.al 2007)

12.1.1 What is Computer?

Computer is an electronic device that is designed to work with Information. *The term computer is derived from the Latin term* '**computare**', this means to *calculate* or *programmable machine*. **Computer cannot do anything without a Program**. It represents the decimal numbers through a string of binary digits. The Word 'Computer' usually refers to the Center Processor Unit plus Internal memory.

The Techencyclopedia (2003) defines computer as "a general purpose machine that processes data according to a set of instructions that are stored internally either temporarily or permanently." The computer and all equipment attached to it are called hardware. The instructions that tell it what to do are called "software" or "program". A program is a detailed set of humanly prepared instructions that directs the computer to function in specific ways. Furthermore, the Encyclopedia Britannica (2003) defines computers as "the contribution of major individuals, machines, and ideas to the development of computing." This implies that the computer is a system. A system is a group of computer components that work together as a unit to perform a common objective.

Computer is an advanced electronic device that takes raw data as input from the user and processes these data under the control of set of instructions (called program) and gives the result (output) and saves output for the future use. It can process both numerical and non-numerical (arithmetic and logical) calculations.

A computer is an electronic device which accepts and processes data by following a set of instructions (PROGRAM) to produce a result (INFORMATION). Since the ultimate aim of a computer is to produce information, the art of computing is often referred to as information processing. (Ayo, 1994)

Digital Computer Definition

The basic components of a modern digital computer are: Input Device, Output Device, Central Processor Unit (CPU), mass storage device and memory.

Four Functions about computer are:

1. Input – Accepts Data
2. Processing – Processes Data
3. Output – Produces Output
4. Storage – Stores Results

Input (Data):

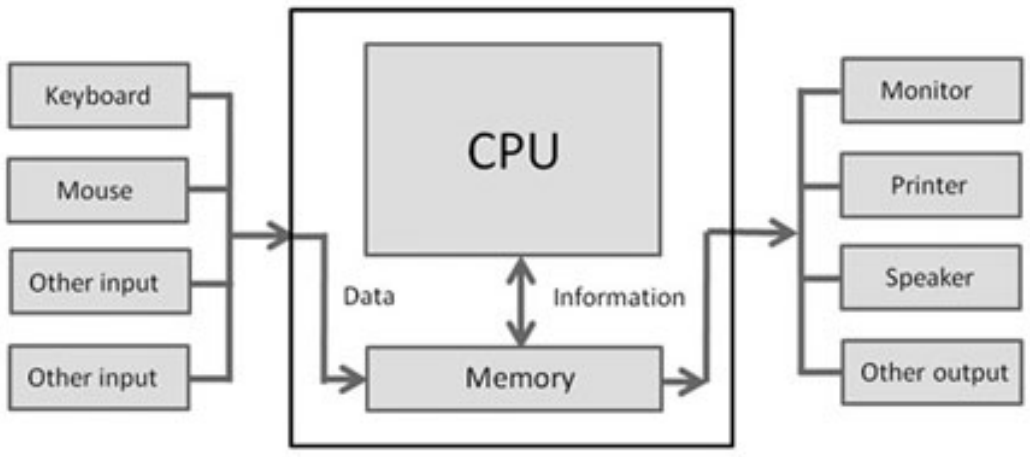
Input is the raw information entered into a computer from the input devices. It is the collection of letters, numbers, images etc.

Process:

Process is the operation of data as per given instruction. It is totally internal process of the computer system.

Output:

Output is the processed data given by computer after data processing. Output is also called as Result. We can save these results in the storage devices for the future use.



12.1.2 Classification of computers

The computers can be classified by the technology from which they were constructed, the uses to which they are put, their capacity or size, the era in which they were used, their basic operating principle and by the kinds of data they process.

On the basis of type of data processed computers are classified which is broadly referred to as types of computers. Statisticians generally classify data collected by counting as discrete data e.g. head count, number of Mangoes etc. Similarly, data collected through measurement are called continuous data. Examples are temperature, height, weight etc. This concept will help us to understand the types of computers. Computers are also classified by their different generations or by their size, or by the different purposes for which they are meant. Some of these classification techniques are discussed as follows:

12.1.2.1 Classification by type of data processed

i. Analog Computers

Analog computers are those that represent data in a continuous manner using physical variables such as pressure, temperature etc. An example is the analog watch. The outputs of analog computers are usually represented in the form of smooth curves or graphs from which information can be read. These computers are less accurate than digital output since their accuracy depends on the user or reader of such output. These classes of computers are used for scientific/engineering purposes and are not concerned with commercial data processing.

A good example of this class of computer is the computer used in hospitals for measuring blood pressure of patients, also a filling station gasoline pumps work purely on analog processes. The volume of fuel pumped out is converted into two measurements (i) price to the nearest kobo, and (ii) volume of fuel to the nearest litre. Other simple devices are the slide-rule, and the car speedometer.

ii. Digital Computers

These are computers that represent data in discrete or discontinuous manner using binary system. A digital watch is an example of a digital device. The output from digital computers are usually in the form of discrete values. This class of computers are commonly found in the business environments, and they include

Desk calculators, Adding machines, and most of the computers we have around (IBM, BBC, Radio Shack Personal Computers (PC), Laptops, Desktops etc.

iii. Hybrid Computers

As the name implies, this class of computers combines the features of both digital and analog computers. Their outputs could be in the form of discrete or continuous value or a combination of both. This class of computers is commonly found in highly scientific environments. Example is an electronic calculating scales used in food stores.

Of these three; digital computers are the most common, since they lend themselves to use in business, scientific and even home environments.

12.1.2.2 Classification by Capacity

Computers can be classified according to their capacity. The term 'capacity' refers to the volume of work or the data processing capability a computer can handle. Their performance is determined by the amount of data that can be stored in memory, speed of internal operation of the computer, number and type of peripheral devices, amount and type of software available for use with the computer.

The capacity of early generation computers was determined by their physical size - the larger the size, the greater the volume. Recent computer technology however is tending to create smaller machines, making it possible to package equivalent speed and capacity in a smaller format. Computer capacity is currently measured by the number of applications that it can run rather than by the volume of data it can process. This classification is therefore done as follows:

I **Micro Computers:** The Microcomputer has the lowest level capacity. The machine has memories that are generally made of semiconductors fabricated on silicon chips. Large-scale production of silicon chips began in 1971 and this has been of great use in the production of microcomputers. The microcomputer is a digital computer system that is controlled by a stored program that uses a microprocessor, a programmable read-only memory (ROM) and a random-access memory (RAM). The ROM defines the instructions to be executed by the computer while RAM is the functional equivalent of computer memory.

The Apple IIe, the Radio Shack TRS-80, and the Genie III are examples of microcomputers and are essentially fourth generation devices. Microcomputers have from 4k to 64k storage location and are capable of handling small, single-business application such as sales analysis, inventory, billing and payroll.

II **Mini Computers:** In the 1960s, the growing demand for a smaller stand-alone machine brought about the manufacture of the minicomputer, to handle tasks that large computers could not perform economically. Minicomputer systems provide faster operating speeds and larger storage capacities than microcomputer systems. Operating systems developed for minicomputer systems generally support both multiprogramming and virtual storage. This means that many programs can be run concurrently. This type of computer system is very flexible and can be expanded to meet the needs of users.

Minicomputers usually have from 8k to 256k memory storage location, and a relatively established application software. The PDP-8, the IBM systems 3 and the Honeywell 200 and 1200 computer are typical examples of minicomputers.

III **Medium-size Computers:** Medium-size computer systems provide faster operating speeds and larger storage capacities than mini computer systems. They can support a large number of high-speed input/output devices and several disk drives can be used to provide online access to large data files as required for direct access processing and their operating systems also support both multiprogramming and virtual storage. This allows the running of variety of programs concurrently. A medium-size computer can support a management information system and can therefore serve the needs of a large bank, insurance company or university. They usually have memory sizes ranging from 32k to 512k. The IBM System 370, Burroughs 3500 System and NCR Century 200 system are examples of medium-size computers.

IV **Large Computers:** Large computers are next to Super Computers and have bigger capacity than the Medium-size computers. They usually contain full control systems with minimal operator intervention. Large computer system ranges from single-processing configurations to nationwide computer-based

networks involving general large computers. Large computers have storage capacities from 512k to 8192k, and these computers have internal operating speeds measured in terms of nanosecond, as compared to small computers where speed is measured in terms of microseconds. Expandability to 8 or even 16 million characters is possible with some of these systems. Such characteristics permit many data processing jobs to be accomplished concurrently.

Large computers are usually used in government agencies, large corporations and computer services organizations. They are used in complex modeling, or simulation, business operations, product testing, design and engineering work and in the development of space technology. Large computers can serve as server systems where many smaller computers can be connected to it to form a communication network.

V **Super Computers:** The supercomputers are the biggest and fastest machines today and they are used when billion or even trillions of calculations are required. These machines are applied in nuclear weapon development, accurate weather forecasting and as host processors for local computer and time sharing networks. Super computers have capabilities far beyond even the traditional large-scale systems. Their speed ranges from 100 million-instruction-per-second to well over three billion. Because of their size, supercomputers sacrifice a certain amount of flexibility. They are therefore not ideal for providing a variety of user services. For this reason, supercomputers may need the assistance of a medium-size general purpose machines (usually called front-end processor) to handle minor programs or perform slower speed or smaller volume operation.

12.1.2.3 Classification by Size

Different **types of computer** systems are nowadays available for different purposes according to the user needs.

(i) Personal computers or Microcomputers

Microcomputers are built to be used by one person. In fact when we talk about *personal* computers or its *common* acronym PC, we always mean **microcomputers**. For surfing the web, playing games or music, editing and many other tasks… we ordinarily use personal computers either at school,

at home or at business.

Personal computers are in two (2) major categories: ***desktop and laptop***.

(ii) Workstations and Servers

If we need a high-end micro computer we should go for a workstation. This type of computer is recommended if we are carrying out game development, scientific calculations, *engineering* or 3D graphics. It is faster than the common personal computer and can be used as a **server** if we need to set up a network client.

The server is generally used for the purpose of allowing many users to work together over a network. Servers require powerful processors, large amount of hard disk drives and random access memory (RAM).

(iii) Mobile Computers

If we prefer the **laptop** we'll go for the *mobile* or *portable* system. Our **notebook**, a common name of laptop, has the advantage to have all the parts built together. The notebook has the same computing power as the desktop machine but it is lightweight enough as to be portable. The mobile computer is relatively more expensive because it costs more to design the small components.

For greater portability. A **handheld** micro computer is now a common option. To manage phone book, diary or to taking notes .etc a Personal Digital Assistant (PDA) is useful for the same purpose.

We can also use the Palmtop, a tinier laptop, for the same purposes and even more. The Palmtop is designed with a small keyboard and a flip-up screen and is useful for surfing the web while we are on the move.

(iv) Mini Computers

These are medium-sized computers that usually have several terminals for input and output, several disk drives and sometimes tape drives for data storage. Minicomputers have greater storage capacity and are faster in speed than microcomputers. Most often, minicomputers can process several programs at a time and can be used by several people simultaneously. Small and medium-sized business also uses minicomputer for their data processing need.

(v) Mainframe Computers

Mainframe computers have many terminals and several disk and tape drives. The components of mainframe computers have greater storage capacity and are faster than those of minicomputers. Mainframe computers can process numerous programs concurrently and can be used by medium and large sized business for their data processing needs. For example the network support for Automatic Teller Machines (ATMs).

(vi) Super Computers

These are the most powerful and most expensive computers. They are designed for very fast processing speeds. Although they have some fewer terminals and their processors can operate much faster, supercomputers are mainly used for complex mathematical calculations such as those needed in scientific research, or space exploration of their complexity and cert.

12.1.2.3 Classification by Purpose

i. Special Purpose computers

These are computers designed to carry out specific tasks. They have in-built programs which are stored in a part of the main memory called Read-Only Memory (ROM). The content of this part of the memory can be accessed and executed by the computer, but cannot be modified by the programmer or the user.

Thus, the operations that can be carried out by this type of computer are pre-determined at the time of manufacture. The computer cannot be used for any other purpose. Microcomputers (small computers) are so inexpensive however; that we can afford to use them to perform specific functions such as telling the time, monitoring human temperature or blood pressure.

ii. General Purpose Computer

These are computers that are not specifically designed or built for specific jobs. They solve various kinds of problems depending on the program or software loaded into them. Their main memory is typically Random Access Memory (RAM) - a temporary storage that looses its contents when the computer is switched off. It is easy to change the contents of RAM, substituting one program for another and this is what makes them general-

purpose computers. Most microcomputers or PCs are general purpose computers.

12.1.3 Characteristics of computer

Basic characteristics about **computer** are:

1. **Speed: - C**omputer can work very fast. It takes only few seconds for calculations that we take hours to complete. Computer can perform millions (1,000,000) of instructions and even more per second. Therefore, we determine the speed of computer in terms of microsecond (10-6 part of a second) or nanosecond (10 to the power -9 part of a second).
2. **Accuracy: -** The degree of accuracy of computer is very high and every calculation is performed with the same accuracy. The accuracy level is **7** determined on the basis of design of computer. The errors in computer are due to human and inaccurate data.
3. **Diligence: -** A computer is free from tiredness, lack of concentration, fatigue, etc. It can work for hours without creating any error. If millions of calculations are to be performed, a computer will perform every calculation with the same accuracy. Due to this capability it overpowers human being in routine type of work.
4. **Versatility: -** It means the capacity to perform completely different type of work. computer may be used to prepare payroll slips. Next moment we may use it for inventory management or to prepare electric bills.
5. **Power of Remembering: -** Computer has the power of storing any amount of information or data. Any information can be stored and recalled as long as we require it, for any numbers of years. It depends entirely upon us how much data we want to store in a computer and when to lose or retrieve these data.
6. **No IQ: -** Computer is a dumb machine and it cannot do any work without instruction from the user. It performs the instructions at tremendous speed and with accuracy. computer cannot take its own decision as we can.
7. **No Feeling: -** It does not have feelings or emotion, taste, knowledge and experience. Thus it does not get tired even after long hours of work. It does not distinguish between users.

8. **Storage: -** The Computer has an in-built memory where it can store a large amount of data. We can also store data in secondary storage devices such as floppies, CDs, Pen drives which can be kept outside our computer and can be carried to other computers.

12.3 The History of Computer

It is difficult to identify any one device as the earliest computer, partly because the term "computer" has been subject to varying interpretations over time. Originally, the term "computer" referred to a person who performed numerical calculations (a human computer), often with the aid of a mechanical calculating device.

The history of the modern computer begins with two separate technologies - that of automated calculation and that of programmability. Examples of early mechanical calculating devices included the abacus, the slide rule and the Antikythera mechanism (which dates from about 150-100 BC). The hero of Alexandria (c. 10–70 AD) built a mechanical theater which performed a play lasting 10 minutes and was operated by a complex system of ropes and drums that might be considered to be a means of deciding which parts of the mechanism performed which actions and when. This is the essence of programmability.

The "castle clock", an astronomical clock invented by Al-Jazari in 1206, is considered to be the earliest programmable analog computer. It displayed the zodiac, the solar and lunar orbits, a crescent moon-shaped pointer traveling across a gateway causing automatic doors to open every hour, and five robotic musicians that played music when struck by levers operated by a camshaft attached to a water wheel. The length of day and night could be re-programmed every day in order to account for the changing lengths of day and night throughout the year.

The end of the middle Ages saw a re-invigoration of European mathematics and engineering, and Wilhelm Schickard's 1623 device was the first of a number of mechanical calculators constructed by European engineers. However, none of those devices fit the modern definition of a computer because they could not be programmed.

In 1801, Joseph Marie Jacquard made an improvement to the textile loom that used a series of punched paper cards as a template to allow his loom to weave intricate patterns automatically. The resulting Jacquard loom was an important step in the development of computers because the use of punched cards to define woven

patterns can be viewed as an early, albeit limited, form of programmability.

It was the fusion of automatic calculation with programmability that produced the first recognizable computers. In 1837, Charles Babbage was the first to conceptualize and design a fully programmable mechanical computer that he called "The Analytical Engine". Due to limited finances, and an inability to resist continuously changing with the design, Babbage never actually built his Analytical Engine.

Large-scale automated data processing of punched cards was performed for the U.S. Census in 1890 by tabulating machines designed by Herman Hollerith and manufactured by the Computing Tabulating Recording Corporation, which later became IBM. By the end of the 19th century a number of technologies that would later prove useful in the realization of practical computers had begun to appear: the punched card, Boolean algebra, the vacuum tube (thermionic valve) and the teleprinter.

During the first half of the 20th century, many scientific computing needs were met by increasingly sophisticated analog computers, which used a direct mechanical or electrical model of the problem as a basis for computation. However, these were not programmable and generally lacked the versatility and accuracy of modern digital computers.

A succession of steadily more powerful and flexible computing devices were constructed in the 1930s and 1940s, gradually adding the key features that are seen in modern computers. The use of digital electronics (largely invented by Claude Shannon in 1937) and more flexible programmability were vitally important steps, but defining one point along this road as "the first digital electronic computer" is difficult Notable achievements include:

EDSAC was one of the first computers to implement the stored program architecture of von Neumann.

Konrad Zuse's electromechanical "Z machines". The Z3 (1941) was the first working machine featuring binary arithmetic, including floating point arithmetic and a measure of programmability. In 1998 the Z3 was proved to be Turing complete, therefore being the world's first operational computer. The non-programmable Atanasoff–Berry Computer (1941) which used vacuum tube based computation, binary numbers, and regenerative capacitor memory.

The secret British Colossus computers (1943) which had limited programmability but demonstrated that a device using thousands of tubes could be reasonably reliable and electronically reprogrammable. It was used for breaking German wartime codes.

The Harvard Mark I (1944), a large-scale electromechanical computer with limited programmability. The U.S. Army's Ballistics Research Laboratory invented Electronic Numerical Integrator and Calculator, ENIAC (1946), which used decimal arithmetic and is sometimes called the first general purpose electronic computer (since Konrad Zuse's Z3 of 1941 used electromagnets instead of electronics). Initially, however, ENIAC had an inflexible architecture which was not flexible essentially requiring rewiring to change its programming.

Several developers of ENIAC, recognizing its flaws, came up with a far more flexible and elegant design, which came to be known as the "stored program architecture" or Von Neumann architecture. This design was first formally described by John von Neumann in the paper First Draft of a Report on the EDVAC, distributed in 1945. A number of projects to develop computers based on the stored-program architecture commenced around this time, the first of these being completed in Great Britain. The first to demonstrate its workability was the Manchester Small-Scale Experimental Machine (SSEM or "Baby"), while the EDSAC, completed a year after SSEM, was the first practical implementation of the stored program design. Shortly thereafter, the machine originally described by von Neumann's paper—EDVAC—w as completed but did not witness full-time use for an additional two years.

Nearly all modern computers implement some form of the stored-program architecture, making it the single trait by which the word "computer" is now defined. While the technologies used in computers have changed dramatically since the first electronic, general-purpose computers of the 1940s, most still use the von Neumann architecture.

Microprocessors are miniaturized devices that often implement stored program central processing units (CPUs).

Computers that used vacuum tubes as their electronic elements were in use throughout the 1950s. Vacuum tube electronics was largely replaced in the 1960s by transistor-based electronics, which are smaller, faster, cheaper to produce,

require less power, and are more reliable. In the 1970s, integrated circuit technology and the subsequent creation of microprocessors, such as the Intel 4004, further decreased size and cost and further increased speed and reliability of computers. By the 1980s, computers became sufficiently small and cheap to replace simple mechanical controls in domestic appliances such as washing machines. The 1980s also witnessed availabilities of home computers and the now common personal computer. With the evolution of the Internet, personal computers are becoming as common as the television and the telephone in the household. In 2005, Nokia started to call its top-line smart phones of the N-series "multimedia computers" and after the launch of the Apple Phone in 2007, many are now starting to add the smart phone category among "real" computers. In 2008, if the category of smartphones is included in the numbers of computers in the world, the biggest computer maker by units sold is no longer Hewlett-Packard (HP), but Nokia.

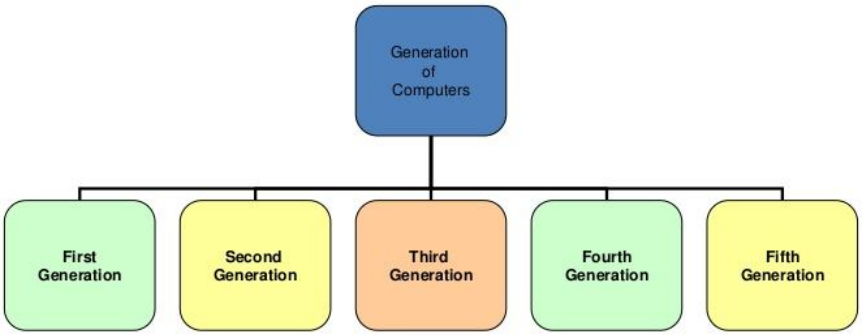
12.4 Generations of computer

Each generation of computer is characterized by a major technological development that fundamentally changed the way computers operate, resulting in increasingly smaller, cheaper, more powerful and more efficient and reliable devices.

The various generations of computers an listed below :

Generation of Computers

Based on the characteristics of various computers developed from time to time, they are categorized as generation of computers.



12.3.1. First Generation (1940-1956): Vacuum Tubes

The first computers used vacuum tubes for circuitry and magnetic drums for memory, and were often enormous, taking up entire rooms. They were very expensive to operate and in addition to using a great deal of electricity, generated a lot of heat, which was often the cause of malfunctions.

First generation computers relied on machine language, the lowest-level programming language understood by computers, to perform operations, and they could only solve one problem at a time. Input was based on punched cards and paper tape, and output was displayed on printouts.

The first computers used vacuum tubes for circuitry and magnetic drums for memory, and were often enormous, taking up entire rooms. They were very expensive to operate and in addition to using a great deal of electricity, generated a lot of heat, which was often the cause of malfunctions.

First generation computers relied on machine language, the lowest-level programming language understood by computers, to perform operations, and they could only solve one problem at a time. Input was based on punched cards and paper tape, and output was displayed on printouts.

In **1946** there was no 'best' way of storing instructions and data in a computer memory. There were four competing technologies for providing computer memory: electrostatic storage tubes, acoustic delay lines (**mercury or nickel**), **magnetic drums** (and disks?), and **magnetic core storage.**

The digital computes using **electronic valves** (Vacuum tubes) are known as first generation computers. the first 'computer' to use electronic valves (ie. vacuum tubes). The high cost of vacuum tubes prevented their use for main memory. They stored information in the form of propagating sound waves.

The vacuum tube consumes a lot of power. The Vacuum tube was developed by Lee DeForest in 1908. These computers were large in size and writing programs on them was difficult. Some of the computers of this generation were:

Mark I : The **IBM Automatic Sequence Controlled Calculator (ASCC)**, called the Mark I by **Harvard University**, was an **electro-mechanical computer**. Mark I is the first machine to successfully perform a long services of **arithmetic and logical operation**. Mark I is the **First Generation** Computer. it was the first operating machine that could **execute long computations automatically**. *Mark I* computer which was built as a partnership between Harvard and **IBM in 1944**. This was the first programmable digital computer made in the U.S. But it was not a purely electronic computer. Instead the Mark I was constructed out of switches, relays, rotating shafts, and clutches. The machine weighed 5 tons, incorporated 500 miles of wire, was 8 feet tall and 51 feet long, and had a 50 ft rotating shaft running its length, turned by a 5 horsepower electric motor.

ENIAC: It was the **first general-purpose electronic computer** built in **1946** at **University of Pennsylvania, USA by John Mauchly and J. Presper Eckert**. The completed machine was announced to the public the evening of **February 14, 1946**. It was named **Electronic Numerical Integrator and Calculator (ENIAC)**. ENIAC contained 17,468 vacuum tubes, 7,200 crystal diodes, 1,500 relays, 70,000 resistors, 10,000 capacitors and around 5 million hand-soldered joints. It weighed more than 30 short tons (27 t), was roughly 8 by 3 by 100 feet (2.4 m × 0.9 m × 30 m), took up 1800 square feet (167 m2), and consumed 150 kW of power. Input was possible from an **IBM card reader**, and an **IBM card punch** was used for output. These cards could be used to produce printed output offline using an IBM accounting machine, such as the **IBM 405**..

EDVAC: It stands for **Electronic Discrete Variable Automatic Computer** and was developed in **1950**.it was to be a vast improvement upon ENIAC, it was **binary** rather than **decimal**, and was a stored program computer. **The concept of storing data and instructions inside the computer was introduced here.** This allowed much faster operation since the computer had rapid access to both data and instructions. The other advantage of storing instruction was that computer could do logical decision internally.

The EDVAC was a **binary serial computer** with automatic addition, subtraction, multiplication, programmed division and automatic checking with an ultrasonic serial memory. EDVAC's **addition time was 864 microseconds** and its **multiplication time was 2900** microseconds (2.9 milliseconds).

The computer had almost 6,000 vacuum tubes and 12,000 diodes, and consumed 56 kW of power. It covered 490 ft² (45.5 m²) of floor space and weighed 17,300 lb (7,850 kg).

EDSAC: It stands for **Electronic Delay Storage Automatic Computer** and was developed by **M.V. Wilkes at Cambridge University in 1949**. Two groups of individuals were working at the same time to develop the first stored-program computer. In the United States, at the University of Pennsylvania the EDVAC (Electronic Discrete Variable Automatic Computer) was being worked on. In England at Cambridge, the EDSAC (Electronic Delay Storage Automatic Computer) was also being developed. The **EDSAC** won the race as the first **stored-program computer** beating the United States' EDVAC by two months. The EDSAC performed computations in the three millisecond range. It performed arithmetic and logical operations without human intervention. The key to the success was in the **stored instructions** which it depended upon solely for its operation. **This machine marked the beginning of the computer age.** EDSAC is the first computer is used to store a program

UNIVAC-1: **Ecker and Mauchly produced it in 1951 by Universal Accounting Computer** setup. it was the **first commercial computer** produced in the United States. It was designed principally by J. Presper Eckert and John Mauchly, the inventors of the ENIAC.

The machine was 25 feet by 50 feet in length, contained 5,600 tubes, 18,000 crystal diodes, and 300 relays. It utilized serial circuitry, 2.25 MHz bit rate, and had an internal storage capacity 1,000 words or 12,000 characters.

It utilized a **Mercury delay line**, magnetic tape, and **typewriter output**. The UNIVAC was used for **general purpose computing** with large amounts of input and output.

The UNIVAC was also the first computer to come equipped with a magnetic tape unit and was the **first computer to use buffer memory**.

Limitations of First Generation Computer

Followings are the major drawbacks of First generation computers.

1. They used valves or vacuum tubes as their main electronic component.
2. They were large in size, slow in processing and had less storage capacity.
3. They consumed lots of electricity and produced lots of heat.
4. Their computing capabilities were limited.
5. They were not so accurate and reliable.
6. They used machine level language for programming.
7. They were very expensive.

Example: ENIAC, UNIVAC, IBM 650 etc

12.3.2 Second Generation (1956-1963): Transistors

Transistors replaced vacuum tubes and ushered in the second generation of computers. The second-generation computer used **transistors** for CPU components & **ferrite cores for main memory** & **magnetic disks** for secondary memory. They used high-level languages such as **FORTRAN (1956), ALGOL (1960) & COBOL (1960 - 1961)**. I/O processor was included to control I/O operations. The transistor was far superior to the vacuum tube, allowing computers to become smaller, faster, cheaper, more energy-efficient and more reliable than their first-generation predecessors. Though the transistor still generated a great deal of heat that subjected the computer to damage, it was a vast improvement over the vacuum tube. Second-generation computers still relied on punched cards for input and printouts for output.

Second-generation computers moved from cryptic binary machine language to symbolic, or assembly, languages, which allowed programmers to specify instructions in words. High-level programming languages were also being developed at this time, such as early versions of COBOL and FORTRAN. These were also the first computers that stored their instructions in their memory, which moved from a magnetic drum to magnetic core technology. The first computers of this generation were developed for the atomic energy industry.

It is in the second generation that the concept of Central Processing Unit (CPU), memory, programming language and input and output units were developed. The programming languages such as COBOL, FORTRAN were developed during this period. Some of the computers of the Second Generation were

1. **IBM 1620**: Its size was smaller as compared to First Generation computers and mostly used for scientific purpose.
2. **IBM 1401:** Its size was small to medium and used for business applications.
3. **CDC 3600**: Its size was large and is used for scientific purposes.

Features:

1. Transistors were used instead of Vacuum Tube.
2. Processing speed is faster than First Generation Computers (Micro Second)
3. Smaller in Size (51 square feet)
4. The input and output devices were faster.

Example: IBM 1400 and 7000 Series, Control Data 3600 etc.

12.3.3 Third Generation (1964-1971): Integrated Circuits

The development of the integrated circuit was the hallmark of the third generation of computers. Transistors were miniaturized and placed on silicon chips, called semiconductors, which drastically increased the speed and efficiency of computers. By the development of a small chip consisting of the capacity of the **300 transistors**. These ICs are popularly known as *Chips*. A single IC has many transistors, registers and capacitors built on a single thin slice of **silicon**. So it is quite obvious that the size of the computer got further reduced

.

Instead of punched cards and printouts, users interacted with third generation computers through keyboards and monitors. These were interfaced with an operating system, which allowed the device to run many different applications at one time with a central program that monitored the memory. Computers for the first time became accessible to many because they were smaller and cheaper than their predecessors.

Some of the computers developed during this period were **IBM-360, ICL-1900, IBM-370, and VAX-750.** Higher level language such as **BASIC (Beginners All purpose Symbolic Instruction Code)** was developed during this period. Computers of this generation were small in size, low cost, large memory and processing speed is very high. Very soon ICs Were replaced by **LSI (Large Scale Integration)**, which consisted about 100 components. An IC containing about 100 components is called LSI.

Features:

1. They used Integrated Circuit (IC) chips in place of the transistors.
2. Semi conductor memory devices were used.
3. The size was greatly reduced, the speed of processing was high, they were more accurate and reliable.
4. Large Scale Integration (LSI) and Very Large Scale Integration (VLSI) were also developed.
5. The mini computers were introduced in this generation.
6. They used high level language for programming.

   Example: IBM 360, IBM 370 etc.

12.3.4 Fourth Generation (1971-Present): Microprocessors

The microprocessor brought the fourth generation of computers, as it became possible to build thousands of integrated circuits were built onto a single silicon chip. An IC containing about 100 components is called LSI (Large Scale Integration) and the one, which has more than 1000 such components, is called as **VLSI (Very Large Scale Integration)**. It uses *large scale Integrated Circuits* (LSIC) built on a single silicon chip called microprocessors. Due to the development of microprocessor it is possible to place computer's *central*

processing unit (CPU) on single chip. These computers are called microcomputers. Later *very large scale Integrated Circuits* (VLSIC) replaced LSICs.

What in the first generation filled an entire room could now fit in the palm of the hand.

The Intel 4004 chip, developed in 1971, located all the components of the computer - from the central processing unit and memory to input/output controls - on a single chip.

In 1981 IBM introduced its first computer for the home user, and in 1984 Apple introduced the Macintosh. Microprocessors also moved out of the realm of desktop computers and into many areas of life as more and more everyday products began to use microprocessors.

As these small computers became more powerful, they could be linked together to form networks, which eventually led to the development of the Internet. Fourth generation computers also saw the development of Graphical User Interface (GUI), the mouse and handheld devices.

Thus the computer which was occupying a very large room in earlier days can now be placed on a table. The personal computer (PC) that we see in our home is a Fourth Generation Computer Main memory used fast semiconductors chips up to 4 M bits size. Hard disks were used as secondary memory. Keyboards, dot matrix printers etc. were developed. OS-such as **MS-DOS, UNIX, Apple's Macintosh** were available. Object oriented language, **C++ etc** were developed.

Features:

1. They used Microprocessor (VLSI) as their main switching element.
2. They are also called as micro computers or personal computers.
3. Their size varies from desktop to laptop or palmtop.
4. They have very high speed of processing; they are 100% accurate, reliable, diligent and versatile.
5. They have very large storage capacity.

   Example: IBM PC, Apple-Macintosh etc.

12.3.5 Fifth Generation - Present and Beyond: Artificial Intelligence

Fifth generation computing devices, based on artificial intelligence, are still in

development, though there are some applications, such as voice recognition, that are already being used today. The use of parallel processing and superconductors is helping to make artificial intelligence a reality. Quantum computation and molecular and nanotechnology will radically change the face of computers in years to come. The goal of fifth-generation computing is to develop devices that respond to natural language input and are capable of learning and self-organization.

5th generation computers use ULSI (Ultra-Large Scale Integration) chips. Millions of transistors are placed in a single IC in ULSI chips. 64 bit microprocessors have been developed during this period. Data flow & EPIC architecture of these processors have been developed. RISC & CISC, both types of designs are used in modern processors. Memory chips and flash memory up to 1 GB, hard disks up to 600 GB & optical disks up to 50 GB have been developed.

Computer Generations

Generation	Device	Hardware feature	Characteristics	System names
First		► Vacuum Tubes ► Punch Cards	► Support machine laguage only ► Very costly ► Generate lot of heat ► Hhuge size ► Consumed lot of electricity	► ENIAC ► EDVAC ► TBM 701
Second		► Transistors ► Magnetic Tapes	► Batch operating system ► Faster, smaller and reliabe than previous generation ► Costly	► Honeywell 400 ► CDC 1604 ► IBM 7030
Third		► ICs ► Large capacity disk and Magnetic Tapes	► Time Sharing OS ► Faster, smaller and reliabe cheaper ► Easier to update	► IBM 360/370 ► CDC 6600 ► PDP 8/11
Fourth		► ICs with VLSI Technology ► Semiconductor Memory ► Magnetic tapes and floppy as portable	► Multiprocessing & GUI OS ► Object oriented programs ► Small, affordable, easy to Use ► Easier to update	► Apple II ► VAX 9000 ► CRAY 1/2
Fifth		► ICs with ULSI Technology ► Large capacity hard disk with RAID Support ► Optical disks as portable read-only storage media ► powerful servers, internet, Cluster computing	► Powerful, cheaper, reliable, easy to use, portable ► Rapid software development possible	► IBM ► Pentium ► PARAM

12.4. Fundamentals of computer

COMPUTER SYSTEM = HARDWARE + SOFTWARE+ USER

12.4.1 Hardware

The hardware is the physical and visible components of the computer system. Some of these hardware components are: Input devices e.g. Keyboard, Mouse, Touch Screen, Light-pen, Joystick, Digitizer Tablet, and Voice Input Device. Output devices such as Monitor, Plotters, Voice Output Device, Computer Output Microfilm (COM) and Printers. Memory e.g. ROM (Read Only Memory) and RAM (Random Access Memory) and Storage Devices such as Magnetic Disk, Magnetic Tape, Optical Disk, Flash Drives and Floppy Disks.

12.4.2. Software

A computer system without software is like the human body without life in it, or like a car without fuel. Without software, what is left of a computer system is hardware that is lifeless. The hardware cannot be used without software. Software directs the hardware by telling it what to do, how to do it, and when to do it. Both are necessary for the computer system to be useful.

Basically, software is programs which enable the hardware components to operate effectively as well as making it provide many useful services. The term, software are usually applied to programs that are written by computer manufacturers and software specialists. Programs are sequences of instructions given to computers to solve a given problem or accomplish a given task. Examples of software are: System Software e.g. Operating System such as windows, Linux etc Utility and Service Programs, Translators e.g. BASIC and Database Management System. And Application Software such as Microsoft Word, Microsoft Access, MS-Publisher, CorelDraw and Statistical Package for Social Scientists (SPSS) etc.

12.4.3. Basic computer Operations

A computer can process data, pictures, sound and graphics. They can solve highly complicated problems quickly and accurately. A computer as shown in Fig. performs basically five major computer operations or functions irrespective of their size and make. These are

1) It accepts data or instructions by way of input,

2) It stores data,

3) It can process data as required by the user,

4) It gives results in the form of output, and

5) It controls all operations inside a computer.

We discuss below each of these Computer operations

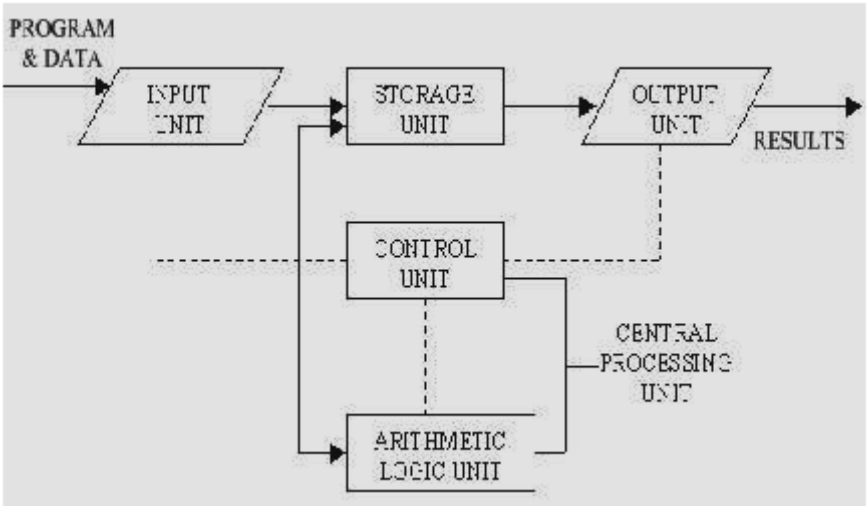


Fig : *Basic computer Operations*

1. **Input:** This is the process of entering data and programs in to the computer system. We should know that computer is an electronic machine like any other machine which takes as inputs raw data and performs some processing giving out processed data. Therefore, the input unit takes data from us to the computer in an organized manner for processing.

2. **Storage:** The process of saving data and instructions permanently is known as storage. Data has to be fed into the system before the actual processing starts. It is because the processing speed of Central Processing Unit (CPU) is so fast that the data has to be provided to CPU with the same speed. Therefore the data is first stored in the storage unit for faster access and processing. This storage unit or the primary storage of the computer system is designed to do the above functionality. It provides space for storing data and instructions.

The storage unit performs the following major functions:

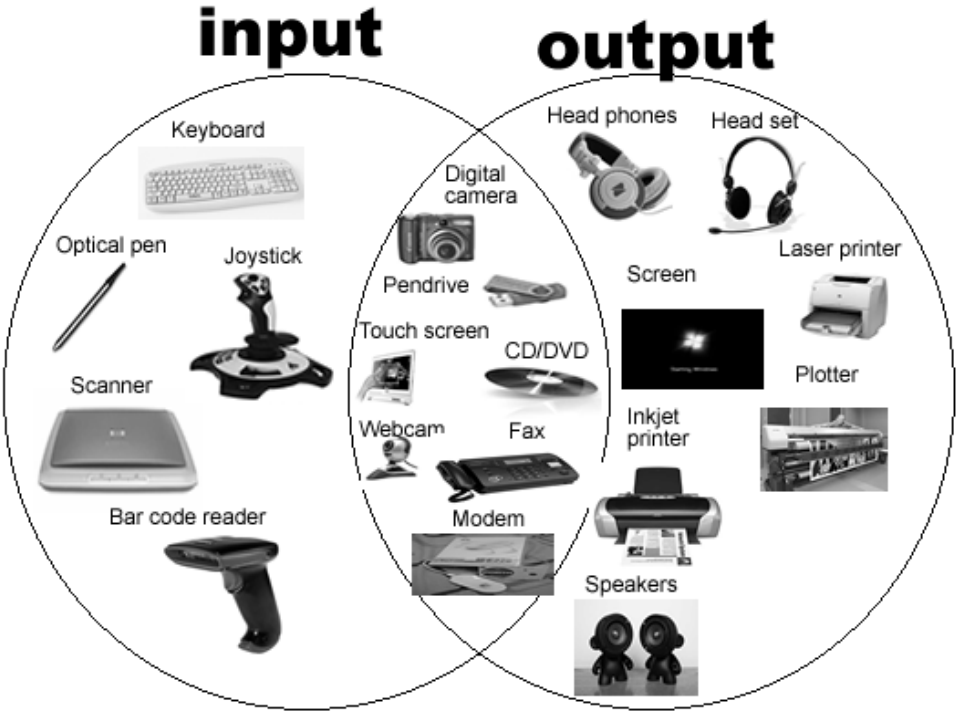
- All data and instructions are stored here before and after processing.

- Intermediate results of processing are also stored here.

3. **Processing:** The task of performing operations like arithmetic and logical operations is called processing. The Central Processing Unit (CPU) takes data and instructions from the storage unit and makes all sorts of calculations based on the instructions given and the type of data provided. It is then sent back to the storage unit.

4. **Output:** This is the process of producing results from the data for getting useful information. Similarly the output produced by the computer after processing must also be kept somewhere inside the computer before being given to we in human readable form. Again the output is also stored inside the computer for further processing.

5. **Control:** The manner how instructions are executed and the above operations are performed. Controlling of all operations like input, processing and output are performed by control unit. It takes care of step by step processing of all operations inside the computer.



12.4.4. Basic Functional Units

In order to carry out the operations mentioned in the previous section the computer allocates the task between its various functional units. The computer system is divided into three separate units for its operation. They are

12.4.4.1. Arithmetic Logical Unit

12.4.4.2. Control Unit.

12.4.4.3. Central Processing Unit.

12.4.4.1 Arithmetic Logical Unit (ALU) Logical Unit

Logical Unit : After we enter data through the input device it is stored in the primary storage unit. The actual processing of the data and instruction are performed by Arithmetic Logical Unit. The major operations performed by the ALU are addition, subtraction, multiplication, division, logic and comparison. Data is transferred to ALU from storage unit when required. After processing the output is returned back to storage unit for further processing or getting stored.

12.4.4.2 Control Unit (CU)

The next component of computer is the Control Unit, which acts like the supervisor seeing that things are done in proper fashion. Control Unit is responsible for coordinating various operations using time signal. The control unit determines the sequence in which computer programs and instructions are executed. Things like processing of programs stored in the main memory, interpretation of the instructions and issuing of signals for other units of the computer to execute them. It also acts as a switch board operator when several users access the computer simultaneously. Thereby it coordinates the activities of computer's peripheral equipment as they perform the input and output.

A *control unit* (or *controller,* same thing) is a piece of *hardware* that manages the activities of *peripherals* (separate devices attached to the computer, such as monitors, hard drives, printers, etc.) Control units found on personal computers are usually contained on a single printed circuit board. The control unit acts as a sort of "go-between," executing transfers of information between the computer's memory and the peripheral. Although the *CPU (central processing unit-the* "big boss" in the computer) gives instructions to the controller, it is the control unit itself that performs the actual physical transfer of data.

The control unit fetches one or more new instructions from memory (or an INSTRUCTION CACHE), DECODES them and dispatches them to the appropriate FUNCTION UNITS to be executed. The control unit is also responsible for setting the LATCHES in various data paths that ensure that the instructions are performed on the correct operand values stored in the REGISTERS.

In a CISC processor the control unit is a small processor in its own right that executes MICROCODE programs stored in a region of ROM that prescribe the correct sequence of latches and data transfers for each type of macroinstruction. A RISC processor does away with microcode and most of the complexity in the control unit, which is left with little more to do than decode the instructions and turn on the appropriate function units.

12.4.4.3 Central Processing Unit (CPU)

The ALU and the CU of a computer system are jointly known as the central processing unit. We may call CPU as the brain of any computer system. It is just like brain that takes all major decisions, makes all sorts of calculations and directs different parts of the computer functions by activating and controlling the operations.

Every things computer does is controlled by its **Central Processing Unit** (CPU). The CPU is the **brains of the computer**. Sometimes referred to simply as the **central processor** or **Nerve Centre** or **heart**, but more commonly called **processor**, the CPU is where most calculations take place.

In terms of computing power, the CPU is the most important element of a computer system. It add and compare its data in CPU chip. A CPU or Processors of all computers, whether micro, mini or mainframe must have three element or parts primary storage, arithmetic logic unit (ALU), and control unit. *Control Unit* (CU) - *decodes the program instruction. CPU chip used in a computer is partially made out of Silica. on other words silicon chip used for data processing are called Micro Processor.*

Central processing unit (CPU) is the central component of the Pc. Sometimes it is called as processor. It is the brain that runs the show inside the Pc. All work that is done on a computer is performed directly or indirectly by the processor. Obviously, it is one of the most important components of the Pc. It is also,

scientifically, not only one of the most amazing parts of the PC, but one of the most amazing devices in the world of technology.

12.4.5 Functions of Processor

The processor plays a significant role in the following important aspects of our computer system;

1. **Performance:** The processor is probably the most important single determinant of system performance in the Pc. While other components also playa key role in determining performance, the processor's capabilities dictate the maximum performance of a system. The other devices only allow the processor to reach its full potential.
2. **Software Support:** Newer, faster processors enable the use of the latest software. In addition, new processors such as the Pentium with MMX Technology, enable the use of specialized software not usable on earlier machines.
3. **Reliability and Stability:** The quality of the processor is one factor that determines how reliably our system will run. While most processors are very dependable, some are not. This also depends to some extent on the age of the processor and how much energy it consumes.
4. **Energy Consumption and Cooling:** Originally processors consumed relatively little power compared to other system devices. Newer processors can consume a great deal of power. Power consumption has an impact on everything from cooling method selection to overall system reliability.
5. **Motherboard Support:** The processor that decides to use in our system will be a major determining factor in what sort of chipset we must use, and hence what motherboard we buy. The motherboard in turn dictates many facets of. The system's capabilities and performance.

12.4.6 Coprocessor

A **coprocessor** is a *chip* that works side-by-side with the computer's main *processor* (the chip called the *central processing unit,* or *CPU).* The coprocessor handles some of the more specialized tasks, such as doing math calculations or displaying graphics on the screen, thereby taking some of the work load off the main processor so it can go on with the business of directing and keeping order

over the whole show. A coprocessor is installed to reduce the burden on a computer's CPU and thus free it for more general duties such as transferring data and handling multiple tasks.

Math coprocessors, for example, are specialized for performing calculations on numbers, and they are much faster at it than the main processor in our computer. So if we have a program that does many math calculations, such as a *spreadsheet* or a *CAD* program, then adding a math coprocessor to our system can sometimes remarkably improve our computing speed.

There are video coprocessors that are used to speed up the display of graphics on our screen. Again, if we use any graphics-based application, including Windows, then adding a video coprocessor to our system on an *add-in board* can speed up our system even more than buying a faster computer.

A coprocessor may be designed to work just with a particular type of CPU, in which case its instructions can be included in the main program and are passed on to the coprocessor by the CPU as it encounters them. In other cases, the coprocessor may require its own separate program and program memory, and communicates with the CPU by interrupts or message passing via a shared memory region.

12.4.7. BIOS (basic input/output system)

A **BIOS** (**Basic Input/Output System**) Short for ROM is boot firmware program that a computer uses to successfully start operating. The BIOS is located on a chip inside of the computer and is designed in a way that protects it from disk failure.

When we turn on a PC, the BIOS first conduct a basic hardware check, called a Power-On Self Test (POST), to determine whether all of the attachments are present and working. Then it loads the operating system into our computer's random access memory, or RAM. The BIOS also manages data flow between the computer's operating system and attached devices such as the hard disk, video card, keyboard, mouse, and printer. The BIOS stores the date, the time, and our system configuration information in a battery-powered, non-volatile memory chip, called a CMOS (Complementary Metal Oxide Semiconductor) after its manufacturing process. The main functions of the BIOS are:

(i) BIOS power on self Test (POST)

(ii) Bootstrap loader

(iii) BIOS Setup utility program

(iv) System service routines

12.4.8. SemiConductor

Semiconductors are widely used in electronics to make components such as diodes, transistors, thyristors, integrated circuits as well as semiconductor lasers.

A semiconductor is usually a solid chemical element or compound that can conduct electricity under some conditions, making it a good medium for the control of electrical current. Its conductance varies depending on the current or voltage applied to a control electrode. Semiconductor is use for manufacturing chips.

A semiconductor device can perform the function of a vacuum tube having hundreds of times its volume. A single integrated circuit (IC), such as a microprocessor chip, can do the work of a set of vacuum tubes that would fill a large building and require its own electric generating plant.

12.4.9. Booting

To **boot** or **boot up** means to start our computer system, usually by turning on the power and/or pushing the "on" button. It's called "booting" because the computer is going inside itself and turning itself on (doing a lot of preliminary checking and adjusting before it's ready to run our programs). Hence the machine is considered to be "pulling itself up by its own bootstraps."

When the computer is first turned on or restarted, it reads the startup instructions found in the ROM BIOS chips. These instructions tell the computer to check the system over (a series of tests called the POST). Certain information (such as the amount of memory and the number and type of disk drives) about the PC is stored in a special chip called CMOS, and that information is also verified during boot. The last thing that happens during boot is the loading of the operating system, which is found on the hard disk drive or on a floppy disk in drive A. The computer cannot do anything without first loading an operating system into memory, because it's the operating system that manages all of the computer's basic functions.

Information in the operating system files continues the booting process. During a PC boot, the CONFIG.SYS file is located, and its instructions are executed. The CONFIG.SYS is a special file that fine-tunes the PC, customizing it so it can

access optional peripherals (such as the mouse or the modem) and unused areas in memory. Next, the AUTO EXECBAT file is located, and its instructions are executed. The AUTOEXEC BAT file contains commands (such as those to start a particular program or change the prompt) that the user wants run at boot. Once the startup files have been found and executed, the computer is fully booted and ready to go.

12.4.10 Data and Information

The words **Data** and **Information** may look similar and many people use these words very frequently, But both have lots of differences between them.

Data: Data are plain facts. The word "data" is plural for "datum." When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information.

12.4.11. Computer Languages

a) Low Level Language

i) Machine Level Language

ii) Assembly Language

Machine language: These language instructions are directly executed by CPU

Assembly language: The endeavor of giving machine language instructions a name structure that means bit strings of instructions of machine language are given name here

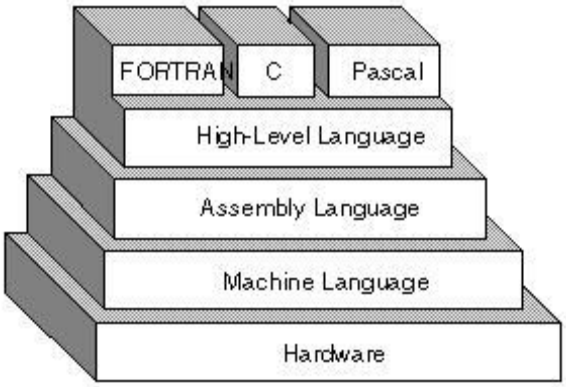
High Level Language: The user friendly language ...more natural language than assembly language.

Assembler is needed to convert assembly language into machine language

Complier is needed to convert high level to machine language

b) High Level Language

COBOL (COmmon Business Oriented Language), FORTRAN (FORmula TRANslation), BASIC (Beginner's All-purpose Symbolic Instruction Code), C, C++ etc. are the examples of High Level Language.



12.4.12 Bad Sector

A disk has two sides (a top and a bottom). Each side of the disk has tracks or concentric rings on the surface. Each ring is divided like a pie into equal wedges, or sectors, which are the smallest units of storage space on the disk. If one of these units is damaged or flawed, it is considered a **bad sector** and cannot be used.

If there was already data in that sector when it got damaged, chances are slim that we can recover that data unless we have the specialized hardware and software necessary for that sort of operation. Almost all hard disks are born with bad sectors, so don't freak out if our software utility reports them. Other bad sectors should not start appearing, though, after we start using the disk.

12.4.13 Applications of computers-

1. Education :

Getting the right kind of information is a major challenge as is getting information to make sense. College students spend an average of 5-6 hours a week on the internet.Research shows that computers can significantly enhance performance in learning. Students exposed to the internet say they think the web has helped them improve the quality of their academic research and of their written work. One revolution in education is the advent of distance learning. This offers a variety of internet and video-based online courses.

2. **Health and Medicine :**

Computer technology is radically changing the tools of medicine. All medical information can now be digitized. Software is now able to computer the risk of a disease. Mental health researchers are using computers to screen troubled teenagers in need of psychotherapy. A patient paralyzed by a stroke has received an implant that allows communication between his brain and a computer; as a result, he can move a cursor across a screen by brainpower and convey simple messages.

3. **Science :**

Scientists have long been users of it. A new adventure among scientists is the idea of a "collaboratory", an internet based collaborative laboratory, in which researchers all over the world can work easily together even at a distance. An example is space physics where space physicists are allowed to band together to measure the earth's ionosphere from instruments on four parts of the world.

4. **Business :**

Business clearly see the interest as a way to enhance productivity and competitiveness. Some areas of business that are undergoing rapid changes are sales and marketing, retailing, banking, stock trading, etc. Sales representatives not only need to be better educated and more knowledgeable about their customer's businesses, but also must be comfortable with computer technology. The internet has become a popular marketing tool. The world of cybercash has come to banking – not only smart cards but internet banking, electronic deposit, bill paying, online stock and bond trading, etc.

5. **Recreation and Entertainment:**

Our entertainment and pleasure-time have also been affected by computerization. For example:

i) In movies, computer generated graphics give freedom to designers so that special effects and even imaginary characters can play a part in making movies, videos, and commercials.

ii) In sports, computers compile statistics, sell tickets, create training programs and diets for athletes, and suggest game plan strategies based on the competitor's past performance.

iii) In restaurants, almost everyone has eaten food where the clerk enters an order by indicating choices on a rather unusual looking cash register; the device directly enters the actual data into a computer, and calculates the cost and then prints a receipt.

6. **Government:**

Various departments of the Government use computer for their planning, control and law enforcement activities. To name a few – Traffic, Tourism, Information & Broadcasting, Education, Aviation and many others.

7. **Defence:**

There are many uses computers in Defence such as:

1) Controlling UAV or unmanned air-crafts an example is Predator. If we have cable I would recommend watching the shows “Future Weapons" and “Modern Marvels". The show future weapon gives an entire hour to the predator.
2) They are also used on Intercontinental Ballistic Missiles (ICBMs) that uses GPS and Computers to help the missile get to the target.
3) Computers are used to track incoming missiles and help slew weapons systems onto the incoming target to destroy them.
4) Computers are used in helping the military find out where all their assets are (Situational Awareness) and in Communications/Battle Management Systems.
5) Computers are used in the logistic and ordering functions of getting equipments to and around the battlefield.
6) Computers are used in tanks and planes and ships to target enemy forces, help run the platform and more recently to help diagnose any problems with the platforms.
7) Computers help design and test new systems.

8. **Sports:**

In today's technologically growing society, computers are being used in nearly every activity.

9. **Recording Information**

Official statistics keepers and some scouts use computers to record statistics, take notes and chat online while attending and working at a sports event.

10. **Analyzing Movements**

The best athletes pay close attention to detail. Computers can slow recorded video and allow people to study their specific movements to try to improve their tendencies and repair poor habits.

11. **Writers**

Many sportswriters attend several sporting events a week, and they take their computers with them to write during the game or shortly after while their thoughts are fresh in their mind.

12. ***Scoreboard***

While some scoreboards are manually updated, most professional sports venues have very modern scoreboards that are programmed to update statistics and information immediately after the information is entered into the computer.

13. **Safety**

Computers have aided in the design of safety equipment in sports such as football helmets to shoes to mouth guards

12.5 Summary

1. Hardware = Internal Devices + Peripheral Devices
2. All physical parts of the computer (or everything that we can touch) are known as Hardware.
3. Software = Programs
4. Software gives "intelligence" to the computer.
5. USER = Person, who operates computer.
6. The technology that enables today's computer industry is called electronics.

   Electronics is concerned with the behaviour and effects of electrons as they pass through devices that can restrict their flow in various ways. The vacuum tube was the earliest electronic device.
7. The first successful large-scale electronic digital computer, the ENIAC, laid the foundation for the modern computer industry.

8. The stored-program concept fostered the computer industry's growth because it enabled customers to change the computer's function easily by running a different program.
9. First-generation computers used vacuum tubes and had to be programmed in difficult-to-use machine languages.
10. Second-generation computers introduced transistors and high-level programming languages, such as COBOL and FORTRAN.
11. Third-generation computers introduced integrated circuits, which cut costs and launched the minicomputer industry. Key innovations included timesharing, wide area networks, and local area networks.
12. Fourth-generation computers use microprocessors. Key innovations include personal computers, the graphical user interface, and the growth of massive computer networks.
13. As computers become more powerful and less expensive, the rise of global networking is making them more valuable. The combination of these two forces is driving major changes in every facet of our lives.

12.6 Self-Learning Exercise

1. Define Computer. Explain its various components in detail.
2. Explain classification of computers.
3. Write an essay on "History of Computers"
4. Explain all generations of computers.
5. Write short notes on
   a. Arithmetic Logical Unit
   b. Control Unit.
   c. Central Processing Unit
   d. Hardware
   e. Software
   f. Booting

g. Semiconductors

h. Bad Sector

i. Computer Languages

6. What are the functions of Processors?

7. What is the difference between the data and information?

8. What are the applications of computers in our daily life?

12.7 Reference Books

- Ayo C.K (1994) “Computer Literacy: Operation and Ap preciation”.
- Henry.C.Lucas, Jr. (2001) 'Information Technology'; Tata Mc Graw Hill Publication Company Limited, New Delhi.
- D.P.Sharma (2008); 'Information Technology'; College Book Centre, Jaipur.
- P.K.Sinha, Priti Sinha,(2007) 'Computer Fundamentals'; BPB Publication, New Delhi.
- Meyers, Jeremy, "A Short History of the Computer" [Online] Available <http://www.softlord.com/comp/>.
- http://www.ieeeghn.org/wiki/images/5/57/Onifade.pdf

Unit-13

Computer peripherals and architecture, elementary idea about operating system, DOS and window environment, Applications of MS-Office

Structure of the Unit

13.0. Objectives

After completing the unit, you will be able to understand about:

- Different Input, output and storage computer peripherals
- Brief idea about computer architecture
- Brief idea about types and functions of operating system
- Features of MS-DOS and MS Windows
- Applications of MS-Office- Excel for various Statistical Analysis
- Brief idea about MS-power point

13.1. Computer peripherals and architecture

13.1.1. Introduction

A **peripheral** is a "device that is used to put information into or get information out of the computer."

There are three different types of peripherals:-

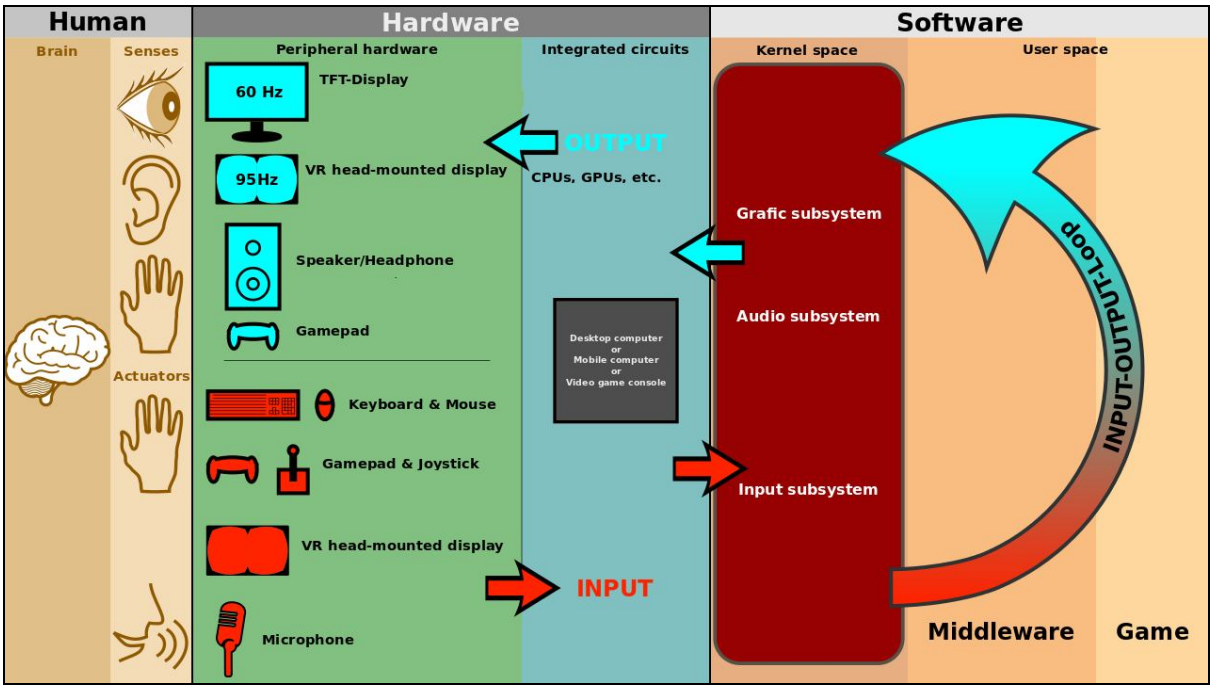
1. Input peripheral is used to interact with, or send data to the computer.
2. Output is a peripheral, which provides output to the user from the computer.
3. Storage is a peripheral, which stores data processed by the computer.

A peripheral device is generally defined as any auxiliary device such as a computer mouse or keyboard that connects to and works with the computer in some way. Other examples of peripherals are expansion cards, graphics cards, image scanners, tape drives, microphones, loudspeakers, webcams, and digital cameras. New devices such as digital watches, smart phones and tablet computers have interfaces which allow them to be used as a peripheral by a full computer, though they are not host-dependent as other peripheral devices are.

Devices that exist outside the computer case are called external peripherals, or auxiliary components, Examples are: Scanner and printer, connect to the peripheral ports on the back of computer. Devices that are inside the case such as internal hard drives or CD-ROM drives are also peripherals in technical terms and are called internal peripherals.

In a system on a chip, peripherals are incorporated into the same integrated circuit as the central processing unit. They are still referred to as "peripherals" despite being permanently attached to their host processor.

RAM - random access memory - straddles the line between peripheral and primary component; it is technically storage peripheral, but is required for every major function of a modern computer and removing the RAM will effectively disable any modern machine



13.1.2. Common peripherals

Input peripherals

1. Keyboard
2. Computer mouse
3. Graphic tablet
4. Touch screen
5. Barcode reader
6. Image scanner
7. Microphone
8. Webcam
9. Light pen
10. Scanner

Output peripherals

1. Computer display
2. Printer

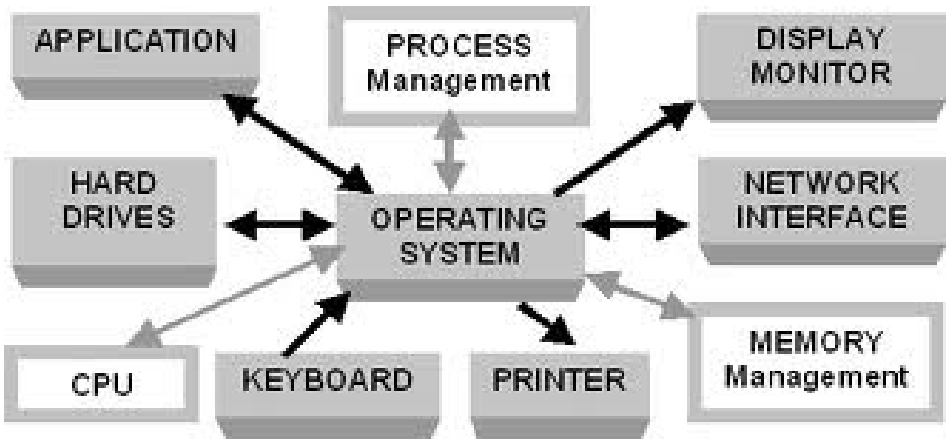
3. Projector
4. Speaker

Storage peripherals

1. Floppy disk drive
2. Flash drive
3. Disk drive
4. CD/DVD drive
5. Smartphone or Tablet computer storage interface

Both Input and Output peripherals

1. Modem
2. Network interface controller (NIC)



13.1.3. Input peripherals

Any device that allows information from outside the computer to be communicated to the computer is considered an input peripheral. All computer input peripherals and circuitry must eventually communicate with the computer in discrete binary form because CPU (Central Processing Unit) of a computer can understand only discrete binary information

A few common computer input peripherals are

1. Computer keyboard

A keyboard is a typewriter-style device, which uses an arrangement of buttons or keys, to act as mechanical levers or electronic switches. In past there was use of punch cards and paper tape, interaction via teleprinter-style keyboards later on keyboard became the main input device for computers.

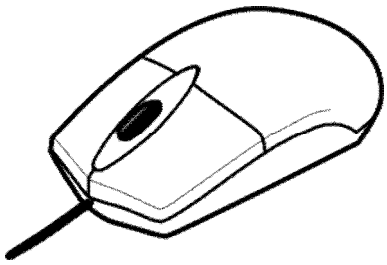
A keyboard typically has characters engraved or printed on the keys and each press of a key typically corresponds to a single written symbol. However, to produce some symbols requires pressing and holding several keys simultaneously or in sequence. While most keyboard keys produce letters, numbers or signs (characters), other keys or simultaneous key presses can produce actions or execute computer commands.

A computer keyboard distinguishes each physical key from every other and reports all key presses to the controlling software. Keyboards are also used for computer gaming, either with regular keyboards or by using keyboards with special gaming features, which can expedite frequently used keystroke combinations. A keyboard is also used to give commands to the operating system of a computer.

2. Mouse

A computer mouse with the most common standard features: two buttons and a scroll wheel, which can also act as a third button. In computing, a mouse is a pointing device that detects two-dimensional motion relative to a surface. This motion is typically translated into the motion of a pointer on a display, which allows for fine control of a graphical user interface.

Physically, a mouse consists of an object held in one's hand, with one or more buttons. Mice often also feature other elements, such as touch surfaces and "wheels", which enable additional control and dimensional input.



3. Graphics tablet

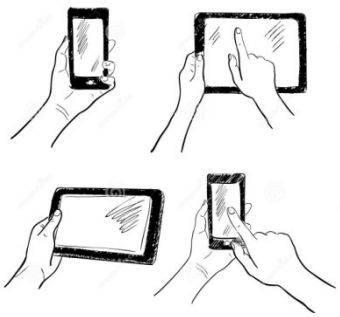
A graphics tablet or digitizer is a computer input device that enables a user to hand-draw images, animations and graphics, similar to the way a person draws images with a pencil and paper. These tablets may also be used to capture data or handwritten signatures. It can also be used to trace an image from a piece of paper which is taped or otherwise secured to the surface. Capturing data in this way, by tracing or entering the corners of linear poly-lines or shapes, is called digitizing.

The device consists of a flat surface upon which the user may "draw" or trace an image using an attached stylus, a pen-like drawing apparatus.

4. Touch screen

A touch screen is an electronic visual display that the user can control through simple or multi-touch gestures by touching the screen with a special stylus (pen) and-or one or more fingers. The user can use the touch screen to react to what is displayed and to control how it is displayed.

The touch screen enables the user to interact directly with what is displayed, rather than using a mouse, touchpad, or any other intermediate device. Touch screens are common in devices such as game consoles, personal computers, tablet computers, and smart phones. They also play a prominent role in the design of digital appliances such as personal digital assistants (PDAs), satellite navigation devices, mobile phones, and video games and Electronic books.



5. Barcode reader

A barcode reader is an electronic device for reading printed barcodes. It consists of a light source, a lens and a light sensor translating optical impulses into electrical ones. All barcode readers contain decoder circuitry analyzing the barcode's image data provided by the sensor and sending the barcode's content to the scanner's output port.

6. Image scanner

An image scanner is a device that optically scans images, printed text, handwriting, or an object, and converts it to a digital image. Modern scanners typically use a charge-coupled device (CCD) or a contact image sensor (CIS) as the image sensor, whereas *drum scanners*, developed earlier and still used for the highest possible image quality, use a photomultiplier tube (PMT) as the image sensor. A *rotary scanner,* used for high-speed document scanning, is a type of drum scanner that uses a CCD array instead of a photomultiplier. Non-contact planetary scanners essentially photograph delicate books and documents.

In the life sciences research area, detection devices for DNA microarrays are called scanners as well. These scanners are high-resolution systems (up to 1 μm/ pixel), similar to microscopes. The detection is done via CCD or a photomultiplier tube

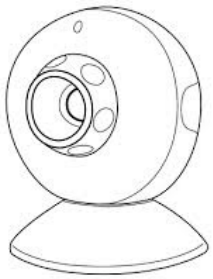
7. Microphone

A microphone, is an acoustic-to-electric transducer or sensor that converts sound in air into an electrical signal. Microphones are used in many applications such as telephones, hearing aids, public address systems for concert halls and public events, motion picture production, live and recorded audio engineering, two-way radios, megaphones, radio and television broadcasting, and in computers for recording voice, speech recognition for non-acoustic purposes such as ultrasonic checking or knock sensors.

8. Webcam

A webcam is a video camera that feeds or streams its image in real time to or through a computer to computer network. In computer, the video stream is captured by the webcam which may be saved, viewed or sent on to other networks via systems such as the internet, and email as an attachment.

Their most popular use is the establishment of video links, permitting computers to act as videophones or videoconference stations. Other popular uses include security surveillance, computer vision, video broadcasting, and for recording social videos.



9. Light pen

A light pen is a computer input device in the form of a light-sensitive wand used in conjunction with a computer's cathode ray tube display. It allows the user to point to displayed objects or draw on the screen in a similar way to a touch screen but with greater positional accuracy. A light pen detects a change of brightness of nearby screen pixels when scanned by cathode ray tube electron beam and communicates the timing of this event to the computer. Since a CRT scans the entire screen one pixel at a time, the computer can keep track of the expected time of scanning various locations on screen by the beam and infer the pen's position from the latest timestamp.

13.1.4. Output peripherals

An output peripheral is a device which accepts results from the computer and displays them to user. The output peripheral also converts the binary code obtained from the computer into human readable form.

1. Computer monitor

A monitor or a display is an electronic visual display for computers. The monitor comprises the display device, circuitry and an enclosure. The display device in modern monitors is typically a thin film transistor liquid crystal display (TFT-

LCD) thin panel, while older monitors used a cathode ray tube (CRT) about as deep as the screen size.

In past computer monitors were used for data processing while television receivers were used for entertainment. From the 1980s onwards, computers (and their monitors) have been used for both data processing and entertainment, while televisions have implemented some computer functionality.

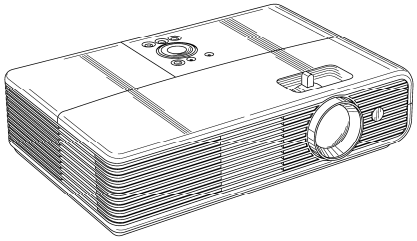
2. Printer

A printer is a peripheral which makes a persistent human-readable representation of graphics or text on paper or similar physical media. The two most common printer mechanisms are black and white laser printers used for common documents, and color ink jet printers which can produce high-quality photograph-quality output.

3. Projector

A projector or image projector is an optical device that projects an image or videos onto a surface, commonly a projection screen. Most projectors create an image by shining a light through a small transparent lens or can be project the image directly, by using lasers. A virtual retinal display, or retinal projector, is a projector that projects an image directly on the retina instead of using an external projection screen.

The most common type of projector is a video projector. Movie theaters use a type of projector called a movie projector. The newest types of projectors are handheld projectors that use lasers or LEDs to project images. Their projections are hard to see if there is too much ambient light.



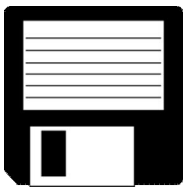
4. Computer speaker

Computer speakers are speakers external to a computer that disable the lower fidelity built-in speaker. They often have a low-power internal amplifier. The computer speakers typically packaged with computer systems are small, plastic, and have mediocre sound quality. Some computer speakers have equalization features such as bass and treble controls. The internal amplifiers require an external power source, usually an AC adapter. They can have a subwoofer unit, to enhance bass output, and these units usually include the power amplifiers both for the bass speaker, and the small satellite speakers.

13.1.5. Storage peripherals

1. Floppy disk

A floppy disk (diskette) is a disk storage medium composed of a disk of thin and flexible magnetic storage medium, sealed in a rectangular plastic carrier lined with fabric that removes dust particles. Floppy disks are read and written by a floppy disk drive. By 2010, computer motherboards were rarely manufactured with floppy drive support. While floppy disk drives now have some limited uses.



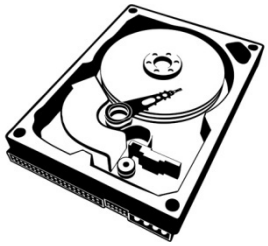
2. USB flash drive

A USB flash drive, is a data storage device that includes flash memory with an integrated Universal Serial Bus (USB) interface. USB flash drives are typically removable and rewritable, and physically much smaller than an optical disc. A one-terabyte (TB) drive was launched in 2013.

USB flash drives are often used for the same purposes for which floppy disks or CDs were used, i.e., for storage, data back-up and transfer of computer files. They are smaller, faster, have thousands of times more capacity, and are more durable and reliable because they have no moving parts. Additionally, they are immune to magnetic interference (unlike floppy disks), and are unharmed by surface scratches (unlike CDs). Some devices combine the functionality of a digital audio player with USB flash storage; they require a battery only when used to play music.

3. Disk storage

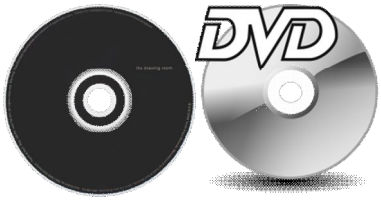
Disk storage is a general category of storage mechanisms where data are recorded by various electronic, magnetic, optical, or mechanical changes to a surface layer of one or more rotating disks. A disk drive is a device implementing such a storage mechanism and is usually distinguished from the disk medium. Notable types are the hard disk drive (HDD) containing a non-removable disk, the floppy disk drive (FDD) and its removable floppy disk, and various optical disc drives and associated optical disc media.



4. CD & DVD

A CD is a pre-pressed optical compact disc which contains data. The name stands for "Compact Disc". Computers can read CD but cannot be erasable. Until the mid-2000s, CDs were popularly used to distribute software for computers and video game consoles.

DVD (digital video disc) is a digital optical disc storage format, invented and developed by Philips, Sony, Toshiba, and Panasonic in 1995. DVDs can be played in multiple types of players, including DVD players. DVDs offer higher storage capacity than compact discs while having the same dimensions.



5. Smartphone

A **Smartphone** is a mobile phone with an operating system. Smartphones typically include the features of a phone with those of another popular consumer device, such as a personal digital assistant, a digital camera, a media player or a GPS navigation unit. Later smart phones include all of those plus a touch screen interface, broadband internet, web browsing, Wi-Fi, 3rd-party applications, motion sensors and mobile payment mechanisms.

13.1.6. Both Input and Output peripherals

1. Modem

A modem (**mo**dulator-**dem**odulator) is a device that modulates signals to encode digital information and demodulates signals to decode the transmitted information. The goal is to produce a signal that can be transmitted easily and decoded to reproduce the original digital data. Modems can be used with any means of

transmitting analog signals, from light emitting diodes to radio. A common type of modem is one that turns the digital data of a computer into modulated electrical signal for transmission over telephone lines and demodulated by another modem at the receiver side to recover the digital data.

2. Network interface controller

A network interface controller (network interface card, network adapter, LAN adapter) is a computer hardware component that connects a computer to a computer network.

13.1.7. Brief idea about computer architecture

- Computer architecture is the conceptual design and fundamental operational structure of a computer system. Computer architecture is defined as the functional operation of the individual hardware unit in a computer system and the flow of information among the control of those units.
- It is a blueprint and functional description of requirements and design implementations for the various parts of a computer, focusing largely on the way by which the central processing unit (CPU) performs internally and accesses addresses in memory. It may also be defined as the science and art of selecting and interconnecting hardware components to create computers that meet functional, performance and cost goals.
- **Cache Memory-** A memory that is smaller and faster than main memory and that is interposed between the CPU and main memory. The cache acts as a buffer for recently used memory location
- **Locality of reference-** Many instruction in localized area of the program are executed repeatedly during some time period and the remainder of the program is accessed relatively infrequently .this is referred as locality of reference.
- **Interrupt-** An interrupt is an event that causes the execution of one program to be

- **Memory access time-** The time that elapses between the initiation of an operation and completion of that operation ,for example ,the time between the READ and the MFC signals .This is Referred to as memory access time.
- **Memory cycle time-** The minimum time delay required between the initiations of two successive memory operations, for example, the time between two successive READ operations.
- **Static Memories-** Memories that consist of circuits capable of retaining the state as long as power is applied are known as static memories.

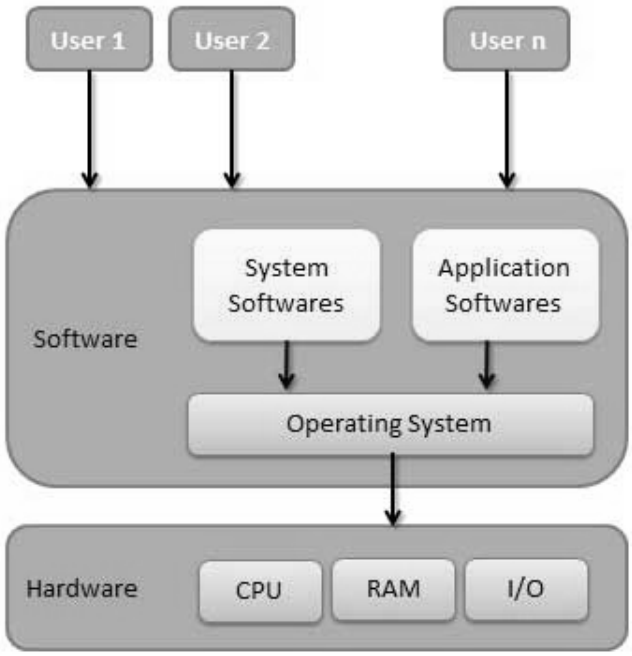
13.2. Operating system

13.2.1. Introduction

An operating system or OS is a software program that enables the computer hardware to communicate and operate with the computer software. Without a computer operating system, a computer and software programs would be useless.

An operating system is a program that acts as an interface between the user and the computer hardware and controls the execution of all kinds of programs. It is comprised of system software, or the fundamental files our computer needs to boot up and function. Every desktop computer, tablet, and Smartphone includes an operating system that provides basic functionality for the device.

Common desktop operating systems include Windows, Mac OS X, and Linux. While each OS is different, they all provide a graphical user interface, or GUI, that includes a desktop and the ability to manage files and folders. They also allow us to install and run programs written for the operating system.



An Operating System provides services to both the users and to the programs.

- It provides programs, an environment to execute.
- It provides users, services to execute the programs in a convenient manner.

Some of the examples of most commonly used computer operating systems are -

1. **Microsoft Windows 7** - PC and IBM compatible operating system. Microsoft Windows is the most commonly found and used operating system.
2. **Apple MacOS** - Apple computer operating system. The only Apple computer operating system.
3. **Ubuntu Linux** - A popular variant of Linux used with PC and IBM compatible computers.
4. **Google Android -** operating system used with Android compatible phones.
5. **iOS** - Operating system used with the Apple iPhone.

13.2.2. Operating system functions

An operating System has some of the following important functions-

- Memory Management
- Processor Management
- Device Management
- File Management
- Security
- Control over system performance
- Job accounting
- Error detecting aids
- Coordination between other software and users

13.2.3. Operating system types

With the progress and development of computers, there are various advancements in operating systems. Some of the types are mentioned below.

1. **GUI -** Graphical User Interface, a GUI Operating System contains graphics and icons and is commonly navigated by using a computer mouse.

   Examples of GUI Operating Systems are:

   a) System 7.x

   b) Windows 98

   c) Windows CE

2. **Multi-user -** A multi-user operating system allows for multiple users to use the same computer at the same time and different times. Examples of operating systems that would fall into this category are:

   a) Linux

b) Unix

c) Windows 2000

3. **Multiprocessing -** An operating system capable of supporting and utilizing more than one computer processor. Examples of operating systems that would fall into this category are:
    a) Linux
    b) Unix
    c) Windows XP
4. **Multitasking -** An operating system that is capable of allowing multiple software processes to run at the same time. Examples of operating systems that would fall into this category are:
    a) Linux
    b) Unix
    c) Windows 7
5. **Multithreading -** Operating systems that allow different parts of software program to run concurrently. Examples of operating systems that would fall into this category are:
    a) Linux
    b) Unix
    c) Windows XP

13.3. DOS and window environment

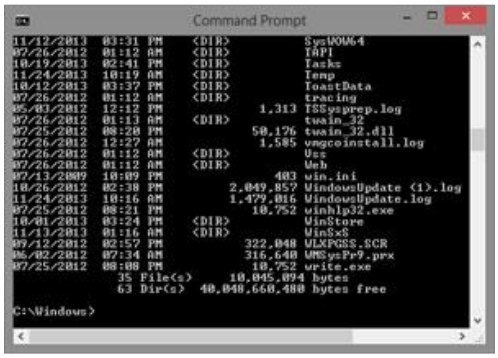
13.3.1. Introduction

Microsoft Disk operating system, MS-DOS is a non-graphical command line operating system derived from 86-DOS that was created for IBM compatible computers. MS-DOS originally written by Tim Paterson and introduced by Microsoft in August 1981 and was last updated in 1994 when MS-DOS 6.22 was released. MS-DOS allows the user to navigate, open, and otherwise manipulate files on their computer from a command line instead of a GUI like Windows.

Today, MS-DOS is no longer used; however, the command shell, more commonly known as the Windows command line is still used by many users. Most computer

users are only familiar with how to navigate Microsoft Windows using the mouse. Unlike Windows, MS-DOS is a command-line and is navigated by using MS-DOS commands. In MS-DOS, to view that same folder we would navigate to the folder using the CD Command and then list the files in that folder using the dir command.

Windows Command Prompt Window



MS-DOS is an operating system for x86-based personal computers mostly developed by Microsoft. It was the most commonly used member of the DOS family of operating systems, and was the main operating system for IBM PC compatible personal computers during the 1980s to the mid-1990s, when it was gradually superseded by operating systems offering a graphical user interface (GUI), in various generations of the Microsoft Windows operating system.

Though UNIX was a powerful operating system available, but it was not suitable for 8-bit 8086 microprocessor based Personal Computers. So there was a need for a small operating system that could work in 640K memory(RAM). DOS was an variant of CP/M (Control Program/Monitor) which ran for the first time on IBM-PC in 1981. It is called so because it resides on Floppy or Hard disk and provides command level interface between user and the computer hardware. The different versions of MS-DOS have evolved over a period of time with Microsoft introducing new features in each new releases. Starting with MS-DOS1.1, the latest version was MS-DOS6.22 released in 1994. There are various versions of DOS like MS-DOS(Microsoft), PC-DOS(IBM), Apple DOS, Dr-DOS etc.

13.3.2. Features of MS-DOS

- **CONFIG.SYS file**: This file contains reference to device drivers which are loaded when Operating System takes control of the computer. These device

drivers are required for configuring operating system for running special devices.

- **AUTOEXEC.BAT file**: This is a special batch program that is automatically executed when the system is started. It can be used to define keys, define the path that MS-DOS uses to find files, display messages on the screen etc. It will be executed only if it exists in the root directory or the diskette from which the system is loaded. Each time the system is started, MS-DOS executes the commands stored in AUTOEXEC.BAT file.
- **Disk or Drives: The** user can store data or programs on secondary storage devices called Hard disk or Floppy disk. Physically disks store data by recording any pattern of magnetic changes on using a tiny read-write head that moves over the surface.
- **Directory:** It is a special type of file that contains other files. The relation between files, directories and disk is very similar to the relation between papers, filing folders and filing cabinets..
- **File:** Information or data is stored is stored on a disk in the form of a file. When storing any file, it must be given a unique name, which can be used for subsequent identification. Filenames should not be longer than 8 characters, can have an extension which should not be longer than 3 characters. Following characters are valid in a filename: A to Z, a to z, 0 to 9,!,@#$%&(){}_-\'. Different files are identified by their extensions.
- File which have extension EXE, COM, BAT are executable files. They can be executed by just typing their name at the command prompt. Extensions TXT, DOC, BAK, BAS, C represent text file, Documentation (MS-Word) file, Backup file, basic program file, C program file respectively

DOS COMMANDS

- **DOS commands-** Any instruction given to the computer to perform a specific task is called command. The DOS has several commands, each for a particular task and these are stored in DOS directory on the disk.
- The commands are of two types :
  (a)Internal Commands: These are in built commands of MS-DOS i.e. these are stored in Command interpreter file (COMMAND.COM). These commands reside in the memory as long as the machine is at he system

prompt(C :\>) level. To use these commands no extra /external file is required. E.g. DATE, TIME, DIR, VER etc.

(b) External commands: These are separate program (.com) files that reside in DOS directory and when executed behave like commands. An external command has predefined syntax. for e.g. HELP, DOSKEY, BACKUP, RESTORE, FORMAT etc.

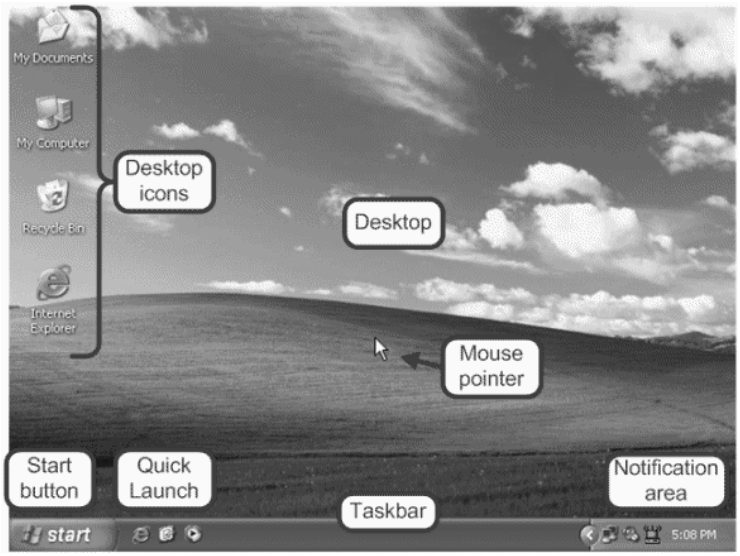
BASIC DOS COMMANDS

- DIR: To list all or specific files of any directory on a specified disk.
- MD: To make directory or subdirectory on a specified disk or drive.
- CD or CHDIR: Change DOS current working directory to specified directory on specified disk or to check for the current directory on the specified or default drive.
- RMDIR or RD: Removes a specified sub-directory only when it is empty. This command cannot remove root directory (C:\) or current working directory.
- TREE: Displays all of the directory paths found on the specified drive.
- PATH: Sets a sequential search path for the executables files, if the same are not available in the current directory.
- COPY : Copies one or more files from source disk/drive to the specified disk/drive.
- XCOPY : Copies files and directories, including lower-level directories if they exists.
- DEL : Removes specified files from specified disk/drive.
- REN : Changes or Rename the name of a file.
- ATTRIB : Sets or shows file attributes (read, write, hidden, Archive).
- BACKUP : Stores or back up one or more files/directories from source disk/drive to other destination disk/drive.
- RESTORE : Restores files that were backed up using BACKUP command.
- EDIT : Provides a full screen editor to create or edit a text file.
- FORMAT : Formats a disk/drive for data storage and use.
- TIME : sets or displays the system time.

- DATE : Sets or displays system date.
- TYPE : Displays the contents of at the specified file.
- PROMPT : Customizes the DOS command prompt.

13.3.3. Features of MS Windows

- The main features of windows are easy to use graphical user interface (GUI), device independent graphics and multitasking support.
- The first version of windows1.0 was introduced in 1985.
- Windows was an application of MS-DOS using the basic commands of DOS.
- WINDOWS-95 was released in 1995 is a 32-bit operating system which includes MS-DOS7.0 and takes control of computer system after starting.
- Windows is easier to learn and use than any of its predecessors.
- Windows and its applications run under the PCs protected mode, which mean that one ill behaved program cannot compromise the memory and resources of another.
- Windows is a pre-emptive multitasking means that programs running in the background do not significantly degrade the interactive program that we are running in the foreground



- The desktop- The center and right empty area we look at is actually the desktop. Whenever we are asked to use the desktop, the request refers to that whole area.

- Icons on the Desktop : The upper left corner contain four icons. Those icons provides access to our files and documents. Four icons are: My computer, Network Neighborhood, Recycle bin and briefcase.
- My Computer : The "My Computer" icon on the desktops opens a view into the resources of the local computer . The contents of the My computer Window depend on the disk drives on our PC and the network support that is installed.
- Network Neighbourhood - This icon displays the computers and shared printers connected on the windows network.
- Folders- Folders on the desktop can contain other folders, documents, applications and shortcuts to devices such as printer. To add a folder to the desktop, move the mouse cursor to an empty spot on the desktop and press the right mouse button. Click the Folder command. A folder icon labelled "New Folder" appears on the desktop. Label can be changed by selecting it. Drag the folder to a convenient place on the desktop.
- Documents- The reference to the current documents are stored in the documents object. The documents list includes Word processing documents, spreadsheets, database files, graphics file etc.

13.3.4. Comparison of DOS and Windows

	DOS	**Windows**
Definition	DOS (Disk Operating System) are simple text command operating systems that were popular from 1981 to 1995.	Windows is a range of graphical interface operating systems that are developed and sold by Microsoft.
GUI	DOS used a text based interface that required text and codes to operate	Windows uses graphics, images and text.
Input System	Text is used as the basic	Uses a mouse for all

	input system commands.	operating system input.
Multitasking	DOS is unable to run multiple processes at the same time.	Windows is a multitasking operating system; allowing more than one process to work simultaneously.
Storage Size	The highest amount of storage size available is 2GB.	Window systems offer storage space up to 2 terabyte.
Demands on System Resources	Booting up system is DOS is less demanding on the CPU.	Booting up Windows is more demanding on the CPU.
Registry and Swap Files	DOS uses a directory system, where all the files are contained within a particular directory or a subdirectory.	Windows uses a different registry compared to DOS, making it difficult to manually delete programs. An excessive number of temporary files and file fragments can cause the system to slow down or crash.
Current Uses	More ideally used for prototyping, testing, and making automated systems.	Used worldwide as the most popular operating system.

13.4. Applications of MS-Office

13.4.1. Microsoft Excel for Statistical Analysis

13.4.1.1. Introduction

Statistics is an area that most life sciences students find difficult. The formulae are often complicated, the calculations tedious, degrees of freedom mysterious, and

probability tables confusing. But infact students need no longer struggle with any of these. In real life, biologists and statisticians rarely use calculation and tables these days, but instead use statistical packages such as spreadsheet software such as *Excel* has most of the common statistical tests built-in. Microsoft Excel provides good capabilities for doing certain basic, frequently used, statistical analyses.

13.4.1.2. Advantages of Microsoft Excel

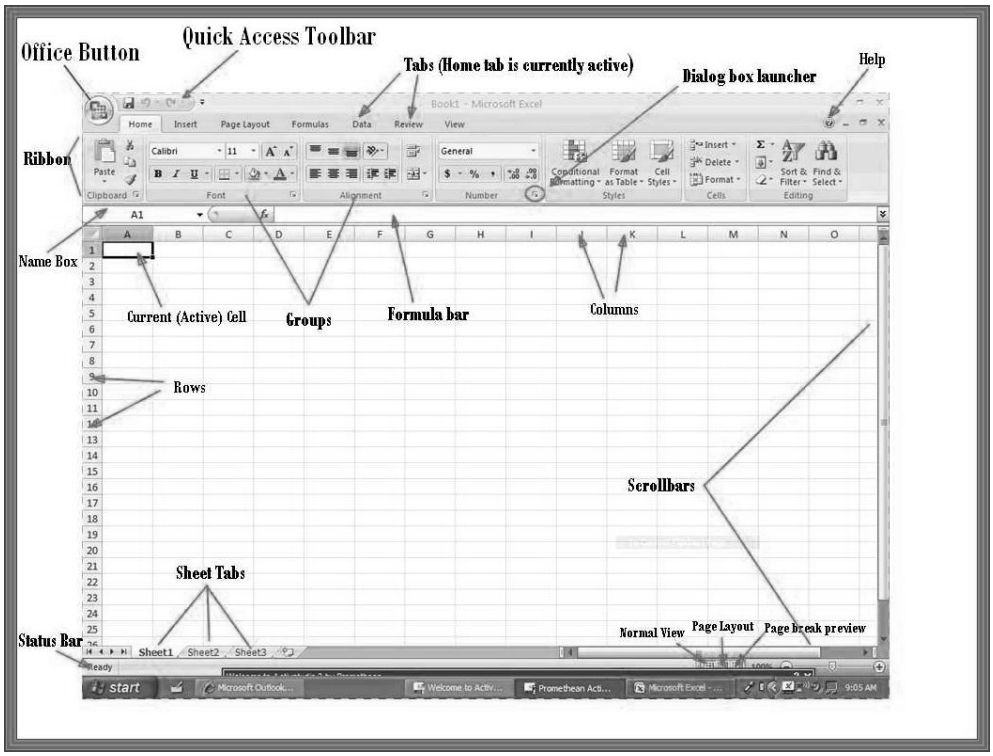
Microsoft Excel allows us to manipulate, manage and analyze data helping assist in decision making and creating efficiencies. Microsoft Excel gives you the right tools to enable you to accomplish all our needs.

The advantages of Excel are wide and varied; here are the main advantages-

- **Easy and effective comparisons -** With the powerful analytical tools included within Microsoft Excel we have the ability to analyze large amounts of data to discover trends and patterns that will influence decisions. Microsoft Excel's graphing capabilities allows us to summarize our data enhancing our ability to organize and structure our data.
- **Powerful analysis of large amounts of data -** Recent upgrades to the Excel spreadsheet enhance our ability to analyze large amounts of data. With powerful filtering, sorting and search tools we are able to quickly and

easily narrow down the criteria that will assist in our decisions. Combine these tools with the tables, Pivot Tables and Graphs we can find the information that we want quickly and easily even if we have hundreds of thousands of data items.

- **Working Together -** With the advent of the Excel Web App we can now work on spreadsheets simultaneously with other users. The ability to work together enhances our ability to streamline processes and allows for 'brainstorming' sessions with large sets of data – the collaboration tools allow us to get the most out of the sharing capabilities of Microsoft Excel.
- **Microsoft Excel Mobile & iPad Apps -** With the advent of the tablet and the smart phone it is now possible to take our worksheets to a research field area without having to bring along our Laptop. The power of these mobile devices now allows us to manipulate data and update our spreadsheets and then view the spreadsheets immediately on our phone or tablet



13.4.1.3. Formulas and Functions

Excel has facility to implement user defined formulas and functions. We may

include built in functions. Both formulas and functions are used for advance computing.

Functions: Functions can be used to perform simple and complex functions. Specific values used to perform functions are known as arguments. Parenthesis used to separate different parts of a formula. We can type desired function in function wizard, formula bar or directly in cell.

Following command sequence may be followed for functions-

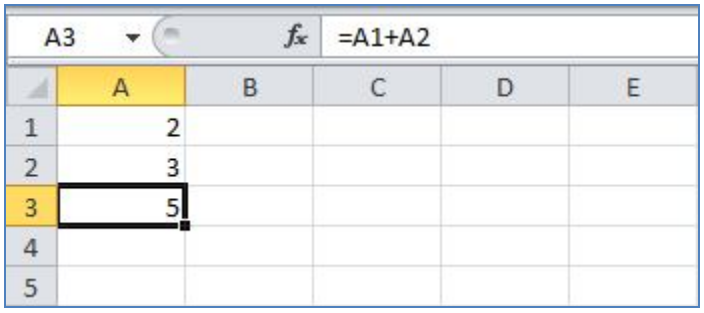
1. Click on the cell in which function is to be used.
2. Choose function from insert menu or click paste function button on the standard tool bar to display the paste function dialogue box.
3. Click on the function category, which has a list of various functions. Different function categories are given below in the table.
4. After this, we will find syntax windows to create a function. Click on the collapse button (marked with red arrow) to the right of the box marked number 1 or value1 (depend on the selected function).
5. Drag the mouse to choose cell range.
6. To insert additional arguments into the function.
7. Drag the mouse to choose cell range.
8. Click ok.

Table: Functions and Categories Used in Excel

Categories	**Functions**
Statistical	Max, min, average etc.
MATHEMATIAL	Sum(num1,num2...); round; sqrt; abs; truc etc
Logical functions	And, not, or
Text functions	Concatenate(text1,text2...), len(text), lower(text), proper(text), upper(text), trim(text)
Date and time	Date(year, month, day),time(hour, minute, second), now()

Using auto sum: Auto sum function is used most frequently. Following command sequence may be followed for auto sum:-

1. Select the desired cell.
2. Click on auto sum button on standard tool bar.
3. Cells get converted into dotted line. It is known as marquee.
4. Press entre key (if range is correct)
5. If not, type or select the correct range.



Formulas: Formulas are mathematical expressions in excel used to perform calculations. It always begins with equal to (=) sign. If formula is wrong, it will display error in formula massage. Following command sequence may be followed for auto sum:-

1. Choose the desired cell for formula.
2. Type equal to sign followed by operation.
3. Type cell names.
4. Press entre key.

Few important Statistical functions-

Function	**Description**
AVERAGE	Returns the average of its arguments
BINOMDIST	Returns the individual term binomial distribution probability

CHIDIST	Returns the one-tailed probability of the chi-squared distribution
CHITEST	Returns the test for independence
CORREL	Returns the correlation coefficient between two data sets
EXPONDIST	Returns the exponential distribution
FORECAST	Returns a value along a linear trend
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an F-test
GAMMADIST	Returns the gamma distribution
GAMMALN	Returns the natural logarithm of the gamma function, $\Gamma(x)$
GEOMEAN	Returns the geometric mean
GROWTH	Returns values along an exponential trend
HARMEAN	Returns the harmonic mean
INTERCEPT	Returns the intercept of the linear regression line
MEDIAN	Returns the median of the given numbers
NORMDIST	Returns the normal cumulative distribution
NORMINV	Returns the inverse of the normal cumulative distribution
NORMSDIST	Returns the standard normal cumulative distribution
POISSON	Returns the Poisson distribution
PROB	Returns the probability that values in a range are between two

	limits
RSQ	Returns the square of the Pearson product moment correlation coefficient
STDEV	Estimates standard deviation based on a sample
STDEVP	Calculates standard deviation based on the entire population
TDIST	Returns the Student's t-distribution
TRIMMEAN	Returns the mean of the interior of a data set

13.4.1.4. Some practical applications of Excel in Biostatistics-

a) Using Excel for t-Tests of Hypotheses

The t-Test for Independent Samples

To use Excel to carry out a t-test for *independent* samples, enter the sample observations into a worksheet. One way to do this is to use two rows for the data, i.e., one row for each set of sample observations. It is handy to have each row begin with a label in the left-most cell. Choose Tools ⟶ Data Analysis ⟶ t-Test: Two-Sample Assuming Equal Variances. In the resulting popup, we will need to specify:

- The ranges for each of the variables, i.e., samples; if the ranges include the cells containing the labels and if we check the box next to labels in the popup, our results will include those labels.
- The hypothesized mean difference. This will be 0 if we are testing the usual null hypothesis that the population means are equal.
- The level of significance, alpha, that we want to use. The default choice is 0.05.
- The output range (or separate worksheet or workbook)

The results will include the observed value of the t-statistic, which – for the purposes of LIS 397.1 – is the one labeled "t Critical two-tail". The results will also include the probability (i.e., the "p-value") of our observing the t value that we did observe, when the null hypothesis is true; this is labeled "P(T<=t) two-tail".

The t-Test for Dependent (and Matched-Pair) Samples

To use Excel to carry out a t-test for *dependent* samples, carry out essentially the same steps as in the independent-sample procedure, except that we will choose Tools ⟶ Data Analysis ⟶ t-Test: Paired Two Sample for Means. Excel assumes that cells in *corresponding* positions in the two rows (or columns) contain the *paired* observations.

b) Using Excel for ANOVA

Enter the data as outlined above for the t-test procedures, using one row (or column) for each set of sample observations (e.g., for a 3-population ANOVA test, we will take a sample of observations from each of the 3 populations, and we will enter each different sample in a different row [or column]). It is easiest to use adjacent rows (or columns) for the data and to begin every row in the same column (or to begin every column in the same row). Choose Tools ⟶ Data Analysis ⟶ Anova: Single Factor. In the Anova: Single Factor popup, we will need to specify:

- the range containing the sample sets of observations, i.e., the entire set of cells containing the whole set of observations
- whether the observations are grouped (divided into sets corresponding to the different samples) by columns or by rows
- whether the range includes cells with labels
- the upper left-hand corner of the range in which we want the output to appear (or we can choose a different sheet in the workbook, or a different workbook)

Excel provides a version of the standard ANOVA table, including the value of the observed F-ratio and the corresponding P-value (i.e., the probability of our observing the value of the F-ratio that we did observe, when the null hypothesis is true). The output also includes the threshold value, labeled "F crit", against which we can compare the observed value of the F-ratio; this is the value that we would find in a table of the F-ratio, such as those contained in many texts on inferential statistics and in this table from the U.S. National Institute of Standards and Technology (NIST).

c) Using Excel for Correlation

Excel does Pearson and Spearman correlation, as well as linear regression. For regression Excel insists that the observations be placed in columns rather than rows, i.e., it insists that the independent and dependent variables have their respective observed values entered in columns. Since often we will want to do both regression and correlation, we may as well develop the habit of entering the pairs of values in successive rows in columns (typically, in adjacent columns, although adjacency is not required).

To do correlation, choose Tools ⟶ Data Analysis ⟶ Correlation. In the Correlation popup, we will have to specify:

- The range containing the observed values; if the range includes the cells containing the labels and if we check the box next to Labels in the popup, our results will include those labels.
- Whether the grouping is by columns or rows; if, as recommended, we have entered the values of each variable into a column, we need to select the radio button for columns.
- The upper left-hand corner of the range in which we want the output to appear (or we can choose a different sheet in the workbook, or a different workbook)

Excel displays the results in a 2x2 table (for the 2-variable case), showing the correlation of each variable with itself (viz., 1) and with the other variable. Only the cells along the diagonal and in the lower half of the table are filled in, since the table is symmetric with respect to the diagonal. The sample Pearson correlation coefficient thus appears in the cell in the lower left corner.

d) Using Excel for Linear Regression

As was noted earlier, in doing regression Excel insists that the observations be placed in columns rather than rows, i.e., it insists that the independent and dependent variables have their respective observed values entered in columns.

To do regression, choose Tools ⟶ Data Analysis ⟶ Regression. In the Regression popup, we will have to specify:

- The range containing the observed values of the dependent (called "Input Y") variable and of the independent (called "Input X") variable. If these

ranges include the cells containing the labels and if we check the box next to Labels in the popup, our results will include those labels.

- The confidence level if we want to choose a level other than 95% (which Excel provides we by default).
- Whether we want to force the regression line to pass through the origin (i.e., the point whose X and Y coordinates are both zero. If we do want to force the regression to pass through the origin, check the box next to "Constant is Zero"; otherwise leave this box unchecked).
- The upper left-hand corner of the range in which we want the output to appear (or we can choose a different sheet in the workbook, or a different workbook).
- Whether we want the output to include displays of other related data, e.g., residuals, which we are ignoring in LIS 397.1. One of the options is a line-fit plot; this plot is not very satisfactory, and this option is probably best left unused for the purposes of LIS 397.1.

Excel displays the results in several tables. The values we are primarily interested in for LIS 397.1 are shown in the column headed "coefficients". Excel uses for this coefficient the label we provided for the independent variable; if we did not choose the Labels option, Excel calls this "X Variable 1." Excel reports the value of the Pearson correlation coefficient as "Multiple R" in the table called "Regression Statistics."

e) Using Excel for Chi-Square Tests

The Chi-Square Goodness-of-Fit Test and the Chi-Square Test of Association

Excel does not provide a particularly convenient means for doing the chi-square goodness-of-fit and association tests. We will have to provide the observed values in one range, and the corresponding expected values in another range, in an Excel spreadsheet. We can, of course, use ordinary spreadsheet functions to calculate the expected values, but it may be quicker and easier to use a calculator to produce the expected-value numbers and then copy the numbers into Excel.

When we have provided a range of observed values and another range of corresponding expected values, we choose another cell and place in it the formula

=CHIINV(CHITEST(range1, range2), df)

where range1 contains the observed values, range2 contains the expected values, and df is the pertinent number of degrees of freedom. The chosen cell will display the observed value of chi-square. If we are working the chi-square goodness-of-fit test, it will be convenient to have each range be a span of cells within a column. If we are working the chi-square association test, it will be convenient to have the ranges take the form of rectangles of cells.

Note that by itself the formula

=CHITEST(range1, range2)

Yields the probability, in the circumstance that the null hypothesis is true, of observing the value of chi-square that was observed.

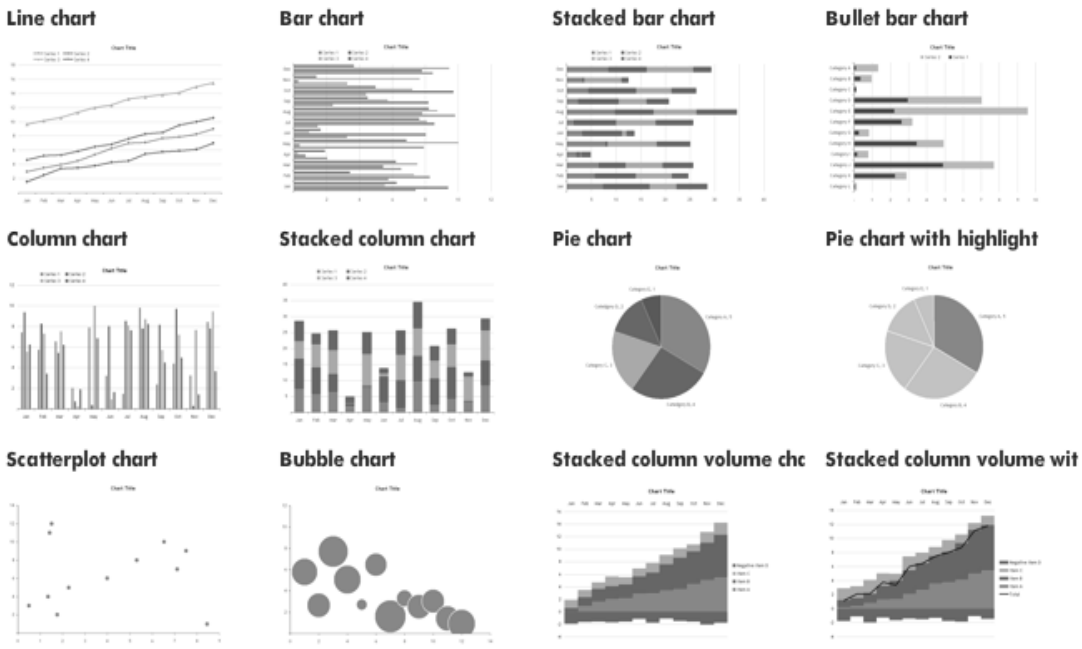
13.4.1.5. Charts

Charts give more visual clarity and meaning to data. It is easier to understand things explained by charts. There are many types of charts available in MS-Excel. Charts are linked to the values given in the worksheet. Charts may be placed in two ways in excel 1) embedded charts 2) chart sheets. Charts are explained in the following table.

Table: Types of Chart in Excel

CHART TYPE	DESCRIPTION
AREA CHART	Used to put emphasis on change over time.
3D SURFACE	Implicates 3d changes.
BAR CHART	Compare values
RADAR	Each category of information radiate form centre.
COLUMN	Looks like bar chart. Bars are aligned vertically.
BUBBLE	Displays three sets of variables. Represents two axes and bubble size.
LINE	To compare trend

SCATTER	To compare set of values with the average or predicted values.
PIE	Compare set of figures
DOUGHNUT	Display more than one figure. Looks like PIE CHART.



Creating charts by Excel

Charts can be created by using chart wizard in excel. We may also adopt following sequence to create charts.

1. Entre data in the excel sheet.
2. Select chart from the insert menu or chose chart wizard available in the standard tool bar.
3. Select chart type according to the values
4. Choose data range.
5. Click on next
6. Entre name of the chart and other related values e.g. x and axis, grid lines,

legends, data table etc.

7. Click next
8. Select location of the chart on the same sheet or new sheet.
9. Press finish button

Resizing and Moving Charts: charts can be resized or moved to other places by using mouse on the border of the chart.

Use of Chart Tool Bar: We can click right button on the menu and select chart. Buttons available on the chart tool bar are explained below:-

Table: Chart Tool Bar Buttons

Command	**Meaning**
Chart object	To select different objects n the chart
Format chart area	To edit the chart area
Chart type	To select different types of charts
Legend	To display of hide chart legend
Data table	To display data table instead of chart
By row	To show data by rows
By column	To show data by column
Angle text downward	To angle text in downward direction
Angle text upward	To angle text in upward direction

Saving a Chart: To save a chart, we may use following command sequence.

1. Select chart and chose chart type.
2. N custom type, select user defined- click add we will find the chart type dialogue box.

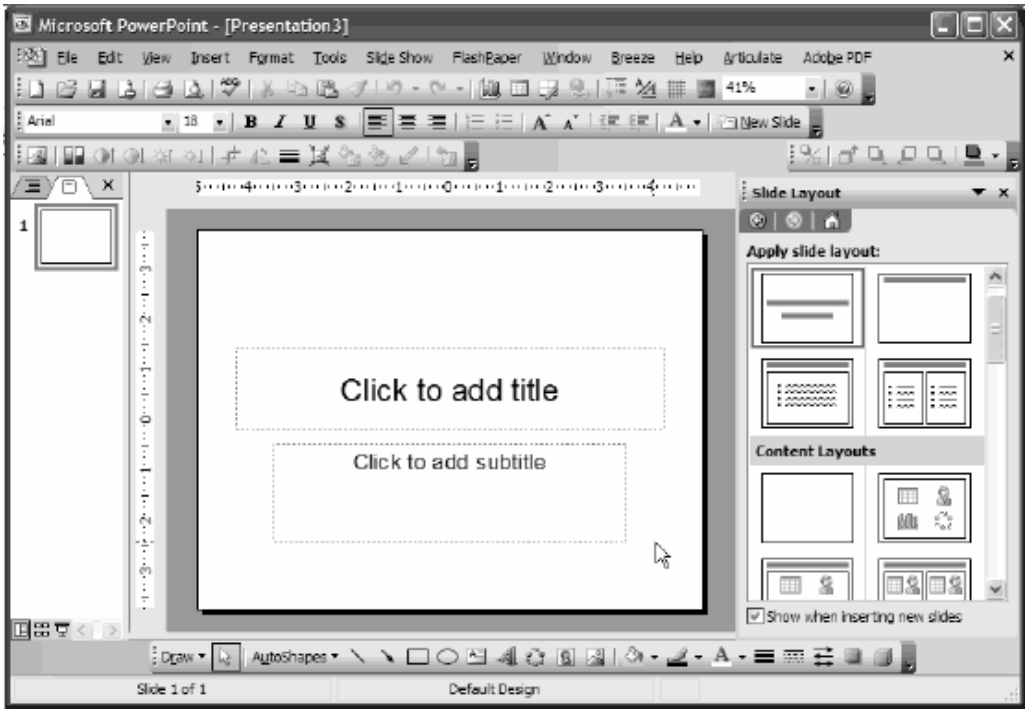
3. Type name of the chart à fill up description.
4. Set as default chart à click yes
5. Click ok to save the chart

13.4.2. Brief idea about MS-power point

PowerPoint is the most popular presentation software of MS-Office. It is regarded by many as the most useful and accessible way to create and present visual aids to the audience.

Uses of MS-power point-

- Quick and Easy: the basic features are easy to master and can make us appear to be organized, even if we are not.
- Easy to create a colorful, attractive design: using the standard templates and themes, even if we do not have much knowledge of basic graphic design principles .
- Easy to modify: when compared to other visual aids such as charts, posters, or objects, it is easy to modify.
- Easily re-order presentation: with a simple drag and drop or using key strokes, we can move slides to re-order the presentation.
- Audience Size: PowerPoint slides are generally easier to see by a large audience when projected than other visual aids.
- Easy to present: we can easily advance the slides in the presentation one after another with a simple key stroke while still maintaining eye contact with the audience.
- No need for Handouts: they look good visually and can be easily read if we have a projector and screen that is large enough for the entire room.



13.5. Summary

Computers used in scientific research have the ability to analyze data in ways and at speeds not possible with the human eye. Computers play a major role today in every field of scientific research from genetic engineering to astrophysics research. This section is a brief overview of computer peripherals and architecture, computer architecture, elementary idea about operating system, DOS and window environment, DOS Commands, Comparison of DOS and Windows and Some practical applications of Excel in Biostatistics

13.6. Self-Learning Exercise

1. What do you mean by operating system? Explain types of operating system.
2. What do you mean by input devices? Explain the functioning.
3. Give brief idea about computer architecture.
4. Define the output units in brief?
5. What are storage peripherals of computer?
6. Write an essay on MS-DOS.
7. What are the various commands of MS-DOS?
8. What are the types of commands of MS-DOS?

9. Give a comparison of DOS and Windows.
10. What are the various applications of MS-Office in relation to statistics?
11. What are the formulas and functions in MS-Excel?
12. What are the advantages of MS-Excel?
13. Write a short note on Charts by MS-Excel.
14. Give a brief idea about MS-power point.
15. Give some practical applications of Excel in Biostatistics.
16. What are the different features of MS Windows?

13.7. Reference Books

- Brian.K.Williams & Stacey C.Sawyer (2005) 'Using Information Technology' ; Tata McGraw Hill Publication company limited, Delhi.
- Henry.C.Lucas,Jr. (2001) ' Information Technology'; Tata McGraw Hill Publication company Limited, New Delhi.
- D.P.Sharma (2008); 'Information Technology' ; College Book Centre , 2008, Jaipur.
- P.K.Sinha, Priti Sinha,(2007) 'Computer Fundamentals' ; BPB Publication, New Delhi.

Unit-14

Software used in biomedical science (image analysis system automation). sound spectrum analysis, computer simulation, digital alternatives of invasive techniques in anatomy and physiology

Structure of the Unit

14.0 Objectives

After going through this unit you will be able to understand the various software used in the field of biomedical sciences and importance of such software in biological field like biomedical laboratories, research laboratories etc.

14.1 Introduction

In clinical diagnosis many Software and digital techniques are developed, these technique and process create visual representations of the interior of a body for clinical analysis and medical intervention. Medical imaging seeks to reveal internal structures hidden by the skin and bones, as well as to diagnose and treat diseases. These techniques also establish a database of normal anatomy and physiology to make it possible to identify abnormalities.

14.2 Software used in biomedical science

Many types of software are being used in the field of biomedical sciences which made the analysis and quantification of data easy and time saving. A few of them are discussed below:

14.2.1 Image analysis system automation

These types of software are developed to study the various techniques applied for analyzing and quantifying biological images.

Image analysis plays an important role in the scientific field due to its wide range of applications in quantitative measurements. Visualization and image analysis methods are critical for understanding various features of cell biology, molecular

biology and neuroscience. With the development of fluorescent probes and the application of high-resolution microscopes biological image processing techniques became more reliable with a profound impact on research in the biological sciences.

One of the most important and frequently overlooked aspects of the cell imaging is image quantification and analysis. In the past, microscopic techniques were applied to study the structural details, but the recent advances in research, demands on determination of the number of cells, its area, perimeter, localization, concentration, densitometry analysis, etc.,. for molecular level studies. Biologists are increasingly interested in using the image analysis protocols to convert the microscopic images into more relatively quantitative measurements.

The difficulties in visual interpretation such as counting of the cell and quantification of specific molecules of interest in the research application, can be overcome by implementing automated methods in these fields. Computerized image analysis has lot of applications over visual analysis, including reproducibility, rapidity, adaptability and the ability to simultaneously measure many features in the image. The goal of image analysis techniques is to combine the results of the wet laboratory techniques with image analysis software, thereby providing more quantitative information.

A large number of image analysis software packages have been developed for biological applications due to their usability in biological sciences. These software packages help to extract useful information from the specimens (image) of interest. In fact, most of this software is expensive and often requires high performance computers to function. Throughout this lab, ImageJ is referred as standard image analysis software, since it is freely available, platform independent and is applicable to the biological researchers to quantify the results obtained in the laboratory techniques.

Applications of Image Analysis Software:

Image analysis software has multiple applications, including the analysis of microscopy, gel, fluorescent stained tissues, and in the medical analysis of specimens obtained from the patient's sample. Usually, image processing techniques are applied to correct problems such as uneven illumination and to enhance images for further analysis, display, and publication.

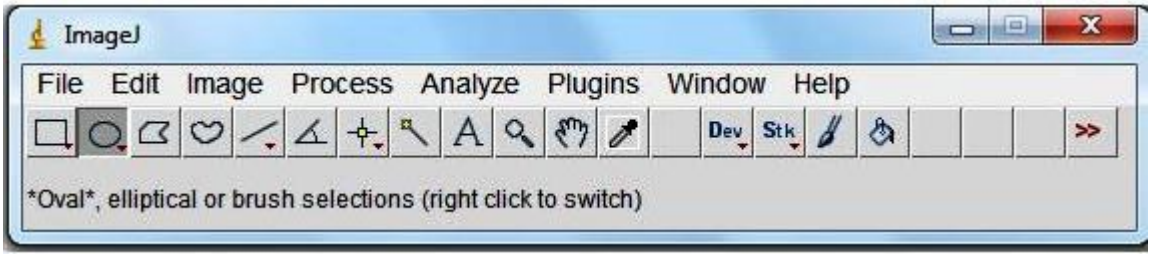
Densitometry, that is, the determination of intensity of apparent amounts of a specific molecule at a certain position inside the sample, can be analyzed with the help of the image analysis software. Image analysis software is used compare the bands detected on the gel, for example, in PAGE, AGE, and Western blot, and also to detect the spot developed on the TLC plate. Here, the image analysis techniques are applied to quantify the endogenous expression of target protein (in case of Western blot and PAGE), presence of DNA in specific regions of the gel, depending on its molecular size (in case of AGE) and to quantify the amount of amino acids present in an unknown sample (in case of TLC). In Western blot, the densitometry analysis can also be applied to find the expression of various proteins, for example, cancer causing proteins, by determining the relative intensity of proteins in the patient's sample.

In the medical field, image analysis can also be applied to measure or count the nuclear or cytoplasmic stain which can be applied for prognosis, diagnosis and evaluation of response to antibiotic therapies. It can measure differences in staining intensity of a tissue sample.

Quantification of fluorescent images is widely used in molecular biology and biochemistry laboratories for a wide variety of experimental, analytical and quality control applications. In fluorescent microscopes, the image analysis techniques can be used to determine the location of fluorophores in tissue sections that can be applied for the future purposes. The quantification of fluorescence imaging techniques has potential applications in both biomedical research and clinical practices. Moreover, quantitative image analysis of DNA staining with respect to its intensity reveals whether a cell is in G1 or G2 phase of the cell cycle. ImageJ help in the Cell count (used to probe cell proliferation/apoptosis/death) that can be difficult to obtain by manual cell counting processes.

Image analysis can also help to identify the phenotypes, such as shape and texture of the cell of interest, that are not otherwise easily measured. Quantitative measurements of the cell morphology are important in studying the normal cellular physiology and in disease diagnosis. Over all, the accuracy of the information obtained through the analysis software depends not only on the quality of the image, but also on the ability of the analysis software to distinguish the image features from the provided data and to convert it into meaningful measurements.

ImageJ is a free public domain image analyzing software used for image processing and analysis techniques. The ImageJ window is shown below.



	Rectangular selection tool
	Oval, elliptical or brush selection tool
	Polygon selection tool
	Freehand selection tool
	Free hand line selection tool
	Angle tool
	Point selection tool
	Wand (Tracing) tool
A	Text tool
	Magnifying tool
	Scrolling tool

ImageJ Toolbars

14.2.2 Sound spectrum analysis

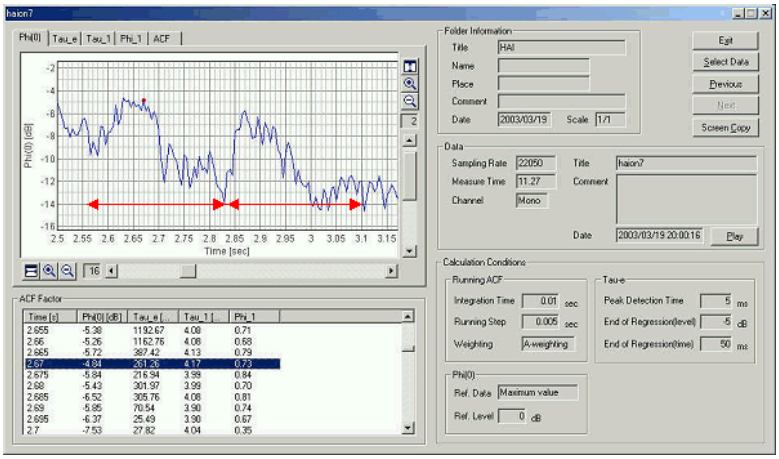
In clinical diagnosis, auscultation is very effective in early detection of heart disease or a lung disease, the importance of auscultation sound is much recognized. In the conventional method, it is required to transmit a good quality auscultation sound to perform a diagnosis. Since the high sampling rate of 44kHz is required for that purpose, highly efficient communication equipment is required for both sides (doctor and patient). To overcome such a problem, new technology proposed for transmitting auscultation sound efficiently. In this the biomedical sound analysis system developed that can analyze various sound using the advanced sound measurement and analysis technologies of temporal and spatial properties of sound.

When a patient is examined, physical assessment is very important. Basic physical assessment techniques are Interview, Inspection, Palpation, Percussion, and Auscultation. Inspection means to examine by looking visually. Palpation is examination by touch. The texture, consistency and tenderness of tissue are determined by palpation. Percussion is the technique of tapping the surface of a body to make a determination by sound. Auscultation means examining by listening to the sounds coming from organs. The sound coming from internal organs gives us much information very effective in diagnosis. The purpose of this research is to carry out the maximum use of the information of sound to contribute to the remote medical treatment.

14.2.2.1 Auscultation of neck

Much information is contained in the sound of a neck. Many diseases can be diagnosed by the auscultation of the neck.

1. It is said that the air passing a tracheal gill would make unusual wind sound if a malignant neoplasm is made in a bronchus. So the bronchus neoplasm is found by auscultation of the Adam's apple. This sound is called rhonchi. Rhonchi are low pitched, snore-like sounds. They are caused by airway secretions and airway narrowing.



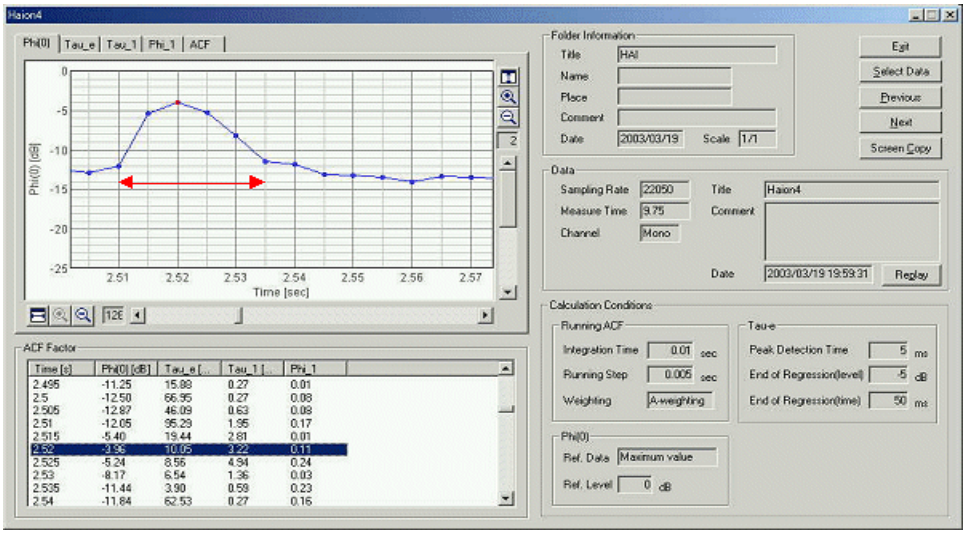
Rhonchi sound: measurement of abnormal lung sound

14.2.2.2 Auscultation of chest

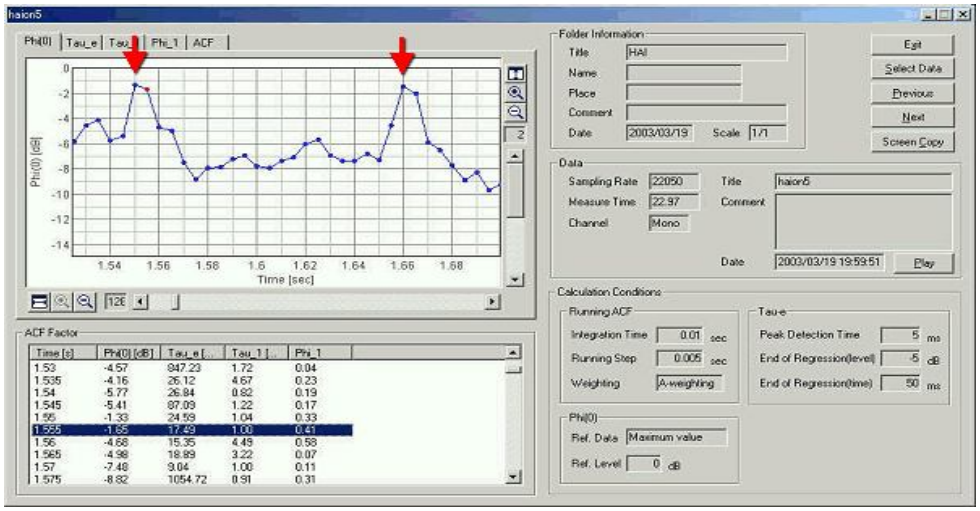
Auscultation of chest is very common. The organs which make sound are lungs and the heart. Other disease may be diagnosed. The sound of lungs and the heart is very complicated. So the qualities of diagnosis are very different between an experienced doctor and an inexperienced doctor.

1. **Pneumonia**

   This is the state in which bacteria and the virus were infected and bred in lungs. From the state of respiratory sound, it can be diagnosed about the grade and cause bacillus of pneumonia. The feature of sound is coarse crackle, which was analyzed in Lung sound measurement 2. More precisely, it is heard by bronchitis, pneumonia, pulmonary tuberculosis, pulmonary infarction, pulmonary suppuration, a lung congestion, the lung blister, bronchiectasis. Crackle sound is described as fine, popping, crackling, discontinuous, non-musical noises. It lasts about 10-25 ms.



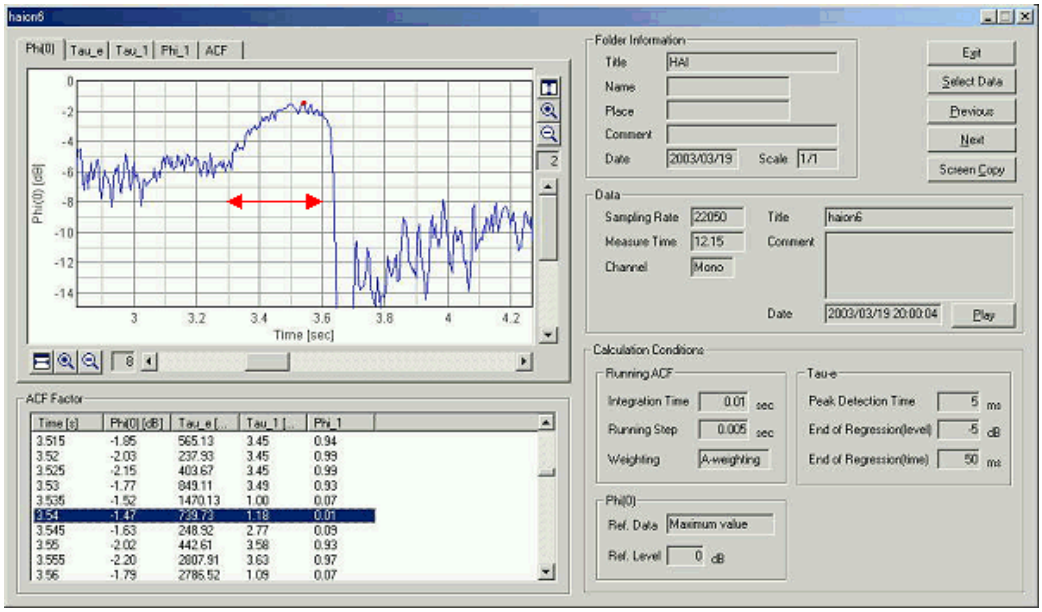
Coarse crackle: lung sound measurement



Fine crackle Lung sound measurement

2. **Bronchial asthma**

A characteristic sound can be heard. So it can be found easily. Auscultation may be the most effective diagnostic method. The feature of sound is wheezing. Wheezes are caused by narrowing, constriction, or spasm in the very small airways. They can occur because of asthma, congestive heart failure, fibrosis, pneumonia, and tuberculosis, also. The lasting time of the wheezing sound is about 250 ms.



Wheezing sound Lung sound measurement

3. **Cardiac insufficiency**

   This is in the state where the function of the heart is weakened and the flow of the blood of the whole body is overdue. Blood accumulates in lungs and the skin. It can be diagnosed by the medical examination of whole body including auscultation.

4. **Valvular heart disease**

   The heart has the function which continues sending blood to the one direction. The structure for adverse current prevention called a valve is in each part of heart so that blood may always flow in the fixed direction. If this valve breaks or it becomes hard, it will become that blood cannot flow or flow backwards, finally it will become heart failure. There are four valves in the heart. It can be diagnosed by auscultation which one is abnormal, whether it is broken or harden.

14.2.2.3 Auscultation of abdomen

The belly is a very busy place when a stethoscope is applied. From the state of the sound, the state of abdomen internal organs including the stomach and intestines is guessable. Even if not disease, the function of the internal organs of the belly can be grasped. For example, it can be seen that this man is constipation or this man has arteriosclerosis, etc.

1. **Ileus**

   A motion of intestines is understood very well by auscultation. Ileus has big influence on a motion of intestines. In diagnosis of ileus, auscultation is the most important method.

2. **Enteritis**

   Enteritis is also found by the auscultation.

3. The abnormalities of other abdomen blood vessels

   By auscultation, an unusual sound called blood vessel noise can be heard.

14.2.2.4 Construction of sound database using Internet

A skill of auscultation depends on practice and experience. It is necessary to hear sound and memorize the feature of the characteristic sound related to some disease. By using the technology, the characteristic sound components contained in auscultation sound can be caught correctly. It can support diagnosis exactly. Furthermore, by applying the technique established by the speech recognition, it becomes possible to discover abnormalities by comparing the measured data with the database of auscultation sound. The database is regarded as a kind of medical sound dictionary. Finally, it is possible to construct a database of all the information about auscultation sound. It is also possible by constructing this database on a network to contribute to support the research and the remote medical examination.

14.2.3 Computer simulation

A computer simulation is a simulation, run on a single computer, or a network of computers, to reproduce behaviour of a system. The simulation uses a computer model, or a computational model to demonstrate the system. Computer simulations have become a useful part of biological system, human systems in psychology, technology and engineering. Simulation of a system is represented as the running of the system's model. It can be used to explore and gain new insights into new technology and to estimate the performance of systems too complex for analytical solutions.

Computer simulations vary from computer programs that run a few minutes to network-based groups of computers running for hours to ongoing simulations that run for days. The scale of events being simulated by computer simulations has far exceeded anything possible (or perhaps even imaginable) using traditional paper-

and-pencil mathematical modeling. Simulation has been used since long time. Some examples include a 1-billion-atom model of material deformation; a 2.64-million-atom model of the complex maker of protein in all organisms, a ribosome, in 2005; a complete simulation of the life cycle of *Mycoplasma genitalium* in 2012; and the blue brain project at EPEL(Switzerland), begun in May 2005 to create the first computer simulation of the entire human brain, right down to the molecular level.

14.2.3.1 Data preparation

The external data requirements of simulations and models vary widely. For some, the input might be just a few numbers (for example, simulation of a waveform of AC electricity on a wire), while others might require terabytes of information (such as weather and climate models).

It required input sources:

- Sensors and other physical devices connected to the model;
- Control surfaces used to direct the progress of the simulation in some way;
- Current or historical data entered by hand;
- Values extracted as a by-product from other processes;
- data can be entered into the simulation when it starts up, for example by reading one or more files;
- data can be provided during the simulation run, for example by a sensor network.

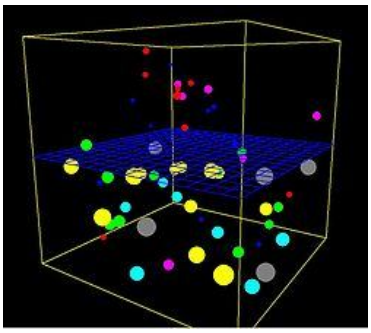
14.2.3.2 Visualization

The output data from a computer simulation was presented in a table or a matrix showing how data were affected by numerous changes in the simulation parameters. The use of the matrix format was related to traditional use of the matrix concept in mathematical models. However, psychologists and others noted that humans could quickly perceive trends by looking at graphs or even moving-images or motion-pictures generated from the data, as displayed by computer generated imagery (CGI) animation. Such intense graphical displays, which transcended the world of numbers and formulae, sometimes also led to output that lacked a coordinate grid or omitted timestamps, as if straying too far from numeric data displays. Similarly, CGI computer simulations of CAT scans can simulate

how a tumor might shrink or change during an extended period of medical treatment, presenting the passage of time as a spinning view of the visible human head, as the tumor changes.

14.2.3.3 Computer simulation in science

Computer simulation in the field of science has great applications like the process of osmosis.


Process of osmosis

Generic examples of types of computer simulations in science, which are derived from an underlying mathematical description:

- A stochastic simulation, typically used for discrete systems where events occur probabilistically and which cannot be described directly with differential equations. Phenomena in this category include genetic drift, biochemical or gene regulatory networks with small numbers of molecules.

Specific examples of computer simulations follow:

- Agent based simulation has been used effectively in ecology, where it is often called "individual based modeling" and is used in situations for which individual variability in the agents cannot be neglected, such as population dynamics of salmon and trout.
- Computer simulations have also been used to formally model theories of human cognition and performance, e.g., ACT-R
- Computer simulation using molecular modelling for drug discovery.

14.2.4 Digital alternatives of invasive techniques in anatomy and physiology

As a field of scientific investigation, digital alternative of invasive technique constitutes a sub-discipline of biomedical engineering, medical physics or medicine depending on the context: Research and development in the area of

instrumentation, image acquisition (e.g. radiography), modelling and quantification are usually the preserve of biomedical engineering, medical physics, and computer science; Research into the application and interpretation of such techniques is usually the preserve of radiology and the medical sub-discipline relevant to medical condition or area of medical science (neuroscience, cardiology, psychiatry, psychology, etc.) under investigation.

Many of the techniques developed as an alternative to prevent invasive technique for anatomy and physiology also have scientific and industrial applications. Some of them are discussed below:

14.2.4.1 Radiography

Two forms of radiographic images are in use in medical imaging; projection radiography and fluoroscopy, with the latter being useful for catheter guidance. These 2D techniques are still in wide use despite the advance of 3D tomography due to the low cost, high resolution, and depending on application, lower radiation dosages. This imaging modality utilizes a wide beam of X-ray for image acquisition and is the first imaging technique available in modern medicine.

Fluoroscopy produces real-time images of internal structures of the body in a similar fashion to radiography, but employs a constant input of x-rays, at a lower dose rate.

Projectional radiographs more commonly known as x-rays, are often used to determine the type and extent of a fracture as well as for detecting pathological changes in the lungs.

14.2.4.2 Magnetic Resonance Imaging (MRI)

A magnetic resonance imaging instrument (MRI scanner), or "nuclear magnetic resonance (NMR) imaging" scanner as it was originally known, uses powerful magnets to polarise and excite hydrogen nuclei (single proton) in water molecules in human tissue, producing a detectable signal which is spatially encoded, resulting in images of the body. The MRI machine emits a RF (radio frequency) pulse that specifically binds to hydrogen. The system sends the pulse to the area of the body to be examined. The pulse makes the protons in that area absorb the energy needed to make them spin in a different direction. This is the "resonance" part of MRI. The RF pulse makes them (only the one or two extra unmatched protons per million) spin at a specific frequency, in a specific direction. The particular frequency of

resonance is called the Larmour frequency and is calculated based on the particular tissue being imaged and the strength of the main magnetic field.

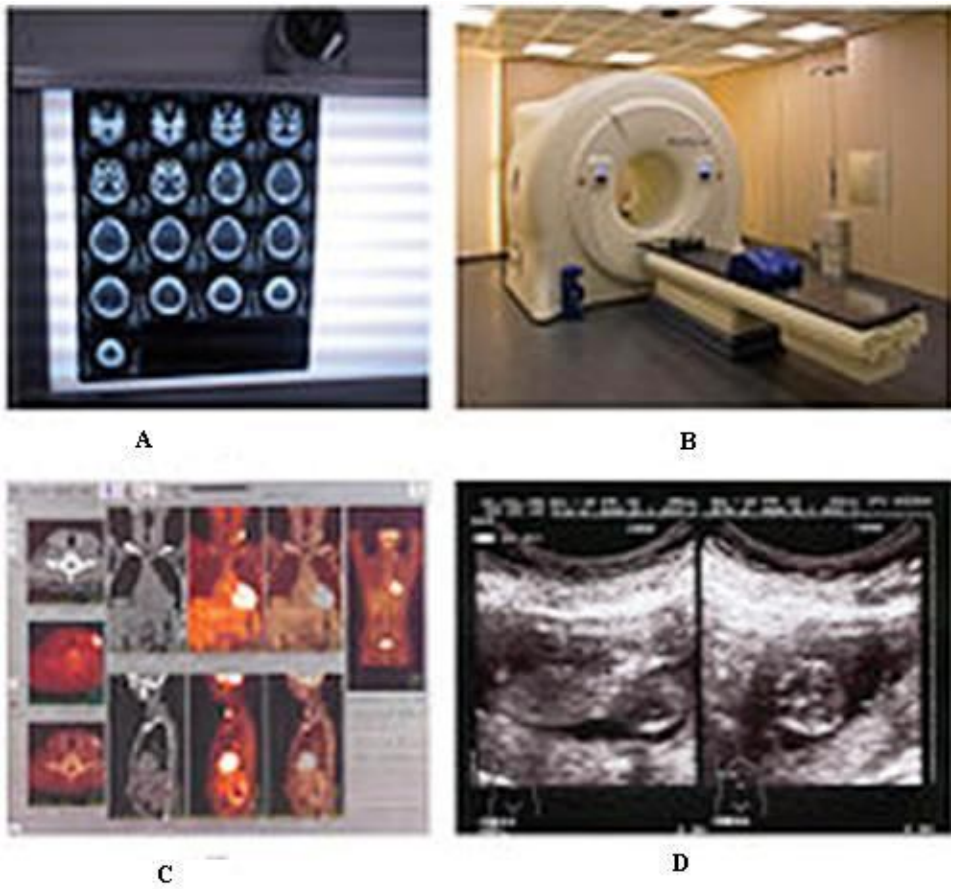


Figure (A) shows the results of a CT scan of the head are shown as successive transverse sections.

Figure (B) An MRI machine generates a magnetic field around a patient.

Figure (C) PET scans use radiopharmaceuticals to create images of active blood flow and physiologic activity of the organ or organs being targeted.

Figure (D) Ultrasound technology is used to monitor pregnancies because it is the least invasive of imaging techniques and uses no electromagnetic radiation.

Because CT and MRI are sensitive to different tissue properties, the appearance of the images obtained with the two techniques differ markedly. In CT, X-rays must be blocked by some form of dense tissue to create an image, so the image quality when looking at soft tissues will be poor. In MRI, while any nucleus with a net nuclear spin can be used, the proton of the hydrogen atom remains the most widely used, especially in the clinical setting, because it is so ubiquitous and returns a

large signal. This nucleus, present in water molecules, allows the excellent soft-tissue contrast achievable with MRI.

(A) A brain MRI representation

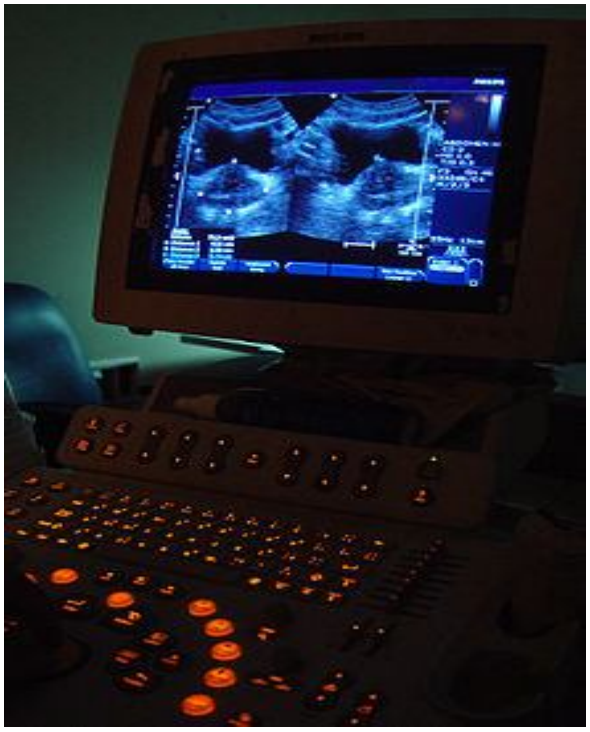
(B) X-ray picture

14.2.4.3 Ultrasound

Ultrasonography uses high frequency broadband sound waves in the megahertz range that are reflected by tissue to varying degrees to produce (up to 3D) images. This is commonly associated with imaging the foetus in pregnant women.

Uses of ultrasound are much broader, however. Other important uses include imaging the abdominal organs, heart, breast, muscles, tendons, arteries and veins. While it may provide less anatomical detail than techniques such as CT or MRI, it has several advantages which make it ideal in numerous situations, in particular that it studies the function of moving structures in real-time, emits no ionizing radiation, and contains speckle that can be used in elastography.

The high frequency sound waves are sent into the tissue and depending on the composition of the different tissues; the signal will be attenuated and returned at separate intervals. A path of reflected sound waves in a multilayered structure can be defined by an input acoustic impedance (ultrasound sound wave) and the Reflection and transmission coefficients of the relative structures.



Ultrasound representation

It is very safe to use and does not appear to cause any adverse effects. It is also relatively inexpensive and quick to perform. Ultrasound scanners can be taken to critically ill patients in intensive care units, avoiding the danger caused while moving the patient to the radiology department. The real time moving image obtained can be used to guide drainage and biopsy procedures. Doppler capabilities on modern scanners allow the blood flow in arteries and veins to be assessed.

14.2.4.4 Elastography

Elastography is a new imaging modality that maps the elastic properties of soft tissue. This modality emerged in the last decade. Elastography can use ultrasound, magnetic resonance imaging and tactile imaging.

14.2.4.5 Tomography

Tomography is the method of imaging a single plane, or slice, of an object resulting in a tomogram. There are two principal methods of obtaining such images, conventional and computer assisted tomography. Conventional tomography uses mechanical means to record an image directly onto X-ray film, while in computer assisted tomography, a computer processes information fed to it

from detectors then constructs a virtual image which can be stored in digital format and can be displayed on a screen, or printed on paper or film.

1. Conventional tomography

In conventional tomography, mechanical movement of an X-ray source and film in unison generates a tomogram using the principles of projective geometry. Synchronizing the movement of the radiation source and detector which are situated in the opposite direction from each other causes structures which are not in the focal plane being studied to blur out. This was the main method of obtaining tomogaphic images until the late-1970s. It is now considered obsolete (except for certain dental applications), having been replaced with computer assisted tomographic techniques.

2. Computer-assisted tomography

In computer-assisted tomography, a computer processes data received from radiation detectors and computationally constructs an image of the structures being scanned. Imaging techniques using this method are far superior to conventional tomography as they can readily image both soft and hard tissues (while conventional tomography is quite poor at imaging soft tissues). The following techniques exist:

- X-ray computed tomography (CT), or Computed Axial Tomography (CAT) scan, is a helical tomography technique (latest generation), which traditionally produces a 2D image of the structures in a thin section of the body. In CT, a beam of X-rays spins around an object being examined and is picked up by sensitive radiation detectors after having penetrated the object from multiple angles. A computer then analyses the information received from the scanner's detectors and constructs a detailed image of the object and its contents using the mathematical principles laid out in the Radon transform. It has a greater ionizing radiation dose burden than projection radiography; repeated scans must be limited to avoid health effects.
- Positron emission tomography (PET) also used in conjunction with computed tomography, PET-CT, and magnetic resonance imaging PET-MRI.

14.2.4.6 Echocardiography (ECG)

When ultrasound is used to image the heart it is referred to as an echocardiogram. Echocardiography allows detailed structures of the heart, including chamber size, heart function, the valves of the heart, as well as the pericardium (the sac around the heart) to be seen. Echocardiography uses 2D, 3D, and Doppler imaging to create pictures of the heart and visualize the blood flowing through each of the four heart valves. Echocardiography is widely used in an array of patients ranging from those experiencing symptoms, such as shortness of breath or chest pain, to those undergoing cancer treatments.

Echocardiography is one of the most commonly used imaging modalities in the world due to its portability and use in a variety of applications. In emergency situations, echocardiography is quick, easily accessible, and able to be performed at the bedside, making it the modality of choice for many physicians.

14.2.4.7 Endoscopy

Endoscopy is looking inside the body for medical reasons using an **endoscope**, an instrument used to examine the interior of a hollow organ or cavity of the body. Unlike most other medical imaging devices, endoscopes are inserted directly into the organ.

Components of endoscope

An endoscope can consist of:

- a rigid or flexible tube.
- a light delivery system to illuminate the organ or object under inspection. The light source is normally outside the body and the light is typically directed via an optical fiber system.
- a lens system transmitting the image from the objective lens to the viewer
- an eyepiece. Modern instruments may be videoscopes, with no eyepiece, a camera transmits image to a screen for image capture.
- an additional channel to allow entry of medical instruments or manipulators.

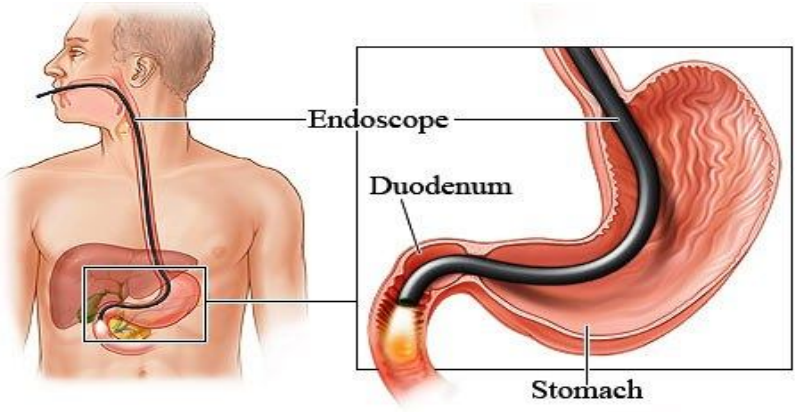


Figure shows endoscopy of stomach

Applications

Health care providers can use endoscopy to review any of the following body parts:

- The gastrointestinal tract(GI tract):
- The respiratory tract
- The ear
- The urinary tract (cystoscopy)
- The female reproductive system (gynoscopy)
- Normally closed body cavities (through a small incision):
    - The abdominal or pelvic cavity
    - The interior of a joint
    - Organs of the chest

Endoscopy is also used for many procedures:

- During pregnancy
    - The amnion
    - The fetus
- Plastic surgery
- Orthopedic surgery

14.2.4.8 Creation of three-dimensional images

Recently, techniques have been developed to enable CT, MRI and ultrasound scanning software to produce 3D images for the physician. Traditionally CT and MRI scans produced 2D static output on film. To produce 3D images, many scans are made, then combined by computers to produce a 3D model, which can then be manipulated by the physician. 3D ultrasounds are produced using a somewhat similar technique. In diagnosing disease of the viscera of abdomen, ultrasound is particularly sensitive on imaging of biliary tract, urinary tract and female reproductive organs (ovary, fallopian tubes). As for example, diagnosis of gall stone by dilatation of common bile duct and stone in common bile duct. With the ability to visualize important structures in great detail, 3D visualization methods are a valuable

14.2.4.9 Use in pharmaceutical clinical trials

Medical imaging has become a major tool in clinical trials since it enables rapid diagnosis with visualization and quantitative assessment.

Imaging techniques such as PET and MRI are routinely used in oncology and neuroscience areas. For example, measurement of tumour shrinkage is a commonly used surrogate endpoint in solid tumour response evaluation. This allows for faster and more objective assessment of the effects of anticancer drugs. In Alzheimer's disease, MRI scans of the entire brain can accurately assess the rate of hippocampal atrophy, while PET scans can measure the brain's metabolic activity by measuring regional glucose metabolism.

14.3 Self learning exercise

1. How image analysis system play important role in biomedical sciences?
2. What do you understand by computer simulation?
3. Discuss the various digital techniques useful in biomedical sciences.
4. Write short note on MRI.
5. How radiology become mile stone in biomedical sciences?

14.4 References

- http:/vlab.amrita.edu
- http://en.wikipedia.org
- www.ymec.com

Unit - 15

Bioinformatics

Structure of the Unit

15.0 Objectives

After going through this unit you will be able to understand:

- Elementary idea of bioinformatics
- Interdisciplinary relationship between biotechnology and information technology
- Applications of bioinformatics
- Role of bioinformatics
- Brief study of genomics & proteomics
- Advantages of bioinformatics in genomics and proteomics
- Study of tools and software involve in secondary structure biomoleclues visualization

15.1 Introduction

Bioinformatics is an emerging field in science which applies the application of information technology, statistics and computer science to solve the mysterious problems of life science in the field of molecular biology. According to NIH definition Bioinformatics is defined as “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data” (http://www.bisti.nih.gov/CompuBio.Def). There is another related term, Computational Biology also define as “the development and application of data-analytical and theoretical methods, and mathematical modeling and computational simulation techniques to the study of biological, mathematical, statistical, behavioral, and social systems”. Whereas the terms

bioinformatics and computational biology get used interchangeably, the former is geared more toward the development of algorithms for analysis of biological data while the latter is concerned with the discovery of new biological knowledge by applying computing resources to large scale modeling and simulation coupled with experimental data. First bioinformatics term was coined by Paulien Hogeweg in 1979. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. Biological databases are always useful to organize and manipulate the data for user requirement. It provides an opportunity to user so that they can easily extract the data and can use it according to the requirement. The science of bioinformatics also involve in the development of the algorithms, tools and software to manage the biological data apart from databases, to analyze the biological problem for example the data of genes and genome, proteins and of several other biomolecules in form of sequence, structure and functions. Bio-informatics is also used to create advancement in databases and uses statistical and computational techniques to finds the answers of the problems which arise from the biological data. Biological data is available inform of sequences, structures and functions. So according the availability of the biological data, bioinformatics works in different research forms for the scientists for example in genetics & genomics, especially in large scale DNA sequencing, next generation sequencing, mapping of DNA, RNA and proteins and so many other combined activities in terms of sequences. In form of structures bioinformatics used to study 3D structure prediction of proteins, 3D structure analysis of proteins, also involve in the creation of software for visualization of biomolecules like proteins. These biological data also incorporate in the comparison of between the species or within the species to get information about the protein sequences and structure of proteins. Though scientists prior got all the information of proteins and DNA in the form of biological data but getting information from this data was very time consuming. Now enormous computational programs have made it easy for the scientists to analyze the data computationally. Programs which are involved in DNA sequencing of the genome are constantly making improvements for an instance BLAST, FASTA and many other programs are useful in sequence similarity for the DNA, protein and so on. Apart from maintaining the large databases, mining useful information from these set of primary and secondary databases is very important. Lot of efficient algorithms has been developed for data mining and knowledge discovery. These

are parallel computing and need fast and computation intensive facilities for handling multiple queries simultaneously.

Structure and Function prediction with great accuracy is one of the aspects in which people are working in the area of bioinformatics. This area is known as structural and functional genomics, where the structure and function of proteins from sequences are identified using a host of similarity search criteria. Visualizing the structure plays an important role in understanding. However, they require computing facilities and special software tools such as MOLMOL, Rasmol, WebLab, etc. Protein structure prediction (protein folding) of the native 3D structure of the protein from its sequence is one of the genuine problem in biological science. A striking observation that has been made is that, though there are more than 10,000 entries of structures in the protein databank, there are only about 1500 unique structure of proteins. Thus the 3D structure of proteins is more or less restricted to a relatively small structure space and any change in structure dramatically alters the function of a protein to make protein diseases when even the slightest change happen in the folding process can turn a desirable protein into a disease. The scientific community considers protein folding as one of the most significant and fundamental problem in biological science that has broad economic and scientific impact and whose solution can be advanced only by applying high-performance computing technologies. Better understanding of how proteins fold will give scientists and doctors better insight into diseases and ways to combat them. Pharmaceutical companies have also design high-tech prescription drugs customized to the specific needs of individual people that are called personalized medicines and doctors could respond more rapidly to changes in bacteria and viruses that cause them to become drug-resistant.

To conclude, today it is possible to perform (using heuristic algorithms) 80% accurate searches perhaps 90 -95% accuracy from the leading software systems. Sensitive algorithms which improve the accuracy in search methodology such as hidden Markov models and Smith-Waterman algorithm are also available but time consuming to execute the search because of the local search methodology. Now to handle these demanding needs, computers are being designed on the biologist required demands. Leading bioinformatics companies are developing software systems which permit research scientists to integrate their diverse data and tools under Common Graphical User Interfaces (GUIs). This creates more opportunity

for research and discovery, through savings in time and data co-ordination. It also permits scientists to share information and provides a powerful solution to archive data.

Thus Bioinformatics has become synonymous with biological research. New academic programs which train students in bioinformatics are providing them with background in molecular biology and in computer science, including programming of the software, database design and analytical approaches. Bioinformatics tools for efficient research will have significant implications in medical sciences and betterment of human lives.

15.1.1 History of Bioinformatics

As millions of biological data is available in raw form and it is always need to manage and stored the data in form of stored houses known as biological databases. First biological database was developed in a very short time span for insulin protein to manage the sequence of this protein as it was first protein to be sequenced in 1956 that to smaller in size up to 51 residues.

Haemophilus influenzea was the first microorganism to be sequenced by The Institute of Genomic Research (TIGR). Later in the mid of nineteen sixties, the first nucleic acid sequence was sequenced which is of Yeast tRNA with 77 bases. In 1965 Margaret Dayhoff, discovered first protein sequence database Atlas of Protein Sequence and Structure (now PIR). Later in 1970, the Needleman-Wunsch algorithm for sequence comparison was published. In 1973, The Brookhaven Protein Data Bank was announced (Robert Metcalfe). The full description of the Brookhaven PDB (http://www.pdb.bnl.gov) was published 1974, First protein structure prediction algorithm Chou and Fasman. 1981 The Smith-Waterman algorithm for sequence comparison was published. Three dimensional structures of proteins were also studied in the same duration, so to maintain the data biological databases were required. PDB was the first structural database which used to store the structural information of the protein in form of two dimensional, three dimensional forms of proteins and so on. Later in 1986, the SWISS PROT database was developed to keep protein sequences for all known organisms whose biological sequences are available.

15.1.2 Applications

Bioinformatics is merger of interdisciplinary fields like statistics, mathematics and computer science & information technology to solve complex biological problems which is related to molecular biology. This interesting field of science has many applications in several research areas where it can be applied.

Sequence Analysis

Sequence analysis means to determine the function of genes in form of encoding or decoding the gene language (Sequence) whether it is regulatory gene, functional or non functional. Application of sequence analysis also involved in the identification of the functional and non functional proteins. There are several powerful tools and software available which perform the duty of analyzing the genome of various organisms. These computational tools also see the DNA mutations in an organism as well as detect those sequences which are related. Shotgun sequence techniques are one of the most important and versatile technique used for sequence analysis of numerous fragments of DNA also some software to see the overlapping of fragments and their assembly.

Genome Annotation

In bioinformatics the other way of sequence analysis can be determine through annotation. Annotation is the technique of bio-informatics which involves the marking of genes within the DNA sequence as well as in the finding of protein coding genes, RNA genes computationally. It is a very important part of the human genome project as it determines the regulatory sequences.

Comparative Genomics

Comparative genomics as its name implies that comparison of genome within the or between the species which determines the genomic structure and functional relationship in form of pattern matching, sequence matching, structure comparison and their functional comparison. For this purpose, intergenomic maps are constructed which enable the scientists to trace the processes of evolution that occur in genomes of different species. These maps are responsible for providing information about the point mutations as well as the information about the duplication of large chromosomal segments. Comparative genomics make ease the research in terms of finding the unknown sequence for non sequenced species through the availability of the known data of model organism by comparing it.

Drug Development and Discovery

The tools of bioinformatics are also helpful in drug discovery, in drug development, in disease diagnosis, in disease cure and disease management. After successful completion of human genome project (HGP) whole genome known sequencing of human genes has enabled the scientists to make medicines and drugs which can target more than 500 genes. Different computational tools, software and drug targets have made ease in drug delivery system and specific because now the mutated or diseased cells can be easily optimized which is easy to know the molecular basis of a disease. Drug designing works in several ways like Knowledge-based drug design which allows the interaction between the known ligand and receptor those can be best fit to each other. *Insilico* drug designing reduce the time and cost for the developed drugs.

Forensic DNA analysis

Bioinformatics involved in Forensic study for the detection of crime spots with the help of following techniques like DNA fingerprinting, Bayesian statistics and likelihood-based methods, personalized healthcare.

Agricultural biotechnology

Agricultural biotechnology is growing field of life science which requires the vast amount of new techniques for the enhancement of research and development. Bioinformatics science has already developed enormous databases based on plants and animals data. Further, other gene expression profiling techniques are also used in agricultural biotechnology to know about the gene expression whether silencing of genes and working of genes to develop the new crop varieties or to enhance the productivity of genes.

Analysis of mutation in Cancer

In case of cancer, effected cells of the genome are rearranged in unpredictable or complex ways. Bioinformatics techniques make it possible to identify the point mutations in various genes of different varieties in cancer that cause multiplication in several neighbor cells. Bio-informatics experts create new software and algorithms to compare the results of sequencing for the collection of germ line polymorphisms and human genome. Some techniques like microarrays are used to identify the gains and losses of chromosomes. To know the point mutations, a technique is called single nucleotide polymorphism arrays. Several hundreds of

SNP have been measured simultaneously from the sequence sites in the whole genome by using these methods.

Analysis of Gene Expression & Protein Expression

Gene expression is basically to know how genes are expressed within the genome which can be determined by measuring the level of messenger RNA (mRNA) using several expression techniques like microarray, serial analysis of gene expression (SAGE) and expressed cDNA sequence tag sequencing. Data of cancerous cells can be compared with the data of non-cancerous cells through microarray. Mass spectrometry and protein microarray are used to obtained proteins present in the genome. The protein microarrays and mass spectrometry is the main concern of bioinformatics in case of protein identification. The basic source for using Protein microarrays are the mRNA for dealing with the similar problems like with the microarrays. The large amount of mass data can be compared through the mass spectrometry with databases of protein sequences.

Prediction of Protein Structure

In the beginning it was very easy to get to know about the primary structure of protein which is in form of linear protein sequence made up of amino acids (building block of protein) but is difficult to predict the secondary, tertiary and quaternary structure of proteins. Crystallography is the technique which is useful in the structure prediction of protein as well as so many bioinformatics software also available in the 2 Dimensional and 3 Dimensional structure prediction.

15.2 Introduction about Genomics & Proteomics

15.2.1 Genome

As genome is made up of the word “Ome” which is collection of the whole, defined as it is the complete genetic material in the cell of an organism, which may be referred as the total genetic code. The language of genetic code is made up of the nucleotide language that is A (Adenine), T (Thymine), G (Guanine) and C (Cytosine) or U stands for (Uracil) in place of T when it comes in form of RNA. The genetic code consists of three pair of nucleotides, which form the one amino acid. These amino acids are the building blocks of proteins that are involved in coding of different forms of protein structures. The 20 amino acids are responsible

in the formation of complex forms of secondary structures of proteins and in many other forms as well like tertiary and quaternary.

15.2.2 Genomics

The study of the genome in form of its formation, storage and incorporation of the genome in to other cellular activities is under comes with genomics. In a cell the genetic material DNA stored the information that flows in form of RNA and protein as well. DNA is made up of hereditary units of the cell called genes. The genes code the information flow in form of "central dogma" to produce protein. The process involves the DNA to RNA to Protein synthesis. This require the need to develop a understanding for protein existence to their role and understanding the role of genome in transcription, where the initiation, elongation and termination takes place to make precise models of RNA later the RNA splicing when the introns (Non coding part of genome) are removed and exons (Coding part of genome) are joined together for protein production. Beside that there is an essential process also takes place that is signal transduction in which a series of signals passed through receptors in the cell membrane to activate transcription are called signal transduction pathways. This involve in the determining of mechanistic understanding of protein evolution by protein to DNA, protein to RNA and protein to protein interaction. To understand such complex model there is a need to evaluating the several genes and protein with their respective sequences (Snyder and Champness 2007; Lewin 2007)

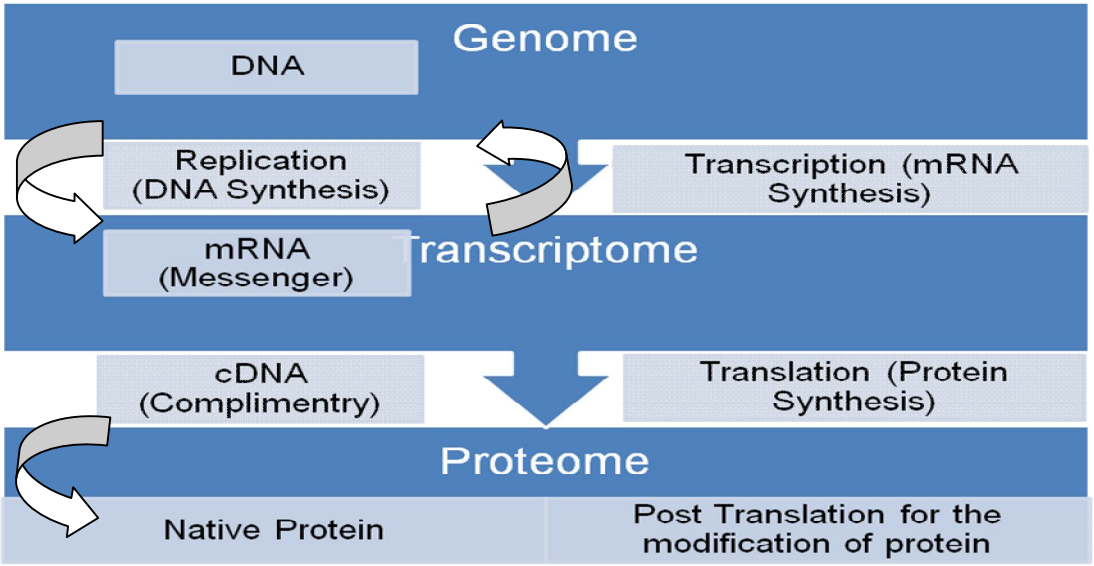


Fig: 1 Story, the Genome to Proteome synthesis in form of flow chart, this process is called Central Dogma in molecular biology. Here in this figure conversion of genome to proteome takes place. How gene is producing the protein, how many processes are involved.

Genomics consists of the following divisional areas:

15.2.2. I Functional genomics

15.2.2. II Structural genomics

15.2.2. III Comparative genomics

There are numerous functional genomics approaches to identify the unknown genes and proteins and several expression techniques to know the expression of genes and proteins. Functional genomics is the revolutionary area of genomics which is major involved in identification of genes and proteins like in case of making comparison between the known and unknown gene data and also to get to know the difference between diseased and normal genes through the microarray technique and Serial Analysis of Gene Expression (SAGE) analysis for example on cancer disease. Scientists are currently using these functional approaches very rapidly in the study of phylogenetic analysis at the evolutionary level, also to identify the relationship between the intergenic and intragenic species of the organism.

In structural genomics we study about the structural analysis of genes and proteins to identify the location information, interaction, interrelationship of many genes. The activity of genes in form of function like co expression and association of more than one gene with other neighbor genes can easily be identified. In structural bioinformatics mainly deals with approaches related to traits which are controlled by one or only a few genes. Together, in a combination of functional genomics and structural genomics together scientist will be efficiently able to create new traits in form of novel genes in the species by using the combination of information. Genetic engineering is the one of the method through which new genes are introducing in to organism genome known as transgenes. For example, both functional and structural genomic information are valuable in the production of new verities of species of plants and animals. **2http://www.cimmyt.org/Journal of Computing Science and Engineering, Vol. 1, No. 1, September 2007**

Comparative genomics is one of latest upgraded area of genomics which is booming now days by making comparison between the species of same genera or

of different genera. Sequence comparison is the most common way to compare the genome of species through the availability of data in form of genes and proteins. So many new facts have already been drawn on the basis of comparative analysis. For example Due to the difficulty in understanding the sequence conservation in viral proteins, certain crucial approaches of sequence comparison had to be laid down. Protein sequence comparison is frequently used rather than nucleotide sequences because of the reliability of protein sequence (no noncoding or junk sequences are present in protein sequence). Multiple sequence alignment is most reliable than the pairwise sequence alignment for an instance identification of conserved patterns or motifs in multiple sequences are responsible for the potential relationships in sequences or structures. Sequence similarity is not only useful to find the common ancestry but also contributing to know the existence of the convergence relationships between unrelated sequences with very few similarities of the organisms. So whenever there is statistically significant sequence and structural similarity found between proteins and genes it indicates the evidence of homolgs in between of divergent sequences as well (Koonin and Galperin 2003). Homologs can be of two types, namely Orthologs and Paralogs, and crucial to the understanding of evolutionary relationships between genomes and gene functions.

15.2.3 Proteomics

Proteome is the whole protein content, the word “Ome” comprises the all from universe named as proteome. The study of the whole protein content is under comes with proteomics, the term was coined by “Marc Wilkins”. The aim of proteomics is qualitative and quantitative measurement of protein expression specifically under the influence of disease perturbations and for the drug development (Anderson and Anderson 1998). In proteomics it is very important to know about the basic knowledge of protein composition which is made up of amino acids and their existence in form of motifs and conserved domains, responsible for the protein structures. Domains provides the nomenclature for proteins because it exists from the evolution and independently involve in the protein functioning. This particular biological activity of domain is very often found in more than one organism which again indicates the relationship between the organisms (Lones and Tyrrell 2005; Seehuus et al. 2005). The proteomic consist of the post translational study of proteins like how protein is transported, filtered, in actual structure and in functional state.

15.3 Software for Visualization of Secondary Structures of Biomolecules

Software and tools of bio-informatics vary from the simple command line tools to the complex databases and programs. SOAP and REST based interfaces are used for variety of purposes and applications of bio-informatics. They allocate an application on one computer in some part of the world to use the algorithms, computing resources and data on servers in other parts of the world. Computers and software tools are extensively used for creating these databases and to predict the function of proteins, model the structure of proteins, determine the coding (useful) regions of nucleic acid sequences, find suitable drug compounds from a large pool and optimize the drug development process by predicting possible targets. Few are the following software and tools which involved in the molecular structure prediction such as Cn3D, Chime, MolView, Protein Viewer, Swiss PDB, RasMol, Protein Explorer, Kinemage, WebMol and so on.

15.3.1 These programs provide the graphical view of the molecular structure by incorporating the following graphical features:

- Representation of structures in form of wireframe with full atomic detail.
- Color coding schemes in different colors in form of different shapes of molecule
  atom type
- In secondary structure
- Alpha & Beta chain
- Burial state
- HSSP (High Sequence Similarity Pattern) provides sequence conservation.
- side stereo view with adjustable separation of the images
- graphical facilities in from mouse driven zoom, translate and drawing slab manipulations
- to select the image through selecting focus; i.e. setting point of rotation to a desired position

- to select the angles through selecting residues according to type or burial state and regions within proteins
- to drawing of the animated structure by animations of rocking motions with adjustable angular range
- to show the clarity in the view of structure by variable background colors

Tools

- Measurement of distance, angle, dihedral angle
- Detection of steric conflicts
- Analysis of peptide bond planarity analysis
- Distance matrix plot and other 2D-projection plots for the interaction
- Interactive Ramachandran plot
- dot surface for the interaction and Vander Waals surface area
- Numerical calculation of solvent accessible and volume detection of cavities
- Axes assignments for secondary structural elements
- Secondary structural element packing analysis
- Main chain hydrogen bond detection

15.3.2 Cn3D

Cn3D is an application to view the 3 dimensional structure of protein. It is a web browser application which is based on windows platform and can also run on Macintosh and UNIX platform as well that allow anybody to perform work with it. This application is provided by NCBI's Entrez retrieval service system. Cn3D runs on Windows, Macintosh, and UNIX. Through Cn3D user can display structure, sequence, and alignment simultaneously and now has powerful annotation and alignment editing features as well. The detailed working of Cn3D is available in below mentioned link where you can get user guide. There are several examples available in the tutorial, along with instructions for the users to get started.

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

15.3.3 (Aka MDL Chime):

Chime is an application to show the 3 dimensional molecular structures. Its features are very similar to RasMol because of its origin. The difference between Chime and RasMol are based on their availability. RasMol is standalone program which means it can be downloaded on the personal computers (perform outside your browser) while Chime can work from sites directly on a webpage. Chime works on author provided molecules information where as RasMol perform for any molecule for which an atomic coordinator PDB file should be known. Chime is very helpful in the study of structural biochemistry.

Chime is available on the below mentioned hyperlink which is produced by MDL Informatics Systems, Inc. It is also known as MDL Chime, it is plug-in that interactively displays 2D and 3D molecules directly in web pages. In its working user can rotate, reformat and save the molecules for use in other programs as well but it will be required to accept a license agreement for installation.

http://www.umass.edu/microbio/chime/

15.3.4 MolView:

Molview is Macintosh based program for the molecular structure visualization. It was developed by Smith Labs which is freely available. It is a collection of tools which work upon few *ab initio* program packages (Gaussian, Molpro, Molcas).

15.3.5 Protein Explorer:

Protein Explorer is freely available software used for visualizing the three dimensional structure of bio molecules such as DNA, RNA, proteins and for several other macromolecules. It is used to identify the interactions and binding between the ligands and receptors, inhibitors and in drug composition as well.

It is suitable for the students, yet it is also widely used by research students and researchers.

http://www.umass.edu/microbio/chime/explorer/

15.3.6 Swiss PDB:

Swiss-PDB viewer is user friendly application that makes the interface which allows analyzing several proteins at the same time. It provides the facility to superimpose the protein molecule to deduce the structural alignments also to

compare their active sites. Mutations in Amino acids, H-bonds angles and distances between atoms are also easy to obtain because of the intuitive graphic and menu interface facility. Swiss-PdbViewer is linked to Swiss-Model, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) in collaboration between GlaxoSmithKline R&D and the Structural Bioinformatics Group at the Biozentrum in Basel. Swiss PDB involves generating the primary thread onto a 3D template. Swiss-PdbViewer can also read electron density maps. In addition, various modeling tools are integrated and command files for popular energy minimization packages can be generated.

http://www.expasy.org/spdbv/

15.3.7 RasMol:

RasMol is a program initially developed by Roger Sayle for molecular graphics visualization. This application provides the graphic visualization of three dimension structure of macromolecules to prepare publication-quality images. RasMol is a standalone application which can be run anywhere without web browser on the computers once after downloading. It's a user friendly application for windows, Macintosh and UNIX platform as well. When a scientific tool works as software source code is an important element in achieving full understanding of that tool for accessing it. Later the series of RasMol 2.7 were released in 1999, when RasMol formally became an open source program. Most scientific software source code was freely and openly available with a minimum of formalities.

www.rasmol.org/

15.3.8 WebMol:

WebMol is java based application which is available on sites and in standalone form as well. WebMol was designed to display and analyze structural information contained in the Brookhaven Protein Data Bank (PDB) which means it uses the URL FTP file, text file, string file as an input in PDB format. This application implements in CGI interface

15.3.9 NAMD: Sacalable Molecular Dynamics

NAMD program designed for high performance simulation of macromolecular system. NAMD provides a parallel molecular dynamics code to hundreds of processors and tens of processors on low cost commodity clusters as well. It is also used to visualize the graphic representation of three dimensional molecular

structures of biomolecules (Protein, DNA & RNA). NAMD can run on individual machines (Laptop & Desktop computers) but it supports too few particular file formats like AMBER and CHARMM which code for potential functions etc. NAMD is mainly involved in the study of classical molecular dynamic force field, equations of motion integration methods along with the efficient electrostatics evaluation algorithm.

Most commonly NAMD use is illustrated with representative applications to small, medium, and large biomolecular system. To discuss the key features of NAMD and their benefits of combining with the molecular graphics/sequence analysis software VMD and the grid computing/collaboratory software BioCoRE the information is available at © 2005 Wiley Periodicals, Inc. J Comput Chem 26: 1781–1802.

http://www.ks.uiuc.edu/Research/namd/

The above mentioned software is the most commonly used for the visualization of three dimensional structures of biomolecules mainly for proteins. Apart from that there are few following tools and software mentioned below in the list which is available in the market for user flexibility. The list is distinguished on the basis of functional property if the software.

15.4 Table for Visualization tools & Software

Visualization

Garlic	A freely available molecular visualization program
Gene View II	Interactive GenBank Entry Visualization tool
gff2ps	By following postscript converting genomic annotations in GFF format
ModView	In real time visualization and analysis of either in form of multiple biomolecule structures and/or in form of sequence alignments
PyMol	PyMOL is a molecular graphics system designed for

	visualization and rapid generation of high-quality molecular graphics images and animations with the help of embedded Python interpreter
RasMol	Freely available program which displays molecular structure in different visualization.
VEGA	VEGA was developed to create a bridge between most of the molecular software packages, like BioDock, Quanta/CHARMm, Insight II, MoPac, etc.
VMD	VMD is a molecular visualization packages for displaying, visualizing, animating, and analyzing biomolecular systems using 3-D graphics and inbuilt scripting in which it is designed.

Structure

ABaCUS	ABaCUS is a program to investigate the significance of the putative correspondence between exons and units of protein structure.
caRNAsta	Comparative Analysis of RNA structures by Tree Alignment
CNS	Crystallography & NMR System for biomolecular structure analysis and visualization.
COVE	COVE is a program involves in the implementation of stochastic context free grammar methods for RNA sequence/structure analysis.
Garlic	A freely available molecular visualization program
LIBELLULA	LIBELLULA is a web server program based on neural network programming to evaluate fold recognition results in protein structure prediction.

MODELLER	Tool for homology protein structure modelling by satisfaction of spatial restraints.
ModView	In real time visualization and analysis of either in form of multiple biomolecule structures and/or in form of sequence alignments.
MOPAC7	It is a molecular orbital package for the study of chemical structures and reactions in semi-empirical manner.
PyMol	PyMOL is a molecular graphics system designed for visualization and rapid generation of high-quality molecular graphics images and animations with the help of embedded Python interpreter
RNA GENiE	A web based program for the prediction of RNA patterns in genomic DNA sequences
RnaViz	A GUI program for producing publication-quality secondary structure drawings of RNA molecules. It is a user-friendly and portable program.

Structural Alignments

Deep View Swiss-PdbViewer	Swiss-PdbViewer is a user based program that provides an interface allowing to analyse several proteins at the same time.

Structural Biology

DINO	A 3D visualization program based on real time for structural biology data.

Protein Structure

ABaCUS	ABaCUS is a program to investigate the significance of the

	putative correspondence between exons and units of protein structure.
Garlic	A freely available molecular visualization program
LIBELLULA	LIBELLULA is a web server program based on neural network programming to evaluate fold recognition results in protein structure prediction.
MODELLER	Tool for homology protein structure modelling by satisfaction of spatial restraints.
ModView	In real time visualization and analysis of either in form of multiple biomolecule structures and/or in form of sequence alignments.

Protein Structure Modelling

MODELLER	Tool for homology protein structure modelling by satisfaction of spatial restraints.

Protein Structure Prediction

LIBELLULA	LIBELLULA is a web server program based on neural network programming to evaluate fold recognition results in protein structure prediction.

Protein/DNA/RNA Family Clustering

ModView	In real time visualization and analysis of either in form of multiple biomolecule structures and/or in form of sequence alignments.

Protein Interaction

ZDOCK	Program for the docking of biomolecules like Protein-protein complex structure prediction

Protein Modelling

BRAGI	BRAGI is a protein modelling program to display the visualized interactive 3Dimensional structure. It was developed for the special purpose to model unknown proteins from the structure of a known one.

Molecular Mechanics

AMMP	A versatile modeling program for the molecular dynamic approach.

Molecular Modeling

Oslet	A molecular modeling and simulation program designed in Java environment. It is commonly used for the learning for the students.
WHAT IF	It is a protein structure analysis program that can be used for prediction of mutant structures, structure verification and molecular graphics, etc

Molecular Modelling

Babel	A program designed to the conversion of file formats internally which is currently used in molecular modeling.

Molecular Rendering

Garlic	A freely available molecular visualization program

Molecular Simulations

Oslet	A molecular modeling and simulation program designed in Java environment. It is commonly used for the learning for the students.

Molecular Visualization

Garlic	A freely available molecular visualization program

RasMol	Freely available program which displays molecular structure in different visualization.
VEGA	VEGA was developed to create a bridge between most of the molecular software packages.
VMD	VMD is a molecular visualization packages for displaying, visualizing, animating, and analyzing biomolecular systems using 3-D graphics and inbuilt scripting in which it is designed.

Molecular Visualization Program

Garlic	A freely available molecular visualization program

Modeling

AMMP	A versatile modeling program for the molecular dynamic approach.
Kintecus	Program for run chemical kinetics/fitting of catalyst reactor, and enzyme reactions
MMTK-2.2	This is a molecular modeling tool-kit 2.2.
Oslet	A molecular modeling and simulation program designed in Java environment. It is commonly used for the learning for the students.
WHAT IF	It is a protein structure analysis program that can be used for prediction of mutant structures, structure verification and molecular graphics, etc

Molecular Dynamics High Performance Simulator

Gromacs	The fastest Molecular Dynamics program of world.

Molecular Dynamics Modeling PDB Force Field

MMTK-2.2	This is a molecular modeling tool-kit 2.2.

Molecular Dynamics Simulations

GDIS	This program is based on GTK and has been successfully useful for the visualization of isolated molecule.

Electron Density Maps

Deep View Swiss-PdbViewer	Swiss-PdbViewer is a user based program that provides an interface allowing analysing several proteins at the same time.

Chemical Structures

MOPAC7	It is a molecular orbital package for the study of chemical structures and reactions in semi-empirical manner.

Bond and Atom Rendering

Spock	A molecular graphics program

3d Structure Visualization

PyMol	PyMOL is a molecular graphics system designed for visualization and rapid generation of high-quality molecular graphics images and animations with the help of embedded Python interpreter

15.5 Summary

Now days there are thousands of organisms have already been sequenced, their genomic and proteomic data is available in biological databases. In past as well several nations have already planed their research projects on bioinformatics after

successful completion of human genome project. Bioinformatics is the merger of many disciplines in a combination of computer science and information technology, statistics and mathematics which is helpful to solve several numbers of life science, biotechnology research problems with the help of molecular biology. Bioinformatics science takes the input of the problem in form of molecular language either sequences or structure, with the aid of sequencer techniques and tools & software. In future as well research will be continued to cater their service to the world.

15.5 Glossary

- **Biotechnology:** Technology of biological science which is useful to solve the biological problem.
- **Bioinformatics:** Combinatorial study of biological science and information technology with the combination of statistics, mathematics and computer applications to make biological research efficient and time consuming.
- **Computational Biology:** It involves the applications of computer programs for the development of analytical data by using the simulation and mathematical modeling.
- **Genomics:** study of whole genome content of cell.
- **Proteomics:** Study of the whole proteome content of the cell.
- **Replication:** Synthesis of DNA molecule from the DNA threads itself.
- **Transcriptome:** Study of the whole transcriptome (RNA material) content of the cell.
- **Transcription:** Synthesis of RNA molecule from DNA thread.
- **Translation:** Synthesis of protein molecule from the mRNA thread.

15.6 Self-Learning Exercise

Section -A (Very Short Answer Type)

1. Who coined the term bioinformatics?
2. Who coined the term proteomics?

3. What is the full form of SAGE?
4. Write the name of the first microorganism that is fully sequenced?
5. GUI stands for _____
6. PDB stands of _____
7. In which year Protein sequence database ATLAS was discovered?
8. In nucleotides,in the conversion of DNA to RNA which base is replaced to Thymine?

Section -B (Short Answer Type)

1. Write a note on the applications of bioinformatics.
2. Briefly explain about central dogma.
3. Mention the role of RasMol.
4. Define comparative genomics.
5. Write down the role of bioinformatics as interdisciplinary stream.

Section -C (Long Answer Type)

1. Write in detail about the evolution of bioinformatics? How it is involve in human welfare?
2. Write an explanatory note on tools and software for visualization of biomolecules?
3. Discuss the importance of bioinformatics in genomics & proteomics.
4. Define the Applications of genomics & proteomics?

Answer Key of Section-A

1. Paulien Hogeweg
2. Marc Wilkins
3. Serial Analysis of Gene Expression
4. Haemophilus influenzae
5. Graphic User Interface
6. Protein Data Bank
7. 1965

8. Uracil

15.7 References

- Snyder L and Champness W, 2007, *Molecular Genetics of Bacteria*, ASM Press, Journal of Computing Science and Engineering, Vol. 1
- © 2005 Wiley Periodicals, Inc. J Comput Chem, 26: 1781–1802, 2005
- http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
- http://www.umass.edu/microbio/chime/
- http://www.umass.edu/microbio/chime/explorer/
- http://www.expasy.org/spdbv/
- http://www.umass.edu/microbio/rasmol/
- www.rasmol.org/
- http://www.ks.uiuc.edu/Research/namd/