M.Sc. Botany, Semester: II
LBT-201 / LBC-801: Computational Biology -2013
AR- 7161
Scheme of valuation
Section A: Answer all the questions (2 X 10 = 20 marks)

1 (i). Formula of mean

(a) For ungrouped data

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Where $\bar{x}$ is arithmetic mean, $x_1+x_2+X_3$………$+X_n$ are observations for x variable and n is number of observations

(b) For grouped data

$X = \Sigma fx / \Sigma f$

Where X is arithmetic mean, Σf is sum of all frequencies and x is variable.

(ii). Formula of median

(a) For ungrouped data

Median (M) = $n+ 1/2_{th}$ term (n is no of observations and it is odd)

M = {n/2 +(n/2+1)}/ 2 (When n is even)

(b) Grouped data

Discrete series

M = (n+1/2)th N = Σ*f*

Continuous series

M = {L + (N/2-C/fm)} i

Where, L = lower limit of the class in which median lies.
N = Total number of frequencies
fm= Frequency of the class in which the median lies.
C = Cumulative frequency of the class preceding the median class.
i = Width of the class interval in which the median lies.

(iii). Formula of student *t* test

$$t = [X_1-X_2] / SE$$

Where, X_1 and X_2 are mean values and SE is standard error value = SD $\sqrt{1/n_1+1/n_2}$ (SD = standard error value and n_1 and n_2 are number of observations for two different sets).

(iv). Formula of Chi-square test

$$x^2 = \Sigma [(O-E)- ½]^2 / E$$

Where O = Observed numbers of frequencies
E = Expected number of frequencies
½ = Yates correction (If E is less than 5)

(v). c
(vi). d
(vii). b
(viii). d
(ix). b

(x). a

Section B : Answer any four question (Each question carry 10 Marks)

2. Short notes on (a)Measures of central location

A measure of central tendency (or location) is a single value that attempts to describe a set of data by identifying the central position within that set of data. They are also classed as summary statistics.

The most common measures of central location are

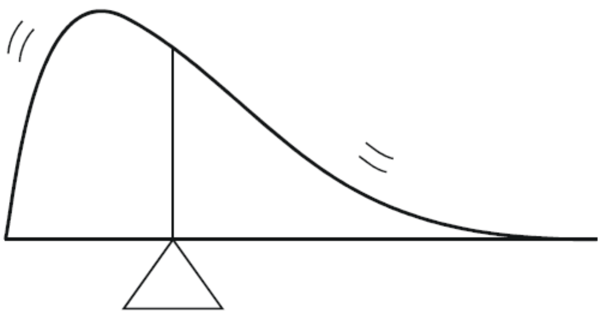
1. Mean or Arithmetic mean 2. Median 3. Mode

Mean

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. the sample mean, usually denoted by $\bar{x}$ (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x}{n}$$

Mean is the center of gravity of the distribution



Median

The median is the middle score for a set of data that has been arranged in order of magnitude. It can be calculated as following for different conditions.

For ungrouped data

Median (M) = n+ 1/2$_{th}$ term (n is no of observations and it is odd)

M = {n/2 +(n/2+1)}/ 2 (When n is even)

Grouped data ; Discrete series **M = (n+1/2)th N = Σ*f***

Continuous series **M = {L + (N/2-C/fm)} i**

Where, L = lower limit of the class in which median lies.
N = Total number of frequencies
fm= Frequency of the class in which the median lies.
C = Cumulative frequency of the class preceding the median class.
i = Width of the class interval in which the median lies.

Mode

The **mode** is the value that occurs most often in a set of data. For example, in the following parity data the mode is 1, because it occurs 4 times, which is more than any other value:

0, 0, 1, 1, 1, 1, 2, 2, 2, 3, 4, 6

Characteristics of Central location

1. It should be rigidly defined
2. It should be based on all items
3. It should be easily understood
4. It should not be unduly affected by the extreme values
5. It should be least affected by the fluctuations of the sampling
6. It should be easy to interpret
7. It should be easy to subject to further mathematical calculations.

(b) Scope of Biostatistics

Bio-statistics is a branch of applied statistics that is management and analysis of numerical data on people, health, disease, medical treatments and procedures. It includes vital statistics, public health statistics, and demography. Biostatistics is divided into 2 branches: descriptive and analytic.

(1) Descriptive statistics deals with collection, organization, presentation, and summarization of data.

(2) Analytic statistics deals with drawing logical and objective conclusions about a sample or a population.

Biostatistics provides the tools for the summary and digestion of a lot of numerical laboratory and clinical data including critical reading and understanding of scientific literature.

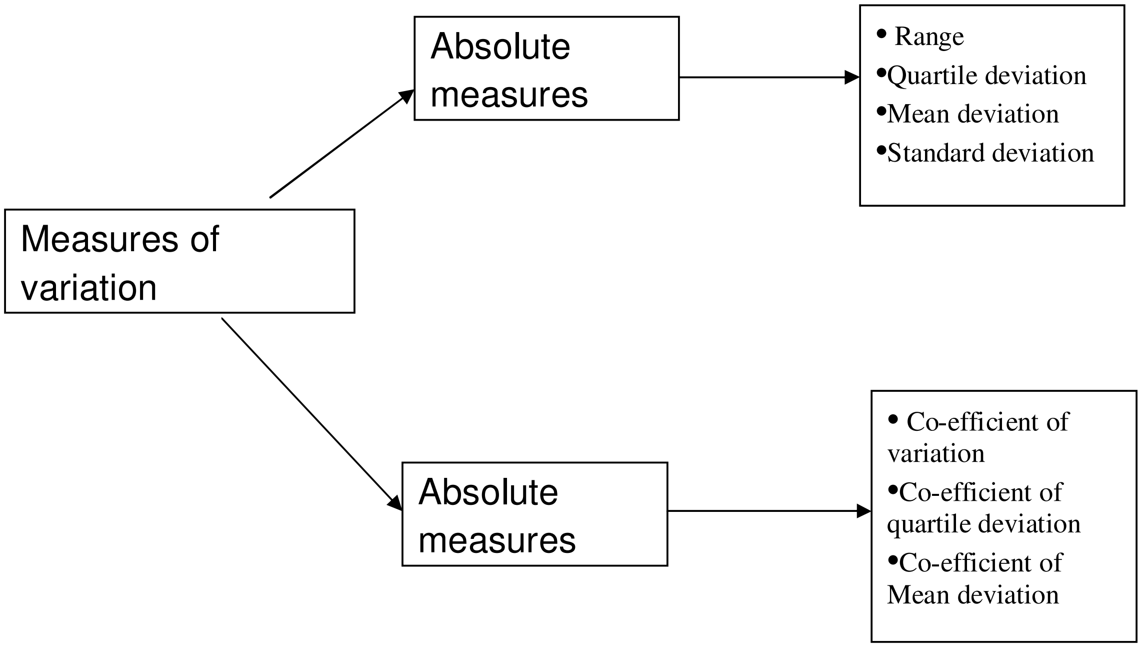
Following are the main applications of biostatistics:

- Public health, including epidemiology, health services research, nutrition, environmental health and healthcare policy & management.
- Design and analysis of clinical trials in medicine
- Population genetics, and statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals (animal breeding). In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics
- Analysis of genomics data
- Ecology, ecological forecasting
- Biological sequence analysis
- Systems biology for gene network inference or pathways analysis

3. Short notes on

(a) Measures of Dispersion (or variation)

Measure of variation is a measure that describes how spread out or scattered a set of data. It is also known as measures of dispersion or measures of spread. It can be shown by a schematic diagram as following:



The most commonly used measure of dispersion is standard deviation (SD). SD is the square root of the arithmetic mean squares of deviation from arithmetic mean or " root mean square deviation from mean

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Where **σ** is SD, N = number of observations, μ is mean of all the values, and x_i is individual values of observations.

Requirement of a good measures of Dispersions

1. It should be easily calculated.
2. It should be rigidly defined.
3. It should be based on all observations.
4. It should not be unduly affected by the extreme values.
5. It should have sampling stability
6. It should have ability to subject to further algebraic treatment.

(b) Sampling techniques

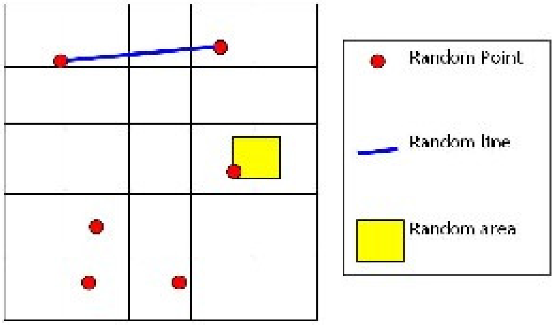
- A shortcut method for investigating a whole population
- Data is gathered on a small part of the whole parent population or sampling frame, and used to inform what the whole picture is like

Three main types of sampling strategy:

- Random
- Systematic
- Stratified

Random sampling

- Least biased of all sampling techniques, there is no subjectivity - each member of the total population has an equal chance of being selected
- Can be obtained using random number tables



A random number grid showing methods of generating random numbers, lines and areas.

Advantages and disadvantages of random sampling

Advantages:

- Can be used with large sample populations
- Avoids bias

Disadvantages:

- Can lead to poor representation of the overall parent population or area if large areas are not hit by the random numbers generated. This is made worse if the study area is very large
- There may be practical constraints in terms of time available and access to certain parts of the study area

Systematic sampling

Samples are chosen in a systematic, or regular way.

- They are evenly/regularly distributed in a spatial context, for example every two metres along a transect line
- They can be at equal/regular intervals in a temporal context, for example every half hour or at set times of the day
- They can be regularly numbered, for example every 10th house or person

Advantages and disadvantages of systematic sampling

Advantages:

- It is more straight-forward than random sampling
- A grid doesn't necessarily have to be used, sampling just has to be at uniform intervals
- A good coverage of the study area can be more easily achieved than using random sampling

Disadvantages:

- It is more biased, as not all members or points have an equal chance of being selected
- It may therefore lead to over or under representation of a particular pattern

Stratified sampling

This method is used when the parent population or sampling frame is made up of sub-sets of known size. These sub-sets make up different proportions of the total, and therefore sampling should be stratified to ensure that results are proportional and representative of the whole.

Advantages and disadvantages of stratified sampling

Advantages:

- It can be used with random or systematic sampling, and with point, line or area techniques
- If the proportions of the sub-sets are known, it can generate results which are more representative of the whole population
- It is very flexible and applicable to many geographical enquiries
- Correlations and comparisons can be made between sub-sets

Disadvantages:

- The proportions of the sub-sets must be known and accurate if it is to work properly
- It can be hard to stratify questionnaire data collection, accurate up to date population data may not be available and it may be hard to identify people's age or social background effectively

4. Short notes on

(a) Testing the significance level

- Every test of significance begins with a ***null hypothesis H_0***. H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.
- The ***alternative hypothesis***, H_a, is a statement of what a statistical hypothesis test is set up to establish.
- The *significance level* α for a given hypothesis test is a value for which a *P-value* less than or equal to α is considered statistically significant. Typical values for α are 0.1, 0.05, and 0.01. These values correspond to the probability of observing such an extreme value by chance.

  For example, if the *P-value* is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.
- Another interpretation of the significance level α, based in *decision theory*, is that α corresponds to the value for which one chooses to reject or accept the null hypothesis H_0. In the above example, the value 0.0082 would result in rejection of the null hypothesis at the 0.01 level.

Steps in Testing for Statistical Significance

1) State the Research Hypothesis
2) State the Null Hypothesis
3) Select a probability of error level (alpha level)
4) Select and compute the test for statistical significance
5) Interpret the results

(b) Difference between correlation and regression

Correlation	**Regression**
1. Relationship between two or more variables which vary in sympathy with the other in the same or the opposite direction	1. Mathematical measure showing the average relationship between two variables
2. Bothe variables x and y are random	2. x is random and y is fixed.
3. Finds our the degree of relationship between two variables.	3. indicates the cause and effect relationship between the variables
4. It is used for testing and verifying the relationship between two variables.	4. It is used for prediction of one value in respect to other given value
5. The co-efficient of correlation in a relative measure and ranges between -1 to +1.	5. Regression co-efficient in an absolute figure. If we know the values of independent variable, we can find the value of dependent variable.
6. It has very limited application considering only linear relationship	6. It has relatively wide application considering the predicting ability.

5. Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. The need for Bioinformatics capabilities has been precipitated by the explosion of publicly available genomic information resulting from the Human Genome Project.

The goal of this project – determination of the sequence of the entire human genome (approximately three billion base pairs) – will be reached by the year 2002. The science of Bioinformatics, which is the melding of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets for drug discovery.

Use of Bioinformatics in Biological sciences:

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

There are two fundamental ways of modelling a Biological system (e.g., living cell) both coming under Bioinformatic approaches.

Static

Sequences – Proteins, Nucleic acids and Peptides

Interaction data among the above entities including microarray data and Networks of proteins, metabolites

Dynamic

Structures – Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides (structures studied with bioinformatics tools are not considered static anymore and their dynamics is often the core of the structural studies)

Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites

Multi-Agent Based modelling approaches capturing cellular events such as signalling, transcription and reaction dynamics

Other uses of Bioinformatics

1. Genome annotation
2. Sequence analysis
3. Computational evolutionary biology
4. Literature analysis
5. Analysis of gene expression
6. Analysis of regulation
7. Analysis of protein expression
8. Analysis of mutations in cancer
9. Comparative genomics
10. Network and systems biology

6. Multiple sequence alignment:

Multiple Sequence Alignment (MSA) is a sequence alignment of three or more biological sequences, generally Protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex.

Various methods of Multiple sequence alignment:

A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as "iterative" because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA. One reason progressive methods are so strongly dependent on a high-quality initial alignment is the fact that these alignments are always incorporated into the final result — that is, once a sequence has been aligned into the MSA, its alignment is not considered further. This approximation improves efficiency at the cost of accuracy. By contrast, iterative methods can return to previously calculated pairwise alignments or sub-MSAs incorporating subsets of the query sequence as a means of optimizing a general objective function such as finding a high-quality alignment score.

A variety of subtly different iteration methods have been implemented and made available in software packages; reviews and comparisons have been useful but generally refrain from choosing a "best" technique. The software package PRRN/PRRP uses a hill-climbing algorithm to optimize its MSA alignment score and iteratively corrects both alignment weights and locally divergent or "gappy" regions of the growing MSA. PRRP performs best when refining an alignment previously constructed by a faster method.

Hidden Markov models

Hidden Markov models are probabilistic models that can assign likelihoods to all possible combinations of gaps, matches, and mismatches to determine the most likely MSA or set of possible MSAs. HMMs can produce a single highest-scoring output but can also generate a family of possible alignments that can then be evaluated for biological significance. HMMs can produce both global and local alignments. Although HMM-based methods have been developed relatively recently, they offer significant improvements in computational speed, especially for sequences that contain overlapping regions.

Typical HMM-based methods work by representing an MSA as a form of directed acyclic graph known as a partial-order graph, which consists of a series of nodes representing possible entries in the columns of an MSA. In this representation a column that is absolutely conserved (that is, that all the sequences in the MSA share a particular character at a particular position) is coded as a single node with as many outgoing connections as there are possible characters in the next column of the alignment. In the terms of a typical hidden Markov model, the observed states are the individual alignment columns and the "hidden" states represent the presumed ancestral sequence from which the sequences in the query set are hypothesized to have descended. An efficient search variant of the dynamic programming method, known as the Viterbi algorithm, is generally used to successively align the growing MSA to the next sequence in the query set to produce a new MSA.

Phylogeny-aware methods:

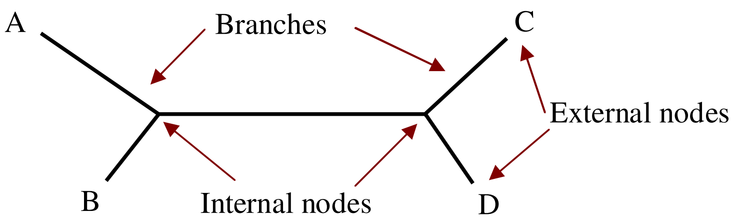
Most multiple sequence alignment methods try to minimize the number of insertions/deletions (gaps) and, as a consequence, produce compact alignments. This causes several problems if the sequences to be aligned contain non-homologous regions, if gaps are informative in a phylogeny analysis. These problems are common in newly produced sequences that are poorly annotated and may contain frame-shifts, wrong domains or non-homologous spliced exons.

Use of MSA:

Multiple sequence alignments can be used to create a phylogenetic tree his is made possible by two reasons. The first is because functional domains that are known in annotated sequences can be used for alignment in non-annotated sequences. The other is that conserved regions known to be functionally important can be found. This makes it possible for multiple sequence alignments to be used to analyze and find evolutionary relationships through homology between sequences. Point mutations and insertion or deletion events (called indels) can be detected.

7. A phylogenetic tree

A tree is an acyclic connected graph that consists of a collection of nodes (internal and external) and branches connecting them so that every node can be reached by a unique path from every other branch.



An unrooted phylogenetic tree joining 4 taxonomic units.

Features of a phylogenetic tree

In the area of phylogenetic inference, trees are used as visual displays that represent hypothetical, reconstructed evolutionary events. The tree in this case consists of:

- internal nodes which represent taxonomic units such as species or genes; the external nodes, those at the ends of the branches, represent living organisms.
- The lengths of the branches usually represent an elapsed time, measured in years, or the length of the branches may represent number of molecular changes (e.g. mutations) that have taken place between the two nodes. This is calculated is from the degree of differences when sequences are compared (refer to “alignments” later)
- Sometimes, the lengths are irrelevant and the tree represents only the order of evolution. [In a dendrogram, only the lengths of horizontal (or vertical, as the case may be) branches count].
- Finally the tree may be rooted or unrooted.

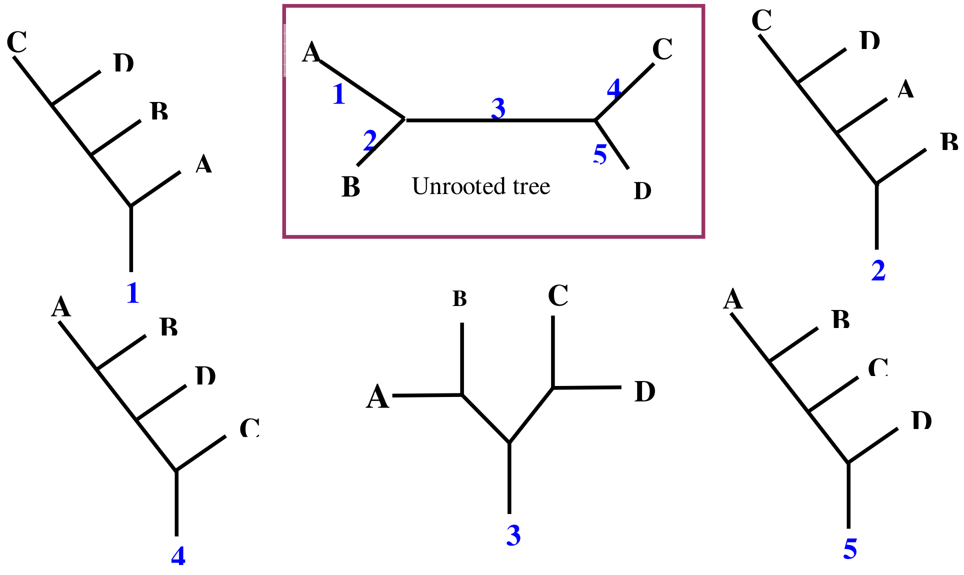
Unrooted trees

An unrooted tree simply represents phylogenetic but doesnot provide an evolutionary path. In **an unrooted tree**, an external node represents a contemporary organism. Internal nodes represent common ancestors of some of the external nodes. In this case, the tree shows the relationship between organisms A, B, C & D and does not tell us anything about the series of evolutionary events that led to these genes (see figure above). There is also no way to tell whether or not a given internal node is a common ancestor of any 2 external nodes.

Rooted trees

Gene trees are not the same as species trees

In case of a **rooted tree**, one of the internal nodes is used as an outgroup, and, in essence, becomes the common ancestor of all the other external nodes. The outgroup therefore enables the root of a tree to be located and the correct evolutionary pathway to be identified. In the above case, five different evolutionary pathways are possible using an outgroup, each depicted by a different **rooted** tree.


C
D
B
A
1
A
1
2
3
4
5
C
B
Unrooted tree
D
C
D
A
B
2
A
B
D
C
4
B
C
A
D
3
A
B
C
D
5