Frequency distribution:

In statistics, a **frequency distribution** is a list, table or graph that displays the frequency of various outcomes in a sample.[1] Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample

Application:

Managing and Operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables.

Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions. This assessment involves measures of central tendency or averages, such as the mean and median, and measures of variability or statistical dispersion, such as the standard deviation or variance.

What is a Mutually Exclusive Event?

Mutually exclusive events are things that can't happen at the same time. For example, you can't run backwards and forwards at the same time. The events "running forward" and "running backwards" are mutually exclusive. Tossing a coin can also give you this type of event. You can't toss a coin and get both a heads *and* tails. So "tossing a heads" and "tossing a tails" are mutually exclusive. Some more examples are: your ability to pay your rent if you don't get paid, or watching TV if you don't have a TV

Binomial

A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has **two possible outcomes** (the prefix "bi" means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Binomial distributions must also meet the following three criteria:

1. **The number of observations or trials is fixed.** In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense—I f you toss a coin once, your probability of getting a tails is 50%. If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.
2. **Each observation or trial is** independent. In other words, none of your trials have an effect on the probability of the next trial.
3. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to anotheR

RULES OF PROBABILITY:

- RULES OF PROBABILITY: The **complement** of an event is the event not occurring. The probability that Event A will not occur is denoted by P(A').
- The probability that Events A and B *both* occur is the probability of the **intersection** of A and B. The probability of the intersection of Events A and B is denoted by P(A ∩ B). If Events A and B are mutually exclusive, P(A ∩ B) = 0.
- The probability that Events A or B occur is the probability of the **union** of A and B. The probability of the union of Events A and B is denoted by P(A ∪ B) .
- If the occurrence of Event A changes the probability of Event B, then Events A and B are **dependent**. On the other hand, if the occurrence of Event A does not change the probability of Event B, then Events A and B are**independent**.

- **DEFINITION of 'Conditional Probability'**

- Conditional probability is the likelihood of an event or outcome occurring based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

  Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation (also called Pearson's *R*) is a **correlation coefficient**commonly used in linear regression.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related

Hypothesis Testing

Hypothesis Testing

The main purpose of statistics is to test a hypothesis. For example, you might run an experiment and find that a certain drug is effective at treating headaches. But if you can't repeat

that experiment, no one will take your results seriously. A good example of this was the cold fusion discovery, which petered into obscurity because no one was able to duplicate the results.

What is a Hypothesis?

A hypothesis is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation. For example:

- A new medicine you think might work.
- A way of teaching you think might be better.
- A possible location of new species.
- A fairer way to administer standardized tests

if you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
"If I…(do tiis to an independent variable)….then (this will happen to the dependent variable).

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

phylogenetics software /tool

is a compilation of computational phylogenetics software used to produce phylogenetic trees. Such tools are commonly used in comparative genomics, cladistics, and bioinformatics. Methods for estimating phylogenies include neighbor-joining, maximum parsimony (also simply referred to as parsimony), UPGMA, Bayesian phylogenetic inference, maximum likelihood and distance matrix methods.
Phylogeny.fr - is a simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.It includes multiple alignment (MUSCLE, T-Coffee, ClustalW, ProbCons), phylogeny (PhyML, MrBayes, TNT, BioNJ), tree viewer (Drawgram, Drawtree, ATV) and utility programs

An attribution **model** is a set of rules which is used to determine how credit for conversions should be attributed to different marketing channels. Through**Model Comparison tool** you can **compare** different default and custom attribution **models** to each other.

What is Multivariate Analysis?

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle. This type of analysis is almost always performed with software (i.e. SPSS or SAS), as working with even the smallest of data sets can be overwhelming by hand. Multivariate analysis can reduce the likelihood of Type I errors. Sometimes, univariate analysis is preferred as multivariate techniques can result in difficulty interpreting the results of the test. For example, group differences on a linear combination of dependent

variables in MANOVA can be unclear. In addition, multivariate analysis is usually unsuitable for small sets of data.
There are more than 20 different ways to perform multivariate analysis. Which one you choose depends upon the type of data you have and what your goals are. For example, if you have a single data set you have several choices:

- **Additive trees, multidimensional scaling, cluster analysis** are appropriate for when the rows and columns in your data table represent the same units *and* the measure is either a similarity or a distance.
- **Principal component analysis (PCA)** decomposes a data table with correlated measures into a new set of uncorrelated measures.
- **Correspondence analysis** is similar to PCA. However, it applies to contingency tables.

Cluster analysis

[110] aims at classifying a set of observations into two or more mutually exclusive *unknown* groups based on combinations of variables. Thus, cluster analysis is usually presented in the context of *unsupervised* classification [111]. It can be applied to a wide range of biological study cases, such as microarray, sequence and phylogenetic analysis [112]. The purpose of clustering is to group different objects together by observing common properties of elements in a system. In biological networks, this can help identify similar biological entities, like proteins that are homologous in different organisms or that belong to the same complex and genes that are co-expressed
It is generally difficult to predict behavior and properties based on observations of behaviors or properties of other elements in the same system, therefore various approaches for cluster analysis emerge. Clustering algorithms may be *Exclusive, Overlapping, Hierarchical* or *Probabilistic*.

Graph theory in bioinfo:

A graph is specified by the set of nodes (the term *vertex* is also sometimes used) V and the set of edges E. Each element of E contains a pair u, v of elements of V. It is allowed that $u = v$, in which case one also speaks of a *self-loop*. The relationships modeled by the edges may be dichotomous (the edge is there or it is not there) or we may consider a more general interpretation of E as a two-place function $f : V \times V \rightarrow F$ with discrete or continuous range F. If $F \subset \mathbb{R}$, then the value $f(u, v)$ is called the weight of the edge from u to v. F can also extend over different discrete categories, for example, a graph with genes as nodes can simultaneously model the homology between genes and their co-citation in the medical literature.

Graphs play roles in three complementary areas. First, graphs provide a data structure for *knowledge representation*. Examples include regulatory, signal transduction, or metabolic networks that are represented in graph form. This might be either in the informal way of the familiar bubbles and arrows cartoons of molecular biology text books, or more formally in knowledge databases such as Reactome [3]. Graphs are also used for knowledge representation in the Gene Ontology (GO) [4], and bipartite graphs between biological

concepts and scientific papers that are written about them [5] are another form of knowledge representation.

A further role for graphs is in *statistical modeling*. For example, one might want to fit a model that describes which sets of proteins can assemble together to form a protein complex, given some data consisting of (usually imperfect and incomplete) observations of pairwise interactions or of the co-precipitation of proteins [6]. Different models might apply to fit the data, and the usual questions of model fitting and discrimination and of hypothesis testing arise.

Probabilistic modeling: Structural bioinformatics is concerned with the molecular structure of biomacromolecules on a genomic scale, using computational methods. Classic problems in structural bioinformatics include the prediction of protein and RNA structure from sequence, the design of artificial proteins or enzymes, and the automated analysis and comparison of biomacromolecules in atomic detail. The determination of macromolecular structure from experimental data (for example coming from nuclear magnetic resonance, X-ray crystallography or small angle X-ray scattering) has close ties with the field of structural bioinformatics. Recently, probabilistic models and machine learning methods based on Bayesian principles are providing efficient and rigorous solutions to challenging problems that were long regarded as intractable.

a **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a randomphenomenon in terms of the probabilities of events. For instance, if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair). Examples of random phenomena can include the results of an experiment or survey.

A **variable** is a quantity whose value changes.

A **discrete variable** is a variable whose value is obtained by counting.

Examples: number of students present
number of red marbles in a jar
number of heads when flipping three coins
students' grade level

A **continuous variable** is a variable whose value is obtained by measuring.

Examples: height of students in class
weight of students in class
time it takes to get to school

distance traveled between classes

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

- A random variable is denoted with a capital letter
- The probability distribution of a random variable *X* tells what the possible values of *X* are and how probabilities are assigned to those values
- A random variable can be discrete or continuous

A **discrete random variable** *X* has a countable number of possible values.

A **continuous random variable** *X* takes all values in a given interval of numbers.

- The probability distribution of a continuous random variable is shown by a **density curve**.
- The probability that X is between an interval of numbers is the area under the density curve between the interval endpoints
- The probability that a **continuous random variable** *X* is exactly equal to a number is zero

Multiple regression:

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when started smoking, smoker type, income and gender.

Fuzzy logic and fuzzy technology are now frequently used in bioinformatics. The following are some examples.

Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean **logic** on which the modern computer is based. The idea of **fuzzy logic** was first advanced by Dr. Lotfi Zadeh of the University of California at Berkeley in the 1960s.

1. To increase the flexibility of protein motifs [43].
2. To study differences between polynucleotides [44].
3. To analyze experimental expression data [45] using fuzzy adaptive resonance theory.
4. To align sequences based on a fuzzy recast of a dynamic programming algorithm [46].
5. DNA sequencing using genetic fuzzy systems [47].
6. To cluster genes from microarray data [48].

To analyze gene expression data
To classify amino acid sequences into different superfamilies
To analyze the relationships between genes and decipher a genetic network

Range of correlation : The Pearson correlation coefficient, r, can take a range of values from **+1 to -1**. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

range: