International Journal of Algebra, Vol. 8, 2014, no. 3, 75 - 113
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ija.2014.312129

Introducing Soft Computing Statistical Measures: Very Useful & Significant Measures by Algebraic Approach

Ranjit Biswas

Department of Computer Science
Jamia Hamdard University
Hamdard Nagar, New Delhi–110062, India

Abstract

In the subject Statistics, the term "universe" refers to the total of the items or units in any field of enquiry, whereas the term "Population" refers to the total of items about which information is desired by a statistician (or by the statisticians). A population may not be just a collection of real number data, in general. Consequently, we define population into two categories: **R-Population** and **NR-Population**. If a population is of real number data only, it is of category 'R-Population', and if a population does not fall into the category of 'R-Population' then it is of the category 'NR-Population'. Thus a 'NR-Population' could be a collection of any type of data, viz. collection of sounds from a bus horn, collection of a large number of handwritten characters of the English character "A", collection of 100 paints of beautiful 'Tajmahal' by 100 number of under-9 children, etc. The main concern addressed in this paper is that the classical measures of statistics, in particular the three fundamental measures : mean (arithmetic mean), median and mode are undefined measures for NR-population data. But there are a lot of analysis, survey, estimations, conclusions, etc. being carried out by the analysts for various important objectives over such type of NR-populations in real life. The failure of these classical measures to play any role over NR-populations has been a hidden truth so far to the statisticians or analysts. In this paper the author introduces a new direction for formulating a number of new statistical measures for both kind of population (not

just restricted to the R-population). It is observed that the fundamental type of classical traditional measures happen to be special cases of our proposed measures. The author also introduces a new type of soft statistical measure called by "Nucleus" which carries a good amount of important information about the population. It is of a new type because of the fact that there is no similar type of statistical measure exists in the current literatures in statistics, although it is a very important measure. The nucleus is neither mean, nor median nor mode. However, in one sense it could be viewed as a kind of fuzzy mode of a population and so it is a kind of soft measure. The notion of nucleus will be very helpful to the statisticians in their way of analysis whenever they are interested to know about the congestion of data at or around various locations of the population. To study and analyze the congestion of data in a population, a number of new measures called by "density", "coefficient of homogeneity", "coefficient of heterogeneity", are introduced. Various characterizations have been done of the existing notion of the term 'population'. Various theoretical characterizations are also done of the multisets and of bags as they are applied in our work here. It is claimed by the author that the notion of nucleus will play a fundamental important role (as exhibited equivalently by classical mean, median or mode) in many areas of engineering, technology, statistics, mathematics, management science, medical science, social science, etc. to list a few only. It is sure that this work will widen the present universe of working-domains of the statisticians, analysts and the scientists as well as will open a new gate for huge and rigorous research in future. The notion of 'plot-function' for a population introduced in this paper is a new philosophical concept to be viewed with philosophical eyes rather than mathematical eyes. Once data be plotted as points, the congestion of data can be visualized with the same philosophy. A further direction is opened here by defining few new type of population mean called by linear mean (LM), region mean (RM) and consequently the measures like linear standard deviation (LSD), linear variance (LV), region standard deviation (RSD), and region variance (RV) are defined.

Mathematics Subject Classification: 62A86, 62-07, 62A99, 11J83, 30L99

Keywords: R-populatiion, NR-population, m-mapping, m_l-mapping, plot function, metric centre (MC), metric mean (MM), metric variance (MV), metric standard deviation (MSD), nucleus, desert point, mean deviation about MM, linear mean (LM), linear variance (LV), linear standard deviation (LSD), coefficient of homogeneity and heterogeneity, region mean, region variance.

1 Introduction

Statistics is a vast subject in Science, part of our everyday life at every moment. By statistic we mean a body of techniques and procedures dealing with the

collection, organization, analysis, interpretation, and presentation of information. Without the use of statistical methods it would be very difficult to make any good decisions if these are based on data. From statistical point of view, the term "Universe" refers to the totality of the items or units (or data elements) in any field of enquiry/survey, whereas the term "Population" refers to the total of items about which information is desired by a statistician (or by the statisticians). The attributes that are the objects of study are referred to as characteristics and the units possessing them are called as elementary units. The aggregate of such units is in general described as 'Population'. Thus all the units in any field of enquiry constitute 'Universe' and all elementary units (on the basis of one or more number of characteristics) constitute 'Population'. Thus, in statistics, by population we mean a large collection of objects of a similar nature which is of interest as a whole - e.g. human beings, households, readings from a measurement device, etc. Whenever we talk about a population P, we can also think of a relevant universe U as explained above. Thus all the members (with repetitions of them respectively) of a population P are also the members of the super object U. But the notion of 'Sample' in statistics is little different. A sample is a subcollection of objects drawn from a population. A sample is chosen to make inferences about the population just by examining or measuring the elements in the sample. Many times the outcome of a situation is difficult, maybe impossible, to predict. All we can say is that there is some range or distribution of possible outcomes. The word dispersion has a technical meaning in statistics. Some parameters attempt to describe the amount of variation between data. For example, consider a population of four data {5, 5 ,5, 5}. Here, each of the values are equal, so there is no variation. Philosophically we say that they are located at one location, not dispersed. The collection {3, 5, 5, 7}, on the other hand, has some variation since some values are different. We can say that they are dispersed as data reside at different location if plotted on the number line. The 'mean' philosophically measures the center of the data. It is one aspect of observations or analysis of data. Another feature of the observations is to know how the data are spread around the center. The data may be close to the center or they may be spread away from the center. If the data are close to the center (usually the arithmetic mean or median), we say that dispersion or scatter or variation of data is small. If the data are spread away from the center, we say dispersion is large. Two distributions having the same mean may differ in spread. *Unless otherwise stated, by the term 'mean' we will understand the 'arithmetic mean' althrough in our present work.*

Suppose we have two groups of students who have obtained the following marks in a class test :

Group-A : 47, 48, 50, 52, 53

Group-B : 20, 40, 50, 60, 80

In both the groups A and B, the means are equal (= 50). But in group A the observations are concentrated around the center. All students of group A have almost the same level of performance. We say that there is consistence in the observations in group-A. Whereas in group-B the observations are not close to the

center. One observation is as small as 20 and one observation is as large as 80. Thus there is greater amount of dispersion in group-B compared to group-A. The study of dispersion is very important in statistical data. If in a certain factory there is consistence in the wages of workers, the workers will be satisfied. But if some workers have high wages and some have low wages, there will be unrest among the low paid workers and they might go on strikes and arrange demonstrations. If in a certain country some people are very poor and some are very high rich, we say there is economic disparity. It means that dispersion is large. The idea of dispersion is important in the study of wages of workers, prices of commodities, standard of living of different people, distribution of wealth, distribution of land among framers and various other fields of life, etc. to list a few only out of infinite.

Some brief definitions of dispersion are:

(i) The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.

(ii) Dispersion or variation may be defined as a statistics signifying the extent of the scatter of items around a measure of central tendency.

(iii) Dispersion or variation is the measurement of the scatter of the size of the items of a series about the average.

Thus, measures of dispersion express quantitatively the degree of variation or dispersion of values in a population (or in a sample). Along with measures of central tendency, measures of dispersion are widely used in practice as descriptive statistics. The two most commonly used measures of dispersion are the range and the standard deviation. Rather than showing how data are similar, they show how data differs (i.e. shows the variation, spread, or dispersion). The range of a population is the simplest measure of spread or dispersion: It is equal to the difference between the largest and the smallest values. The range can be a useful measure of spread because it is so easily understood. However, it is very sensitive to extreme scores since it is based on only two values. The range should almost never be used as the only measure of spread, but can be informative if used as a supplement to other measures of spread such as the standard deviation or semi-interquartile range. The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation. The semi-interquartile range is also a good measure of spread or dispersion. It is computed as one half the difference between the 75th percentile [often called (Q3)] and the 25th percentile (Q1). The formula for semi-interquartile range is therefore: (Q3-Q1)/2. Since half the scores in a distribution lie between Q3 and Q1, the semi-interquartile range is 1/2 the distance needed to cover 1/2 the scores. In a symmetric distribution, an interval stretching from one semi-interquartile range below the median to one semi-interquartile above the median will contain 1/2 of the scores. The semi-interquartile range is little affected by extreme scores, so it is a good measure of spread for skewed distributions. In Statistics there are three basic measures of central tendency which are Mean (AM), Median and

Mode of a population data. These measures are the most primary measures as these reflect a first abstract about the population data without going into further depth. Philosophically, the mean measures the location of the 'center' of the population, the standard deviation measures its "radius". It can be shown that if X has a Gaussian distribution, 68% of the examples will be within one standard deviation of the mean, and 95% will be within two standard deviations of mean.
In this paper, the author introduces a number of new statistical measures of information about a population which are of very basic in nature. The existing measures of central tendency, of dispersion etc. happen to be particular types of our newly proposed measures. For many population data of real life, many of the existing statistical measures are either undefined or non-applicable. This long standing hidden problem is overcome in this work by generalizing the existing and traditional philosophy. It is certain that the various kinds of new measures introduced in this work will play a greater role in statistical analysis, data analysis, population survey, and research works. In fact, by introducing a number of soft and rigid statistical measures the author makes a major expansion of the existing universe of all the domains of populations in statistical analysis; the statisticians will now be able to visit and explore a much bigger domain than the traditional for their various interests. Allthrough our discussion in this paper, we will consider only finite populations.

2 Preliminaries

In this section, first of all we recollect some basic preliminaries of the Yager's Theory of Bags. In classical set theory, two sets A and B are said to be equal if for any $x \in A$ we have $x \in B$, and for any $x \in B$ we have $x \in A$. In this case the ordering of elements or repetition of elements is redundant. But in a statistical population, although the data are well defined objects nevertheless in this collection the redundancy counts, and hence a population can not be viewed as a set in general. For example, collection of books in the Central Library, collection of zeros of an algebraic polynomial, collection of names of students in a class, etc. are not sets in general. Such kind of collections are called multisets. Yager in [30] introduced a very simple and interesting mathematical structure to view the multisets. He called them by a new term 'bags', which facilitates the analysis in better ways, in a much more comfortable ways although both multisets and bags are same objects.

Definition 2.1
Let X be a finite set of elements. A bag B drawn from the set X is characterized by a function given by $C : X \to N$ where N is the set of all non-negative integers.
The function C is called the 'count function' of the bag B. For any $x \in X$, the value C(x) indicates the number of occurrence of the object x in the bag B.

The bag B can also be represented using the following notation

$$B = \{ x / C(x) : x \in X \}.$$

In our present work in this paper, we consider only finite populations, not infinite populations.

Suppose that a population P is given by $P = \{ p_1, p_2, p_3, p_4, ..., p_N \}$ which is a multiset. View P in the form of a bag (Yager's bag). If there is no confusion, let us accept that by P we denote here both the bag P and the multiset P. Let B be the set of which P is a bag (multiset) with the count function C(x), where $x \in B$. The value C(x) reflects the multiplicity (repetition factor) of the element x in the multiset P and is called the count-value of x in the bag P.

The cardinality of the multiset P

= cardinality of the equivalent bag B

$= \sum C(x), \quad x \in B.$

A multiset may be finite or infinite. The "Null Multiset" and "Null Set" are of same concept. For example, $P = \{ 2, 3, 2, 5, 5, 2 \}$ is a multiset, not a bag of Yager's definition. Cardinality of this multiset P is 6. The object $Q = \{ 2/3, 3/1, 5/2 \}$ is a bag of cardinality 6 drawn from the set $X = \{ 2, 3, 5 \}$. $P = \{ 2/3, 3/1, 5/2 \}$ is a bag drawn from the set $B = \{ 2, 3, 5 \}$, not from any other set (for instance, P is a bag, but not drawn from the set $Z = \{ 2, 3, 5, 8 \}$). The collection of all roots of the equation $(x-4)^3 (x-2) (x-1)^5 = 0$ forms a bag P drawn from the set $Z = \{ 2, 4, 1 \}$ having the cardinality 9. This collection can also be well taken as the multiset $P = \{ 4, 4, 4, 2, 1, 1, 1, 1, 1 \}$ of cardinality 9. Obviously, a set is a special case of a bag in which the count function and the characteristic function are to be viewed as same. The application of bags have been found in many areas, viz. relational database systems, mathematical analysis, decision sciences, etc. to list a few only. Thus a multiset is a collection of well defined objects in which objects may occur repeated times, where a bag is a "set associated with a count function" and is called by the phrase "a bag B drawn from a set X". For various operations and properties of multisets and bags one could see [3, 4, 9, 16-21, 26-28].

In our course of the present work, we require few definitions and results from Mathematical Analysis [10,12] too. The notion of metric space is an important basic topic in mathematical analysis. For details about the metric space, one could follow any standard book on mathematical analysis, viz. [10,12,15] to list a few only out of many excellent books. Nevertheless, let me present below few definitions from Mathematical Analysis [10,12] for ready reference to the readers.

Definition 2.2

A **metric** on a non-null set X is a function (called the distance function or simply **distance**)

$$d : X \times X \to \mathbf{R}^*$$

satisfying the following properties $\forall$ $x, y, z \in X$ **:-**

(i) $d(x, y) \geq 0$

(ii) $d(x, y) = 0$ if and only if $x = y$

(iii) $d(x, y) = d(y, x)$
(iv) $d(x, z) \leq d(x, y) + d(y, z)$

where R^* is the set of non-negative real numbers.

And then X is said to form a metric space with respect to the metric d. This metric space is denoted by the notation (X,d). The definition of a "metric" captures the most important and basic elements of what a "distance" should be. The distance from x to y should be the same as that from y to x; different points should be at positive distance from one another, but a point should be at distance 0 from itself, and traveling between two points via an arbitrary third point should not be shorter than the distance between the original two.

Definition 2.3 Closed Metric Ball & Open Metric Ball

Let (X,d) be a metric space, and r be a positive real number. A closed metric ball of radius r centered at $x \in X$ is the collection of all elements y of X such that $d(x,y) \leq r$. Any closed ball of X is a subset of it. An open metric ball of radius r centered at $x \in X$ is the collection of all elements y of X such that $d(x,y) < r$. Clearly, an open ball of X is a subset of the corresponding closed ball.

In the subsequent sections we introduce a number of new kind of statistical measures of fundamental type which are having a different kind of significance from that of the existing statistical measures. The new measures proposed in this work are soft measures and will surely play a different kind of roles, in fact a major role sometimes, in statistical analysis/survey in a new directions and in a new dimensions. For this, we introduce here few notions on population.

3 Characterization of the Object 'Population'

In the subject "Statistics", the term 'Population' is probably the most uttered term. Let R be the set of real numbers. A point x in the space R^n can be represented as $x = (x_1, x_2, x_3, \ldots\ldots, x_n)$, where $x_i \in R \ \forall\, i = 1, 2, 3, \ldots, n$. Let us characterize all the populations into two categories : R-Population and NR-Population.

Definition 3.1 R-Population and NR-Population

A population is called a **R-Population** if it is a collection of points from R^n for some finite positive integer n. If a population is not a R-population, it is called a **NR-Population**. Through our discussion in this paper, by a population we shall mean that it could be R or NR both, unless otherwise stated or specified.

In this section and in Section-5, we make some characterizations of the concept of 'Population' in a new direction which are required in our subsequent work in this paper. Let U be a universal set, may be finite or infinite. Here U is called the 'Universe of discourse' or 'universe' simply. Suppose that U forms a metric space with respect to the metric d**.** Consider a finite population P of size N in this universe U. *In fact while we consider a population P, we do not have, in general, a readymade source of information or knowledge about its universe U.*

However, we can, by intuition, enclose a projected hypothetical collection of same species of all data to form the universe U depending upon the objective of our interest.

For example, if we consider a population P of the diseases of all 40 patients of an ward in a hospital, then we can assume U to be the collection of all type of human diseases. If we consider a population P of marks scored by the 50 students of Class-X of Calcutta St.Xaviers School, then we can assume U to be the closed interval [0,100] or even the set R of real numbers. It is needless to mention here that U is a set, while P is a multiset or bag (P need not be a set in general). If there is no confusion, let us accept that by the notation P we shall denote here the multiset P as well as the bag P of the population P. This is quite obvious that for a given population P, the universe U conceptualized in this way may not be unique in many situations. We next introduce few basic terminologies.

Definition 3.2 Core Set of a Multiset (or of a Population)

Let B be the set of which the population P is one of its bags. Then the set B is called the core set of the population P. Clearly the core set is a subset of the universe U of P. For example, the core set of the population P = { 2, 3, 2, 5, 5, 2 } is set B = { 2, 3, 5 }.

Definition 3.3 "Sub-Population" of a Population

We introduce the definition of "sub-population", which is very simple. In general, a population P is not a collection which can be viewed as a set. It is a collection of data which forms a multiset, in general. Any sub-multiset S of the multiset P is called a sub-population of the population P.

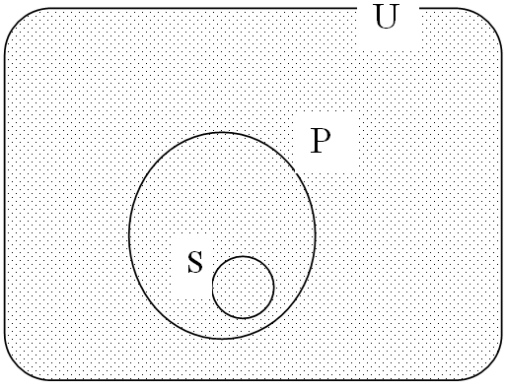

Figure 1. Universe U, a population P and a sub-population S of P

If we view a population P as a bag, then a sub-population will be a sub-bag of this bag. Thus the count value C(x) of any element x in a sub-population will be less or equal to the count value of the same in the population. Obviously, sub-population and sample are not the same concept. Any sample of a population P may be mathematically observed as a sub-population of P, but the converse is not true. A sub-population may be coined from a population mathematically, but a sample is drawn from a population statistically. In other words, a sub-population

may not be a population in general, because a population is a collection which is complete and sound with respect to some interest of the statistician concerned, where a sub-population may not fulfill the same interest to him.

We know that a population is not a set in general, but a multiset (i.e. bag). In classical mathematical analysis, metric spaces are defined over sets, not over multisets or bags. For our next course of work here dealing with statistical populations, we need to extend the definition of classical metric space to the theory of multisets by defining an almost similar mathematical object called by "Multiset Space", which is to be called by "Population Space" for statistical population as a particular case since a population is a multiset. The notion of "Multiset Space" or "Population Space" will be more useful and relevant to the subject 'Statistical Analysis' than the subject 'Mathematical Analysis'.

Definition 3.4 Multiset Space & Population Space

Consider a population P in the universe U, or rather say a multiset P in U. Suppose that U forms a metric space with respect to the metric d**.** If there is no confusion, let us accept that by P we denote here both the multiset P and the bag P. Then the multiset or bag P is said to form a **"multiset space"** in U with respect to the metric d, and is denoted by (P, d). For a statistical population P, the multiset space (P, d) is to be called by **"Population space".** If the cardinality of P is finite, then the population space (P, d) is called a 'finite population space'; and if the cardinality of P is infinite, then the population space (P, d) is called an 'infinite population space'. If P happens to be a set, then the population space (P, d) becomes a classical metric space. If P is not a set, then clearly the population space (P, d) can not be called a metric space. We present below two examples of population space.

Example 3.1

Consider a collection of 50 towns in India which are of some business importance to an industry "ABC INDUSTRY". For this population P of 50 towns, let us choose the universe U as the collection of all towns in India. With relevance to the business point of view, the ABC INDUSTRY, for its various strategies and profit-analysis, considers a metric d in U given by the following description :

d(x,y) = shortest available road distance between the cities x and y through which vehicles can ply (assuming that every road is both way).

Then P forms a population space with respect to the metric d. We denote this population space by (P, d).

Example 3.2

Consider the metric space (R, d) where R is set of real numbers and d is the metric defined by $d(x,y) = |x-y|$ where $x, y \in R$.

Consider a population P given by

$$P = \{ 45, 67, 34, 45, 8, 12, 45, 45, 8, 21, 8, 8, 12 \}.$$

Then P forms a population space with respect to the metric d.

Obviously, if S be a sub-population of the population P and if (P,d) be a popula-

tion space then (S,d) is also a population space. The converse is also true.
The complete work in this paper is based on the initial basic assumptions that :-

(i) the population (be it R-population or NR-population) must form a population space with respect to a highly relevant metric d defined over the universe U. The metric d is to be choosen by the concerned statistician by his best intellectual capability so that the main interest for the pursuance for good results/analysis/ conclusion is expected to be met.

(ii) The population must be finite.

4 Introducing a New Statistical Mean : MM (or MC)

In the subject Statistics, while we talk about 'dispersion' or 'spread' of a population data, about the behavior of 'central tendency' etc., we generally conceptualize them in the ***philosophy of differences (or distances)*** of data from the other data or from the centre measure (say, mean). In many of the cases the population data are real numbers (or of numeric type like scattered dot points on XY-plane). For example, 'Height' of students of class-X of a school, 'Weight' of the patients of 1000 sugar patients in a locality, 'Marks' in mathematic scored by the second-year students of an engineering college in 2009, 'Temperature' recorded for a day at Calcutta Meteorological Department at an interval of every 20 minutes, etc are statistical variates which takes data from the universe U = R, the set of real numbers. For calculating the various measures of central tendency, various measures of dispersion etc. of a population, we need to coin the basic elements like **:** the 'distances' of x_is from x_j s, where x_i, x_j are population data, the 'distances' of x_is from mean, etc. These distances are usually a type of absolute distances or Euclidean RMS distances. But in many real life situations, the population data are not real numbers or of numeric type. They could be images, sounds, graphs, pictures, shapes, noise, etc., to list a few only out of infinite which form NR-populations.

For example, consider the following module of an interesting Project (not any official project, but purely posed by few students at their own curiosity and interest) being carried out by a team as mentioned below :-

Recently a team of student-researchers of Indian Statistical Institute, Calcutta has planned to collect about 20,000 number of handwritten characters of the English character "A" from 20,000 distinct students of class-X from different states of India (in fact this team of Calcutta wants to do the same for all the English characters A to Z, a to z, and also for all the decimal digits 0 to 9, in India and then in China, to finally make some important comparisons in the style of hand-writing; and similarly to do the same in many other common attributes of habits/practices, and finally to make correlation with physiological parameters, how this correlation differs between Indian and Chinese nationals).

The initial purpose of this Calcutta Research team is :-

(i) *to visualize the 'mean' of 20,000 handwritten characters of each English character in India and China independently; how does the mean look like?*

(ii) *To estimate the 'standard deviation' (and 'variance') of 20,000 handwritten characters for each of the English characters.*

(iii) *To compute many other measures time to time according to their works need.*

But there is a basic problem being faced by this research team. It is because of the fact that the mathematical formula (for the classical measures of central tendency or for the classical measures of dispersion) can not be useful to them as these formula are not applicable to the population data for any NR-population. For example, to compute the classical mean one has to calculate the value of the expression $\frac{1}{n}\sum_{1}^{n} x_i$, i.e. of the expression $\frac{1}{n}(x_1 + x_2 + x_3 + \ldots\ldots + x_n)$, but the classical operation "+" used here may not be a valid operation in the multiset data of NR-populations, unlike R-populations. The same is true for the operation of multiplication by the real number $\frac{1}{n}$.

This is a major failure or drawback of the classical statistical measures. In our subsequent work, in particular in Section-4 and Section-10, we overcome this failure by introducing and mathematically modeling few broad kind of statistical measures applicable to wide types of statistical populations, in particular to NR-populations too.

4.1 Metric Centre (MC) or Metric Mean (MM) of a Population

In this section, the author introduces a new kind of mean which will play a role analogous to that of the classical mean but having a stronger potential for application to a bigger domain of the nature or types of populations.

Definition 4.1.2 Metric Centre (MC) or Metric Mean (MM)

Let P be a finite population in a universe U. Let the cardinality of the core set B of the population P is n. Suppose that if P is viewed as a bag then C(x) is the count function where $x \in B$. Now consider an object m of the universe U which is *closest* to all the objects of P compared to all other objects of U; i.e. which is at the *centre* of P. This statement seems to be not precise because of the presence of two imprecise terms/words "*closest*" and "*centre*", but the philosophy carried by this statement can be expressed in another way as below :-

Consider the object m ($\in$ U) such that

(i) it <u>minimizes</u> the expression E given by
$E = \sum d(m, x_i)$, where $x_i \in$ multiset P
(i.e. the expression $E = \sum C(x_i).\, d(m,x_i)$,
where $x_i \in B$), and

(ii) for which the standard deviation of the collection of values $d(m,x_i)$ is minimum, where $x_i \in$ multiset P. (Here the collection of values $d(m,x_i)$ clearly forms a multiset of real numbers).

Then the object m is called the **"d-Metric Centre"** or **"d-Metric Mean"** of the population P. The term d-Metric Centre can be called in short by the term 'Metric Centre' (MC), and similarly the term d-Metric Mean can be called in short by the term 'Metric Mean' (MM).

How to solve the above cited minimization problem is not a part of our work here. It is rested upon the computer programmers mathematicians or who can pose the problem in a good mathematical way beautifully into an optimization problem and can solve it purely on case to case basis either analytically or using appropriate software. It is needless to mention here that the MM value m of a population need not be a real number or of numeric type. It's datatype or characteristic properties will be similar to those of the objects (data) of the population, be it a R-population or a NR-population. Besides that it is also true that although the MM is a member of U, it may or may not be a member of the population P.

Example 4.1.1

The classical 'arithmetic mean' (AM) of a finite population P out of the universe R (the set of real numbers) is a simple and very common example of d-Metric Mean where d is the metric given by $d(x,y) = |x-y|$ for every $x, y \in R$.

The above two conditions in the Definition 4.1.2 of MM (or, MC) are independent. This can be understood by the following two mathematical examples with hypothetical data :-

Example 4.1.2

Consider a hypothetical population data $P = \{7, -7\}$ consisting of two numbers only, where the universal set U is R (the set of real numbers). Clearly $\forall$ m $\in [-7.7]$ the condition (i) is satisfied, whereas both the conditions (i) and (ii) are satisfied by $m = 0$ only. (Considering the metric d given by $d(x,y) = |x-y|$ for every $x, y \in U$).

Example 4.1.3

Consider a population P which is a collection of finite number (say, 100) of 3-D points (x,y,z) lying on a given circular ring in space. Here U is the set of all 3-D points in space.

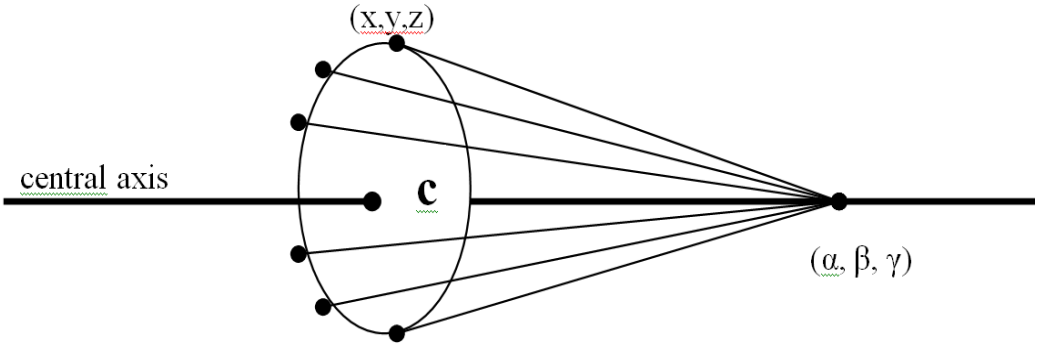

Figure 2. A population P of 100 points (x,y,z), MM being at C

Clearly, $\forall$ m = (α, β, γ) lying on the axis of the ring the condition (ii) is

satisfied, whereas both the conditions (i) and (ii) are satisfied by the object m = 'centre' point C of the ring. (Considering the metric d in U given by the Euclidian Distance in 3-D geometry).

A statistician may choose the classical Mahalanobis Distance, Eucledian Distance, Bhattacharya Distance, Hamming Distance, etc. for d on the basis of suitability and requirements. The choice of a suitable metric d is to be done by the concerned analysts looking at the type of the data objects of the population (R or NR). No absolute method can be suggested on how to choose an appropriate metric d. In case the population is a R-population and the data are real numbers, an obvious choice of a metric is the metric d of Example 4.1.1.

A universe U of a population P (be it R or NR) could be a collection of cities, or a collection of company franchise, or a collection of real numbers, a collection of complex numbers, or a collection of hospitals, or a collection of sounds from a single source, etc. For many such types of population the classical measure "mean" happens to be an 'undefined measure' or an 'absurd measure' because of failure of the definitive mathematical formula for "mean", and hence it does not exist (or, rather say that there does not exist any question for its existence). But the notion of 'MM' proposed here is an interesting concept generalizing the traditional measure "mean" because MM of a population will always exist if (U, d) forms a metric space with respect to a suitably chosen metric d. The following example will explain the significance of MM for a given situation.

Example 4.1.4

Consider the "ABC Petroleum Company" which has signed a contract to supply ten vehicles of petrol to ten cities respectively, one vehicle (full tank petrol of one vehicle) to each city, everyday from its 'Head Quarters' whose location is to be decided. Now the question is **:** "Where should this company's 'Head Quarters' be set up in order to minimize the total transportation cost?". This is a problem of optimization, but a correct answer is that its Head Quarters must be located at the location "MM" of the population P of ten cities, where MM may be computed with respect to the metric "Euclidian Distance". (For solving this problem, one can apply Cartesian Coordinate geometry. The population P will be the collection of the coordinates of the cities assuming that earth surface is plane, where the universe U is the set $R \times R$, R being the set of real numbers. The solution m will be a unique (x,y) coordinate on the XY-plane).

The notion of MM introduced above is a new kind of statistical measure of central tendency over the whole population. It is closest to the complete population considered collectively. Clearly it is a major generalization of the existing concept of classical measure 'mean' of central tendency. For different choices of the metric d, the value of MM of a population will be different, in general. The choice of the metric rests upon the statistician (or Data-Analyst) who will choose it depending upon his interest of work and analysis.

5 Further Characterizations of Population

In Section-4 above, we have made some characterizations of the notion of 'population'. In this section we make further characterizations of population in a new direction, which will be useful for any statistician or analyst while dealing with population whose data could be of any kind, not just real numbers only, i.e. for NR-population too.

Let B be the core set of the population P, and that the cardinality of B is n. Suppose that if we view the population P as a bag, the corresponding count function is C(x), where $x \in B$. We next define the terms 'diameter' of a population, and 'density' of a population, which can be regarded two useful informatic soft measures about any population.

Definition 5.1 Diameter of a Population

Let (P,d) be a population space. The diameter of the population P is the non-negative real number D given by $D = \max \{ d(x,y) \}$, where $x, y \in P$.

Clearly, if all the objects of a population are identical then $D = 0$.

The diameter of a sub-population can be defined similarly. Obviously, the diameter of a sub-population can not exceed the diameter of the parent population.

Example 5.1

The classical statistical measure "Range" of a population P of real numbers data is a simple example of the notion of diameter of a population w.r.t the metric d of Example 5.1.1.

Definition 5.2 Density of a Population

Let P be a population of cardinality N and diameter $D (\neq 0)$. The density ρ of the population P is defined by the real number given by $\rho = \frac{N}{D}$.

In case $D = 0$, let us assume that mathematically the population has an infinite density. The following proposition is straightforward.

Proposition 5.1

Consider n number of populations P_i $(i = 1, 2, 3, \ldots\ldots, n)$ in a common universal metric space (U, d) having dimensions ρ_i respectively and diameters D_i respectively for $i = 1, 2, 3, \ldots\ldots, n$. Consider the merged population P of all the populations P_i. Let ρ be the density and D be the diameter of P.

Then the following results are true :-

(i) $D \leq \sum D_i$

(ii) $\rho \geq \sum \rho_i D_i / \sum D_i$

Example 5.2

Consider the population space (P, d) with respect to the metric d given by $d(x,y) = |x-y|$ where $x, y \in U (= R)$, and the population P is { 5, 9, 21, 8, 5, 5, 42, 5, 5, 5, 9 }.

Clearly, a little computation will show that the diameter (P) = 37, and the

density (P) = $\frac{11}{37} \approx .2973$ which is the overall density of the population.

Although the diameter of a sub-population can not exceed the diameter of the parent population, but density of a sub-population may (or may not) exceed the density of the parent population. The notion of density of a population should not be confused with the notion of correlation. Its significance is to be viewed philosophically, and is straightforward like the concept of density of a liquid, density of a solid mass, etc. as in the subjects Physics and Chemistry. The concept of 'density' of a population defined above is the overall density of the population. Sometimes, a statistician may be curious to know the density of the population "at some locality" of it or "at around a point" of it. To introduce this concept we need to define the term 'population ball'.

Definition 5.3 Closed Population Ball

Let (P, d) be a population space, and r be a non-negative real number. A closed population ball of radius r centered at $x \in P$ is the collection of all elements y of P such that $d(x,y) \leq r$. Such a closed population ball is denoted by B[x,r] which is a multiset, in general. Clearly, a closed population ball B[x,r] of a population P is a sub-population of it, and can never be a null multiset even if r = 0. In fact the cardinality of B[x,0] will be at least one (and it will be more, if the count value $C(x) > 1$).

Definition 5.4 Open Population Ball

Let (P, d) be a population space, and r be a non-negative real number. An open population ball of radius r centered at $x \in P$ is the collection of all elements y of P such that $d(x,y) < r$. Such an open population ball is denoted by B(x,r) which is a multiset in general. Clearly, an open population ball B(x,r) of a population P is a sub-population of it.

For r = 0, the closed population ball encloses only one distinct element with all its repetition (i.e. with all its occurrences), but the open ball is a null multiset. It means that the core set of B[x,0] is a singleton set $\forall x \in P$, but B(x,0) is always a null multiset.

If $r_1 < r_2$, then $B(x, r_1) \subseteq B(x, r_2)$ and $B[x, r_1] \subseteq B[x, r_2]$ $\forall x \in P$.

Also $B[x,D] = P$ $\forall x \in P$, where D is the diameter of the population P.

In Definition 5.2 above, we have defined the overall density of a population P. We now define the notion of density of a population at different locations of it. The density of a population P at some point x of it is defined with respect to the close neighborhood of the point x. It is defined by the amount of congestion of data or by the amount of crowd around the point x.

Definition 5.5 r-density of a Population at some point

Consider the population space (P, d). For any positive real number r, the r-density d_r of the population P at a point x of it is denoted by $d_r(x)$ and is defined by $d_r(x) = \frac{n}{2r}$

where, n is the cardinality of the closed ball B[x,r].

The term 'r-density' may be simply called in short by the term 'density'.

Note : The members of a closed ball B[x,r] constitute a sub-population Z of the population P. It does not necessarily mean that the diameter d of the sub-population Z will be equal to 2r, the diameter of the ball. However, it is obvious that $d \leq 2r$.
In the next section we introduce a new and highly significant, useful important soft measure in Statistics which is called by 'nucleus' of a population.

6 Nucleus : An Important Statistical Measure

The notion of nucleus of a population is a kind of soft measure in statistics. It is neither MM nor the classical mean; It is also neither the classical median nor the classical mode, but it hints at a very significant information about the population. The kind of information available by the proposed soft measure 'nucleus' about a population is not available by any kind of existing classical statistical measures. Consequently, this soft measure will make a major expansion of the existing universe of the working domains of the statisticians in statistical analysis, decision making, in drawing conclusion, etc. Information about the congestion of data at some location of the population P (R or NR) sometimes is very important and useful to the concerned analysts. We might be interested to know which area of a very large land of rubber trees producing every year more rubber (with more density), and to analyze the reasons behind such results. We might be interested to know which locations of a very big city have large number of educated people, which area of a sector receives more rainfall.
Nucleus of a population (R or NR) means a member of the population centered around which there exist a large number of members of the population. In another terminology we say that a nucleus is such a member of the population the neighborhood of which is crowded (i.e. it has a dense population around). The notion of nucleus is related to a local congestion of data and hence related to a local density of population. In a population there may not exist any nucleus. Otherwise, a population may have one nucleus or many nuclei. Actually the number of nuclei of a population depends upon the way an analyst or a statistician or an intelligent agent seeks the population to be analyzed. The nucleus is neither the classical mean nor median nor mode in nature, but it can be viewed as one kind of fuzzy mode of the population. By crisp mode of a population we mean a population data which occurs most frequently in the population P. If m is mode in the multiset P, then the corresponding functional value of the count function C(x) for x = m is greatest. But it is to be noted that the crisp mode of a population may or may not be a nucleus. One may argue that a crisp mode can be viewed as a special case of fuzzy mode and hence it trivially qualifies to become a nucleus. But it is not so. Even, it may happen that while the crisp mode of a population is not a nucleus, but a data nearest or very near to it (in the sense of the metric considered) happens to become a nucleus. We present an example below to understand the significance of this

new statistical measure 'nucleus'.

Example 6.1

Suppose that a teacher teaches Mathematics at Xth level in a school at Calcutta, the students' strength in his class being as high as 190. Last month he conducts a class-test of total marks 100 where pass mark is 35; and then he completes the evaluation. It is observed that the highest mark scored is 100 out of 100, mean of the marks scored by the students is 60, median is 61, mode is 72, and the range is 99. These all are traditional kind of information which are of rigid nature. But the teacher has also given to his Principal a new kind of important soft statistical information **:** "about 50 students of the class have scored 40 or close to 40". The teacher observes that a good number of data is centered around 40. In another way, it is observed that a large number of data are "data(40) centric data". What is the name of this new statistical measure? It is neither the classical mean, nor median nor mode. It is certainly a new type of measure not existing in the literature of Statistics. Such kind of information about a population is very common in many real life events and will be useful for drawing better conclusion, better decision and for planning for better future actions. For extracting such kind of information, we introduce the notion of 'nucleus' of a population, a new kind of soft statistical measure, which could also be viewed as one kind of 'fuzzy mode'.

Definition 6.1 Nucleus

A nucleus of a population is such a member of the population centered around which there exist a congestion of large number of members of the population. Thus we are considering the case of a large number of "data centric data", where the first word 'data' is singular (which is the nucleus defined here) and the second word 'data' is plural.

Since there exists a large number of data closely spread or dispersed around the nucleus, therefore all these data are 'almost equal' or 'almost identical' to the nucleus in some sense of real importance. We can therefore very rightly view a nucleus as a kind of "fuzzy mode" of the population. Fuzzy mode of a population as conceptualized here is not fuzzy, but it is a crisp member of the population such that there exist a large number of members in P which are almost identical or very close to it.

The above definition of nucleus seems to be not precise mainly because of the presence of the following phrase/hedge used :-

(i) 'centered around which' (ii) 'large'

which are vague.

The phrase 'centered around which' is related to a 'population ball' with a center and a radius. But, the obvious questions arise : How much is the radius? How to choose an appropriate value of the radius? The hedge 'large' is also fuzzy, but it is obvious in this context that 'large' is a positive integral number. **Thus nucleus is not a hard measure, but a kind of soft measure.**

But we notice that the above definition of nucleus is complete provided that for computing the nucleus (or nuclei) of a population, the concerned statistician

does choose and fix-up the values of 'radius' and 'large' by applying his best intellectual judgment relevant to his interest and purpose.

Definition 6.2 Dimension of a Nucleus
To compute a nucleus, if exists, of a given population (R or NR), a statistician pre-chooses the values **:** radius = r and large = N which are non-negative real numbers. Then the pair (r, N) is called a dimension of the nucleus.

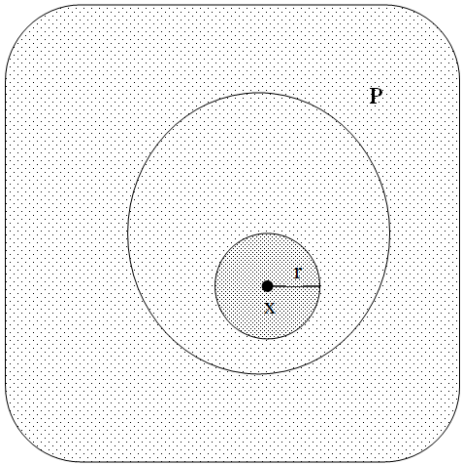

Figure 3. A nucleus x of a population P with dimension (r, N)

For a given dimension there may exist nil or one or many nuclei of a population. Also, for more than one distinct dimensions, the nucleus (or nuclei) of a given population may be common.

Definition 6.3 Nucleus, Multi-nucleus, and Nuclei Multiset
If count value C(x) of a nucleus x in the population bag P is 1, it is a simple **nucleus.** In case the count value of x is more than 1, then x is called a **multi-nucleus**. A multi-nucleus is also a single element, like a simple nucleus. But a multi-nucleus is such a nucleus whose count value in the bag P is more than 1. In this sense, we assume that a 'multi-nucleus' is a singular term, where the term 'multi-nuclei' is the corresponding plural term. For different choice of the dimension (r, N), the multi-nuclei and the number of multi-nuclei may be different in a given population. The multiset of all the multi-nuclei for a given dimension (r, N) is known as **Nuclei Multiset**, and is denoted by the notation M(r, N). Clearly, $M(r, N) \subseteq P$ (in the sense of sub-multiset)**.** If $N > \#(P)$, then $M(r, N) = \Phi$, but the converse is not necessarily true.

The following propositions are obvious.

Proposition 6.1
If $N = 0$ or 1, then every data of the population is a nucleus of it.
i.e. $M(r, 0) = M(r, 1) = P$.

Proposition 6.2
If $N_1 > N_2$ then $M(r, N_1) \subseteq M(r, N_2)$.

Proposition 6.3
If x is a nuclei with dimension (r, N), then x is so with dimension (ρ , M) $\forall\ \rho \geq r$ and $\forall\ M \leq N$.

Proposition 6.4
Let x_1 be a nuclei with dimension (r_1, N_1) and x_2 is a nuclei with dimension (r_2, N_2) for a population P. Then x_1 and x_2 both are nuclei with dimension (r, n), where

$$r \geq \{ d (x_1, x_2) + \max \{ r_1, r_2 \} \} \quad \text{and} \quad n \leq N_1 + N_2.$$

Proposition 6.5
If x is a nuclei with dimension (r, N), then the r-density of the population P at the point x is at least $\frac{N}{2r}$.

Proposition 6.6
If x be the MM of a population P having the diameter D, and r be the minimum value of the radius such that B[x,r] = P, then $r \geq \frac{D}{2}$.

In a scattered diagram of a large number of points (x,y) in a XY-Plane, we fit a straight line using least square method, which we call 'best fit straight line'. In our notion of nucleus, it apparently seems to us that we see that we fit a point around a congestion of a large amount of points (data). There may exist nil or one or more number of such points which can be fit over and around the population, depending upon the scatter of data points. A nucleus is inhabitant in a locality (sub-population) of the population P, be it a R or NR. The importance of the notion of nucleus (nuclei) is not hidden to our daily life. Nuclei carry a good amount of information to the concerned analyst about the population data. Nuclei may exist in finite as well as infinite population both. A nucleus along with its members forms a sub-population of the population. With its members if considered locally, it can play great role as an useful and informatic local measure of central tendency for that particular region of the population. Statistical analysts dealing with a population may become sometimes curious to know whether there is a nucleus in the population, and if 'yes' then how many nuclei are there in the population. The information regarding "existence of a nucleus" or "non-existence of a nucleus" in a population gives an instant message to the analyst.

The notion of 'nucleus' should not be confused with the notion of 'cluster' initiated by Tryon [25] in 1939. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait, often a kind of proximity according to some defined distance measure.

The following two examples will reflect the significance of a nucleus.

Example 6.2

The scores of a cricket batsman (of his 20 years cricket career) is recorded for the 310 matches he has played so far. His performance was continuously poor around his 40^{th} match for about 10 to 12 matches, and also around his 220^{th} match for about 8 to 10 matches. But his career poorest performance happened at his 157^{th} match.

Example 6.3

In an Atmospheric Research Centre at Calcutta, during one day of snowfall, a research team records the amount of snowfall at "nodal times" only. Examples of nodal time according to that research team are : "Node 6 O'clock", "Node 7 O'clock", "Node 14 O'clock", "Node 21 O'clock", etc. By a nodal time "Node X O'clock", they mean the time interval [X – **.**5 hour, X + **.**5 hour]. For instance, the nodal time "Node 23 O'clock" means the time interval from 10.30 P.M. to 11.30 P.M. If the snowfall at a nodal time "X O'clock" is f unit of amount (w.r.t. certain unit followed by them), they record this data in the form of (x,y) coordinate style, like the point (X,f). The distance between two such data points (X_1,f_1) and (X_2,f_2) is measured by them using Euclidean metric.

On a certain day of snowfall, this meteorological research team recorded the following data of 24 hours starting from 6 A.M. of the day till 5 A.M. of the next day :-

Serial No.	Nodal Point	Amount of snowfall (in some unit)
1.	6 O'clock	9
2.	7 O'clock	5
3.	8 O'clock	2
4.	9 O'clock	**30**
5.	10 O'clock	**32**
6.	11 O'clock	**34**
7.	12 O'clock	**35**
8.	13 O'clock	**33**
9.	14 O'clock	7
10.	15 O'clock	8
11.	16 O'clock	1
12.	17 O'clock	0
13.	18 O'clock	1
14.	19 O'clock	1
15.	20 O'clock	0
16.	21 O'clock	3
17.	22 O'clock	**39**
18.	23 O'clock	3
19.	24 O'clock	2
20.	1 O'clock	4
21.	2 O'clock	2
22.	3 O'clock	10
23.	4 O'clock	8
24.	5 O'clock	15

The population P considered here is the collection of data points (Xi, f_i) where i = 1, 2, 3,…….., 23, 24. The statisticians of the Atmospheric Research Centre, Calcutta observed that there is a continuously heavy snowfall around 11 O'clock in the morning, where the mode (peak) of the data stands at around 10 P.M. The congestion of data is observed at around 11 A.M. and also at around 6 P.M. w.r.t. certain appropriate dimension pre-choosen by the team.

Definition 6.4 Core Nucleus

Consider a population P (R or NR) and its core set B. For a given dimension, a nucleus of the population P may or may not qualify to become a nucleus of the set B, if B is independently viewed as a hypothetical population. But any nucleus of B will obviously be a nucleus of P. A nucleus of a population P which is also a nucleus for B is called a 'core nucleus' of the population P. A population P having one or many nuclei may or may not have a core nucleus. It is obvious that the collection of all core nuclei of a population P will be a subset of the nuclei-multiset of P. At the point of a core nucleus of a population, it is expected to have very strong data-congestion and hence to have usually a high density.

7 Homogeneously & Heterogeneously Scattered Population

A population (R or NR) may have different density at different points of it. If the density of the population at each of its points are same or almost same, then we say that the population data are scattered homogeneously. Otherwise we say that the population data are scattered heterogeneously. For example, consider a collection Γ of three data (each being a point on a 2-D XY-plane) which form a equilateral triangle if connected by lines. Consider another collection Π of three data (each being a point on the same XY-plane) which form a triangle of internal angles 10° degree, 87° and 83°, if connected by lines. It is obvious that the above population Γ is homogeneously scattered, where the population Π is heterogeneously scattered. **In general, we can view any population to be "homogeneously scattered" and "heterogeneously scattered" both.** In extreme cases we can say that if homogeneity of data-spread is 100% in amount in a population then the heterogeneity will be 0% in amount and vice-versa.

But the question is "How to measure the amount of homogeneity and the amount of heterogeneity?". This measure could be a very important information to the statisticians or data-analysts or researchers in many real life cases for making decisions or conclusions or future plans. To initiate a study of the notion of homogeneity or heterogeneity, we first of all define a new kind of multiset of a finite population P, which forms a population space (P,d) in the universe U.

Definition 7.1 "Distance Multiset" of a finite Population

Consider a finite population space (P,d) where P is a multiset of cardinality n (say). For every pair of elements x_i and x_j of P, we have a distance $d(x_i, x_j)$. If x_i and x_j are same then $d(x_i, x_j) = 0$, otherwise $d(x_i, x_j) > 0$.

The collection of all such $d(x_i, x_j)$ of the elements of the population P forms a multiset D_P which is called the 'Distance Multiset' of the population P. The corresponding count function is denoted by C_{DP}. Obviously, Cardinality $(D_P) = {}^nC_2$, where n is the cardinality of P.

Proposition 7.1

If $D_P = D_Q$, it does not necessarily mean that P = Q.

Proposition 7.2

Max $C_{DP} \geq$ Max C_P, where C_P is the count function of the bag P and C_{DP} is the count function of the bag D_P, viewing the multisets P and D_P as bags.

Definition 7.2 Coefficient of Heterogeneity & Coefficient of Homogeneity

Let P be a finite population, of which D_P is the distance multiset. Let σ be the standard deviation (classical s.d.) of all the data elements of D_P. Then the **"Coefficient of Heterogeneity"** of the population P is denoted by He(P) and is defined by

$$\text{He(P)} = \frac{\sigma}{Max\{d(xi, xj)\}} = \frac{\sigma}{D},$$

where D is the diameter of the population P.

If the population data are real numbers, then the diameter D is nothing but the Range of the population (with respect to the usual metric d(x,y) = $|x-y|$ on the set of real numbers), and in that case we have

$$\text{He(P)} = \frac{\sigma}{Range}$$

The **"Coefficient of Homogeneity"** of the population P is denoted by Ho(P) and is defined by

$$\text{Ho(P)} = 1 - \frac{\sigma}{Max\{d(xi, xj)\}} = 1 - \frac{\sigma}{D}.$$

The above two coefficients satisfy the following conditions :-

(i) $0 \leq \text{He(P)} \leq 1$.

(ii) $0 \leq \text{Ho(P)} \leq 1$.

(iii) $\text{He(P)} + \text{Ho(P)} = 1$.

If He(P) > Ho(P), it signifies that the spread of data in the population P is more heterogeneous than homogeneous. Similarly, if Ho(P) > He(P), it signifies that the spread of data in the population P is more homogeneous than heterogeneous.

Example 7.1

Consider the populations Γ and Π as considered above at the beginning of Section-8. We see that D_Γ is a multiset of cardinality 3 consisting of one element repeated three times, and D_Π is a set of three elements. It can be calculated that $\text{He}(\Gamma) = 0$ and $\text{Ho}(\Gamma) = 1$. It means that the data in the population Γ is 100% homogeneously spread, the amount of heterogeneity being nil. By little computation, the values of $\text{He}(\Pi)$ and $\text{Ho}(\Pi)$ can also be calculated, and it can be observed that the following

results are obviously satisfied by the population Π :-

(i) $0 < He(\Pi) < 1$

(ii) $0 < Ho(\Pi) < 1$

(iii) $He(\Pi) + Ho(\Pi) = 1$.

While the data in the population Γ is 100% homogeneously spread, but the data in the population Π is not so. In Π, the data is homogeneously as well as heterogeneously spread too, none being nil.

Definition 7.3 Desert Point of a Population

A point x of a population P is called a desert point of dimension (r, M) if the cardinality of the multiset B[x,r] is less than M.

A desert point for 'a large value of r and a small value of M' signifies a lot of information to the statistician.

Definition 7.4 Isolated Point of a Population

A point x of a population P is called an isolated point of radius r (≥ 0) if the core set of the multiset B[x,r] is a singleton set.

For a given dimension (r, M), a population may or may not have any desert point. Similarly, for a given radius r, a population may or may not have any isolated point.

An isolated point, if exists, is also a desert point w.r.t. the same value of r, but the converse need not be true. It does not necessarily mean that for r = 0, every point of the population will be an isolated point, because there are many points whose count value is 2 or more.

Proposition 7.3

If x is a desert point of dimension (r, M) of a population P, then the following results are true :

(i) the point x is also so for the dimension (r, K) $\forall$ $K \geq M$.

(ii) the point x is also so for the dimension (s, M) $\forall$ $s \leq r$.

Proposition 7.4

If x is an isolated point with radius r_1, then x is also so with any radius r_2 ($\leq r_1$).

8 To Compute Nuclei of a Finite Population (R or NR)

In this section we propose two methods for computing all the nuclei of a finite population. The methods may be termed as :-

(i) Statistical Computing

(ii) Fuzzy Computing

It is not necessary that the above two methods will give the same results for a given population.

8.1 Statistical Computing of Nuclei of a Finite Population

For a given dimension (r_o, n_o), we will now calculate all the multi nuclei of a population. Suppose that the population P is given by the multiset (bag)

$$P = \{ p_1, p_2, p_3, p_4, ..., p_N \}.$$

Let X be the core set of the multiset P given by

$$X = \{ x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n \},$$

and let the count value of an element x in the bag P is $C_P(x)$, where $x \in X$.
Now consider the closed population balls $B_i = B[x_i , r_0]$ in the population P for $i = 1, 2, 3, 4, \ldots, n$ and fill-up the entries in the following table: -

Table 8.1.1

Closed Ball B_i	**Number (N_i) of population data lying inside the ball B_i**	**Is $N_i \geq n_0$?**
$B_1 = B[x_1 , r_0]$	N_1	Yes/No
$B_2 = B[x_2 , r_0]$	N_2	Yes/No
$B_3 = B[x_3 , r_0]$	N_3	Yes/No
……	……..	Yes/No
……	……..	…….
……	……..	……..
$B_n = B[x_n , r_0]$	N_n	Yes/No

If there is no "Yes" entry in the last column in the above table, it means that there exists no nucleus of this population corresponding to the input pair (r_o, n_o). Total number of "Yes" entry in this column is the total number of nuclei of the population.
Suppose that there is a "Yes" entry corresponding to the ball B_k. Then x_k is a nucleus. Thus there may exist zero, one or more number of nuclei of a population. With no loss of generality, instead of plotting the points in a hyperspace or in an abstract hyperspace, we plot them horizontally in Figure 4 to have a philosophical view to visualize the congestion of data as below :-

. . . . …….. . . . . . .
x_1 x_2 x_3 x_k x_n

Figure 4. A nucleus x_k of a population P

Obviously, the r_0-density of the population P at the point x_i will be the value $\frac{Ni}{2ro}$ $\forall$ i = 1, 2, 3,……,n. The standard deviation of data residing inside a ball centered around a nucleus will be low. To illustrate this method by an example below, let us consider a hypothetical R-population (i.e. with real number data points) as shown below, and compute its all nuclei.

Example 8.1.1
Consider a population consisting of the following data in a statistical survey :-

52.8, 9.3, 18.2, 6.12, 32.1, 7, 40.8, 5.25, 21.3, 6.18, 99, 19.9, 5.2, 82, 34.7, 6.2, 22.2, 70, 6.3, 6.14, 90, 15.6, 65.1, 6.1, 14.4, 1, 20, 82, 6.5, 100, 8.2, 17, 5.1, 82, 2, 60, 15.6, 6.13, 50.7, 3.4, 25, 6.15, 95, 12.4, 5.9, 29, 82.

A little statistical computation will yield that for this population the three fundamental classical measures are : mean = 30.999361, median = 17, mode = 82. Since the population data in this example are real numbers, the population balls are intervals here. Now we see that for a given dimension say (0.3, 10) there exists only one nucleus of the population which is 6.2. Thus there is a 'large' number of data centered around 6.2. In another terminology we say that the neighborhood of 6.2 is crowded because it has a dense population around, with **.**3-density equal to 16.666 (approx.). If we vary the dimension, the results are likely to vary.

8.2 Fuzzy Computing of Nuclei

Fuzzy theory [31,7] has been regarded as one of the most appropriate mathematical tools to deal with imprecise data or information.

Let us recollect the definition of nucleus. It is a data centered around which there exists a large amount of data. Thus there is a high density of population around a nucleus.

In fuzzy computing of nucleus, we pre-choose the value of "large" = L (say). The analyst (like a Statistician or any Decision Maker [7] or any intelligent agent) will choose a value for this parameter depending upon the plan of analysis to be carried out by him with the given population, but unlike Statistical Computing Technique we do not use here any pre-chosen value of the radius. Consequently an obvious question does arise : How to decide which data are centered around a given data and which are not? We see that this question automatically gets redressed in the fuzzy computing technique. First of all we define the term 'fuzzy dimension'.

Definition 8.2 Fuzzy Dimension of a Nucleus

In Definition 6.2 we defined dimension of a nucleus. In fuzzy computing we do not require the parameter 'radius', but instead we use a decision level $\alpha \in [0,1]$. The pair of real numbers (α, L) decides the nucleus in fuzzy computing. Let us call this pair (α, L) by **"fuzzy dimension"** of a nucleus. It may be noted that fuzzy dimension is not fuzzy but this is the dimension on the basis of which the fuzzy computing technique of nuclei is constructed.

We now present the fuzzy technique below :-

As mentioned before, we consider only a finite population. Suppose that the population P is given by the multiset $P = \{ p_1, p_2, p_3, p_4, \ldots\ldots, p_N \}$ whose core set is the set $B = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$. Consider P in the form of a bag with the count function $C_P(x)$, where $x \in B$. The value $C_P(x)$ reflects the multiplicity (repetition factor) of the element x in the multiset P and is called the count-value of x in the bag P. Clearly $\sum C_P(x)$

measures the cardinality of the multiset P (i.e. of the bag P) denoted by #(P).
Corresponding to every point x_i we consider a fuzzy set "about x_i" (which can be synonymously called by "closest to x_i" or "approximately identical to x_i"). Since the elements x_i are not real numbers in general, the closeness of two elements x_k and x_r are measured by the amount of distance between them. If the distance is less, then they are close. If the distance is more then they are away from each other, not close. The close elements are almost equal in some sense, or approximately identical. Let the notation A^{xi} denotes the fuzzy set "about x_i" of U. Construction of membership function for the fuzzy set A^{xi} is an important decision to be made by the concerned decision maker. For a suitable decision level α in [0,1], we then compute the α-cut of A^{xi}. Let this α-cut of A^{xi} be denoted by $A^{xi}(\alpha)$. Obviously, $A^{xi}(\alpha) \neq \Phi$. Now we apply the notion of 'fuzzy equal member' as defined by the following definition.

Definition 8.2.2 Fuzzy Equal Members of a population data
Consider the bag (multiset) M^{xi} given by

$M^{xi} = \{ x / C_P(x) : x \in B \cap A^{xi}(\alpha)$ and $C_P(x)$ = count of x in the bag P $\}$.

The members of the multiset M^{xi} are called **"fuzzy equal members of x_i"** in the population P . In another terminology the members of M^{xi} are also called to be the data 'centered around' the element x_i , as they are close neighbors of x_i.
Suppose that the pre-choosen value of "large" is L.
Thus there will be n number of multisets M^{xi}, one for each x_i where i = 1, 2, 3, …..,n. Calculate the cardinality of each of these multisets. Suppose that $\#(M^{xi}) = L_i$ for each i.
Now collect all the values of i for which $Li \geq L$. If for i = k this inequality is satisfied, then we say that x_k is a nucleus of the population. Clearly, there may exist more than one distinct nucleus for a given fuzzy dimension. The following definitions (as defined earlier in Definition 6.3) hold good for the case of fuzzy dimensions too.
If count value of x_k is 1, it is a simple nucleus. In case the count value of x_k is more than 1 in the bag P then x_k is called a **multi-nucleus**. A multi-nucleus is also a single element, like a simple nucleus. But a multi-nucleus is such a nucleus whose count value in the bag P is more than 1. In this sense, we assume that a 'multi-nucleus' is a singular term, where the term 'multi-nuclei' is the corresponding plural term. For different choice of the fuzzy dimensions (α, L), the multi-nuclei and the number of multi-nuclei in a population may be different. The multiset of all multi-nuclei for a given fuzzy dimension (α, L) is known as **Nuclei Multiset**, and is denoted by the notation M(α, L). Clearly, M(α, L) $\subseteq$ P. If L > #(P), then M(α, L) = Φ, but the converse is not necessarily true.

The following propositions are obvious.

Proposition 8.2.1
If L = 0 or 1, then every data of the population is a nucleus of it.
i.e. $M(\alpha, 0) = M(\alpha, 1) = P$.

Proposition 8.2.2
If $L_1 > L_2$, then $M(\alpha, L_1) \subseteq M(\alpha, L_2)$.

We present below an algorithm for this fuzzy computing technique of nuclei multiset of a population P.

Fuzzy Computing Algorithm

1. input the fuzzy dimension (α, L)
2. input the multiset P in the form of a bag P.
3. compute the core set $B = \{ x_1, x_2, x_3, \ldots\ldots, x_n \}$ of the bag P.
4. input the counts $c_i = C(x_i)$, where $x_i \in B$, i = 1, 2, 3, …., n.
5. for i = 1 to n do
6. compute A^{xi} corresponding to x_i.
7. compute $A^{xi}(\alpha)$.
8. compute M^{xi}.
9. compute $L_i = \#(M^{xi})$.
10. endfor
11. $M(\alpha, L) = \Phi$
12. for i = 1 to n do
13. if $L_i \geq L$, then $M(\alpha, L) = M(\alpha, L) \cup \{ x_i / c_i \}$ *: this is bag union*
14. endfor
15. return $M(\alpha, L)$
16. stop

We show by an example below the fuzzy computing technique of nuclei multiset of a population with hypothetical data. For the sake of presentation here, we consider a small size population.

Example 8.2.1
Consider a population consisting of the following data :

82, 51, 18, 17, 68, 16, 25, 17, 82, 16, 82.

A statistician needs to calculate all the multi-nuclei of this population for 'large' L = 5, an information which is required by him in making some sort of analysis and prediction. He considers the fuzzy dimension (.6, 5).
In this population, the multiset P is { 16, 16, 17, 17, 18, 25, 51, 68, 82, 82, 82 }, and the corresponding bag P is {16/2, 17/2, 18/1, 25/1, 51/1, 68/1, 82/3 }, for which the core set is B = { 16, 17, 18, 25, 51, 68, 82 }.
Suppose that he considers the notion of triangular fuzzy number in his computation, the notion which is revised in [5]. Recall that a triangular fuzzy number $\tilde{a} = (a_1, a_2, a_3)$ is represented by the membership function given by

$$\mu_{\tilde{a}}(x) = \begin{cases} 0 & \text{if} \quad x \leq a_1 \\ \dfrac{x-a_1}{a_2-a_1} & \text{if} \quad a_1 \leq x \leq a_2 \\ \dfrac{x-a_3}{a_2-a_3} & \text{if} \quad a_2 \leq x \leq a_3 \\ 0 & \text{if} \quad x \geq a_3 \end{cases}$$

Suppose that the statistician, by his best intellectual capability and judgment, considers the following A^{xi} s :

$A^{x1} = 1\tilde{6} = (13, 16, 19)$, $A^{x2} = 1\tilde{7} = (14, 17, 20)$,
$A^{x3} = 1\tilde{8} = (15, 18, 21)$, $A^{x4} = 2\tilde{5} = (22, 25, 28)$,
$A^{x5} = 5\tilde{1} = (48, 51, 54)$, $A^{x6} = 6\tilde{8} = (65, 68, 71)$,
$A^{x7} = 8\tilde{2} = (79, 82, 85)$.

Therefore, we have the following membership functions :-

$$\mu 1\tilde{6}(x) = \begin{cases} 0 & \text{if} \quad x \leq 13 \\ \dfrac{x-13}{3} & \text{if} \quad 13 \leq x \leq 16 \\ \dfrac{19-x}{3} & \text{if} \quad 16 \leq x \leq 19 \\ 0 & \text{if} \quad x \geq 19 \end{cases}$$

$$\mu 1\tilde{7}(x) = \begin{cases} 0 & \text{if} \quad x \leq 14 \\ \dfrac{x-14}{3} & \text{if} \quad 14 \leq x \leq 17 \\ \dfrac{20-x}{3} & \text{if} \quad 17 \leq x \leq 20 \\ 0 & \text{if} \quad x \geq 20 \end{cases}$$

$$\mu 1\tilde{8}(x) = \begin{cases} 0 & \text{if} \quad x \leq 15 \\ \dfrac{x-15}{3} & \text{if} \quad 15 \leq x \leq 18 \\ \dfrac{21-x}{3} & \text{if} \quad 18 \leq x \leq 21 \\ 0 & \text{if} \quad x \geq 21 \end{cases},$$

and so on.

The value of the decision level choosen here is $\alpha = .6$, which he thinks to be the

most appropriate level of initial presumption. Then it can be computed that :

M^{16} = { 16/2, 17/2 }, and therefore $L_1 = 4$.
M^{17} = { 16/2, 17/2, 18/1 }, and therefore $L_2 = 5$
M^{18} = { 17/2 , 18/1 }, and $L_3 = 3$
M^{25} = { 25/1 }, and $L_4 = 1$
M^{51} = { 51/1 }, and $L_5 = 1$
M^{68} = { 68/1 }, and $L_6 = 1$
M^{82} = { 82/3 }, and $L_7 = 3$.

Clearly, the condition $L_i \geq L$ is true for L_2 only. Therefore there exist only one nucleus of the population which is $x_2 = 17$. There is no other nucleus for this population. It may be noticed that $x_2 = 17$ is a multi-nucleus with count value 2.

Note : In this example, it could be noted that for 'large' $L = 6$ or for 'large' $L \geq 6$, the population does not have any nucleus. However, for $L = 3$ the population has four nuclei/multi-nuclei. Also, considering the serious drawback of the existing notion of fuzzy numbers as pointed out in details in [5], we shall prefer to use the redefined version of fuzzy numbers [5] here.

9 Plot Function

In statistical analysis, we come across various types of population (both R and NR). But for every population, all the data have a common trait. For example, a population may be a collection of numeric data like collection of marks scored by the students of class-X of Calcutta St. Xaviers School, or may be of non-numeric type like a collection of sounds from a car-horn, a collection of diseases of the patients of an ward of a hospital, a collection of paints by 50 artists on a common scenery, a collection of 20 cities in India, etc. Our interest is "how to plot" a population data in the form of "points" on a suitable chosen "base". A collection of real number data can be plotted as points on a number line (recording the respective count value for each point), where the suitably choosen base is the "number line". A collection of points like (x,y) where x, y are real numbers can be plotted on a XY-plane (recording the respective count value for each point), where the suitably choosen base is the "Cartesian XY-plane". A collection of points like (x,y,z) where x, y, z are real numbers can be plotted on XYZ-space (recording the respective count value for each point), where the suitably choosen base is the "3-D XYZ-space". A collection of points like $(x_1,x_2,\ldots..,x_n)$ where $x_1, x_2,\ldots...., x_n$ are real numbers can be abstractly plotted on $X_1X_2X_3\ldots.X_n$ – hyperspace (recording the respective count value for each point), where the suitably choosen base is the "n-dimensional hyperspace". But the general issue is that we can not plot an arbitrary population in the form of points. More precisely speaking, we do not have in our present knowledge any method to plot an arbitrary population as points, and the equally tough issue is that given a population what could be an

appropriate "base" on which we can effort to put our imagination for plotting the population data as points. For instance, with our present knowledge or intellectual ability, we can not plot a collection of diseases or a collection of paints or a collection of sounds abstractly in a so straightforward manner on a suitable base, real or hypothetical, as a collection of points.

If we can not plot a population data as dots on some real or hypothetical or abstract base, then we have a lot of limitations and paralyzed situation, viz. :-

(i) we can not think of visualizing the 'congestion' of data at different locations,
(ii) we can not think of estimating the 'density' of population overall,
(iii) we can not think of estimating the 'density' of population at different locations of it,
(iv) we can not think of visualizing the crowded regions of the population,
(v) and consequently, we can not think of computing any measure of dispersion (spread), any measure of central tendency, etc. for such type of population.

We propose **: Let us apply a philosophical thought over this issue, instead of applying pure mathematics.**

In this section we will introduce the notion of plot function of a population which is to be viewed with philosophical eyes. Before that we need to make some useful characterizations of multisets.

9.1 The Notion of 'Mapping'(Function) Over a Multiset Domain

First of all let us review the concept of "belongingness" in the theory of multiset. When an object is said to be in a multiset ? Let P be a multiset and B be its core set. An object x is said to belong to the multiset P denoted by the notation "$x \in P$" iff $x \in B$. For example, if $P = \{ 5, 9, 3, 9, 9, 3 \}$ then it is true that $9 \in P$, $5 \in P$ etc. If $Q = \{ 8, 3, 3, 4, 3, 3 \}$ then it is true that $3 \in P$, and also $3 \in Q$.

9.1.1 Introducing "m-mapping" and "m_1-mapping" of a Multiset

In classical sense, a mapping is defined from a set to another set. In this section we extend this concept by defining a notion of mapping from a multiset to a multiset, retaining the same nomenclature (terminology) 'mapping' or 'function'.

Let S be the core set of a multiset P. Consider a classical mapping $f : S \rightarrow T_1$ from the set S to a set T_1. Let T be the co-domain of the mapping f. Therefore, the mapping $f : S \rightarrow T$ is a mapping from domain to co-domain. Let us extend this notion of mapping from the multiset P to the multiset T_e as below (without changing the function name) $f : P \rightarrow T_e$ such that

(i) if $x \in P$ with count value C(x), then $f(x) \in T_e$ with a minimum count value C(x),
(ii) the cardinality of the multisets P and T_e are equal, and
(iii) the core set of T_e is T.

Such a mapping of a multiset (or, of a bag) to a multiset (or, to a bag) is

called by **m-mapping** (or, m-function). Here T_e may be termed as an extended multiset being the extension of the core set T. In fact, from the population P we compute the core set S, from the core set S we arrive at the set T (co-domain) via the mapping f, and lastly from the set T we get the multiset T_e by the simple extension as described above.
If the classical mapping $f : S \rightarrow T$ is 1-to-1, then the m-mapping $f : P \rightarrow T_e$ is called to be a **m_1-mapping** (or, m_1-function).

Example 9.1
Consider the multiset P = { 5, -2, 5, 7, -5, 3, -2, -5 }. The core set of P is given by S = { 5, -2, 7, -5, 3 }. Consider the mapping $f : S \rightarrow T_1$ given by $f(x) = x^2$ where T_1 = { 6, 25, 1, 4, 49, 9, 8 }. Clearly T = { 25, 4, 49, 9 }. Then, an example of m-mapping from a multiset to a multiset is given by $f : P \rightarrow T_e$, where P = { 5, -2, 5, 7, -5, 3, -2, -5 } and T_e = {25, 25, 25, 25, 4, 4, 49, 9 }.
However, this m-mapping f is not a m_1-mapping.

9.2 Plot Function of a Population (R or NR)
Plotting a R-population viz. plotting a collection P of 100 points (x,y) on a XY-plane is an easy task and thus we can directly visualize the congestion of data in P, if any. But in real life situation, datatype of the NR-population could be of various nature, and many of such NR-populations can not be plotted in the form of points even in an abstract (or imaginary) way. How to choose a suitable base for plotting? and how to plot population data as points on such a base? There is no absolute method for these tasks. But quite naturally the immediate question is "Why do we need to plot the population data as points on a suitable base?". One answer to this question is that our main interest is to visualize the congestion of data of the population whatever be the datatype of its member. We will deal this problem here with the help of philosophical eyes, without going into any serious debate on the basis of our knowledge in mathematics or science. It is quite natural that while dealing with a population P for some interest, there is no guarantee that the concerned statistician will be able to choose or model a suitable matric d_P over P by which a sense of distance or a sense of proximity can be conceptualized between two data elements in P to attain the main interest of the concerned statistician. It is not because of any lack of knowledge of the statistician always, but because of too much abstractness of the datatype of the population data. We will see that this failure will not be always an obstacle to our interest of studying and searching for the congestion of data in P, or to our interest of estimating the measures of congestion of data in P. Because, it may be sometimes easily possible to think of a suitable m_1-mapping f of the multiset P into a multiset T_e in which it could be easy to think of a metric d_T such that distance between two members of P can be philosophically estimated (or, scaled) by the distance between their f-values in T_e. Since we sometimes can not define a metric d_P in P or since we can not imagine of any idea about a kind of distance in P, we can not mathematically accept so blindly that

the m_1-mapping f is a distance preserving mapping ; but by a good intuitionistic choice of d_T we may become philosophically convinced that what we are doing is OK i.e. with philosophical views, we are expecting to have the following hypothetical equality with the overall situation :-

$$d_P(x,y) = d_T (f(x), f(y)) \quad \forall \; x, y \in P,$$

or to have the following hypothetical inequality with the overall situation :-

$$d_P(x,y) \leq d_T (f(x), f(y)) \quad \forall \; x, y \in P,$$

where the left hand distance may be untraceable or un-earthed in both the cases.

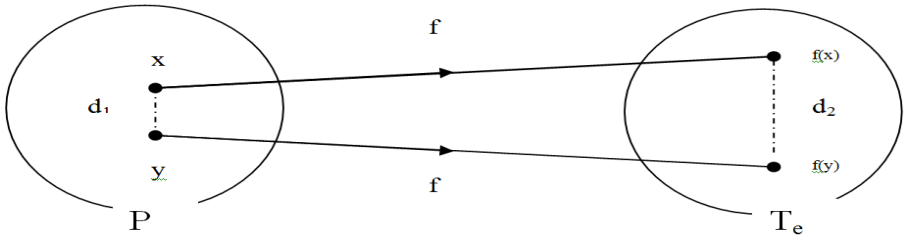

Figure 5. A case of m_1–mapping of P with $d_1 \leq d_2$.

It is needless to mention here that any arbitrary m_1-mapping f of P will not fulfill our purpose, because intuitionistically it may not be a distance-preserving (or, distance expanding) mapping which is our basic requirement. If the core set T of T_e be a subset of R^n for any $n \in N$ (the set of Natural numbers), then the abstract plotting of T_e in the hyperspace R^n is easy to visualize as there are a number of standard and good metric available in R^n. If the core set T of T_e be not a subset of R^n, then we must have a good metric d_T over T with us. Surely, with the above philosophy the congestion of data in the actual population P will be same or more than the congestion of data in the created population T_e. Such a m_1-mapping is called a **plot function** of a population. The introduction of plot function of a population is completely based upon philosophical platform. The following example shows an instance of philosophically acceptable metric over the population P.

Example 9.2

Consider a collection of 100 pieces of hand-written characters of the English alphabet "A". Suppose that these hand-written characters constitute the population P denoted by $P = \{ A_1, A_2, A_3, \ldots\ldots, A_{99}, A_{100} \}$.

Suppose that all these A_i be normalized on 1 cm x 1 cm size square frames putting the character at the centre of the frame like below:

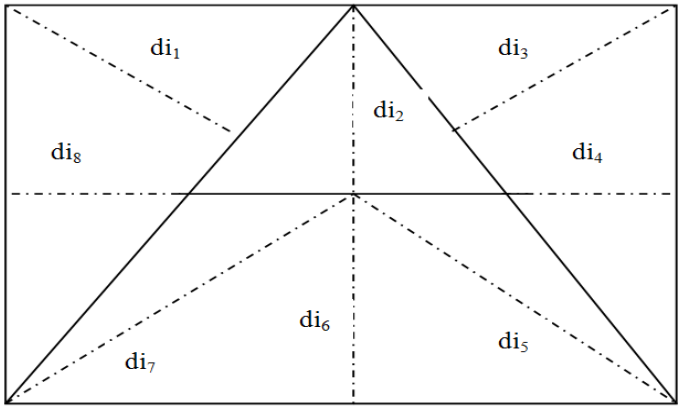

Figure 6. The character A placed on a 1 cm x 1 cm frame.

Consider the object A_i and the corresponding vector $V_i = \{d_{i1}, d_{i2}, d_{i3}, \ldots\ldots, d_{i8}\}$ $\forall$ i = 1, 2, 3, …., 99, 100 where d_{ij} are the eight distances as shown in the frame. Now consider the m-mapping $f : P \rightarrow T_e$ given by $f(A_i) = V_i$ where $T_e = \{ V_1, V_2, V_3, \ldots.., V_{99}, V_{100} \}$, and assume that $d_P (A_r, A_s) = d_T (V_r, V_s)$ $\forall$ r, s = 1, 2, 3, …., 99, 100

where $d_T (V_r, V_s) = 1 - \dfrac{\sum_{k=1}^{8} v_{rk} v_{sk}}{\sqrt{\sum_{k=1}^{8} v_{rk}^2} \cdot \sqrt{\sum_{k=1}^{8} v_{sk}^2}}$

Clearly, f is a m_1-mapping of P into T_e.

(Note : It may be understood that in case the above m-mapping f be not a m_1-mapping, we may increase the dimension of the vector V_i sufficiently high (instead of low number 8) so that f becomes so**).**

An arbitrary choosen m_1-mapping of the population multiset P into a multiset T_e may not be distance preserving (or, distance increasing). The following example will justify it.

Example 9.3

Consider the population P given by { 80, 6, 14, 80, 80, 9, 14, 23 } whose core set is S = { 80, 6, 14, 9, 23 }. Let us sort the elements of S in increasing order, and hence consider the rank of each of these five numbers (where 1 means lowest rank). Consider a 1-1 mapping f of S into the set T = {1,2,3,4,5} of ranks defined by f(x) = rank of x $\forall$ x $\in$ S.

From the set T, let us generate the multiset T_e = {1, 2, 3, 3, 4, 5, 5, 5}.

Now consider the population spaces (P,d) and (T_e , d), where d is the usual metric in the set R of real numbers given by $d(x,y) = |x-y|$ for every x ,y $\in$ R. and the m_1-mapping f from P to T_e (here eventually, the metric d is common to both P and T_e). Surely, congestion of data in P can not be visualized in this case by the congestion of data in T_e.

Thus an arbitrary m_1-mapping may dilute the scene of congestion of the domain data in its co-domain data, or may deepen too or may not be at all useful in any way.

Therefore the m_1-mapping of P is to be suitably choosen by the analyst by his best possible intuition and judgment so that the co-domain set reflects (increases) the congestion of data of the domain set. Very appropriately, if the population data happens to be from the universe $U = R^n$ where n is a natural number, then the m_1-mapping f from P to T_e should be nothing but the identity mapping (i.e. the self-mapping).

10. Statistical Measures : LM, LV, and LSD

Consider a R-population P of real number data. There are three fundamental means (or, centres) of P which are **:** Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM). Out of these three, AM does always exist, but GM or HM may or may not exist. Each of the above three type of means signifies a centre point of the population data in some sense. There is no absolute measure of the centre point of the population data. We have already discussed about the major failure of all these three kind of means while the population data are not real numbers, in particular for the NR-populations.

In this section, we introduce another kind of mean called by "Linear Mean (LM)" which could be sometimes applicable to locate the centre of a population even if the population be NR, and then introduce the corresponding notion of variance and standard deviation. Then we define **region mean (RM)** over a region [6]. For this a preliminary study about the 'Theory of Regions' [6] is recommended. In [6], it is unearthed that the elementary algebra (various rules, formulas, equalities, identities, solution methods, results, etc. which are being fluently used in Mathematics/Statistics) can not be fluently practiced in general in a group, ring, field, module, linear space, algebra over a field, associative algebra over a field, division algebra, or in any existing algebraic system. The minimum platform required for practicing elementary algebra is the region algebra [6]. In the same work [6], the 'Theory of Objects' has also been introduced and it has been justified and analysed that the 'Theory of Numbers' is a particular case of the topic 'Theory of Objects'. In this sense, RM will play a very generalized role in Statistics in the context of the properties of population or big data [8].

Definition 10.1 Linear Mean (LM)

Consider a finite population P of which the universe is the set U. Suppose that U forms a linear space (i.e., vector space) over the field R of real numbers with respect to the addition operator "$\oplus$" and scalar multiplication operator "$\otimes$".

Let the notation $\sum_{1}^{k} Ai$ denotes the object $(A_1 \oplus A_2 \oplus A_3 \oplus \ldots\ldots\ldots \oplus A_k)$ of U where $A_i \in U \quad \forall\, i = 1, 2, 3, \ldots\ldots, k.$

Let the population P is the multiset $\{X_1, X_2, X_3, X_4, \ldots\ldots, X_n\}$ where X_is need not be all distinct. Then the linear mean (LM) of the population P is defined by the object μ of U given by

$$\mu = \frac{1}{n} \otimes \sum_{1}^{n} Xi \,.$$

Example 10.1

The classical arithmetic mean for any population of real number data is an example of LM.

Example 10.2

If the population be a collection of m x n matrices whose elements are real numbers, then we can calculate the LM of the population using the addition operation $\oplus$ as "matrix addition" and the multiplication operation $\otimes$ as the "matrix scalar multiplication".

Example 10.3

Let U = C[0,1] be the set of all real valued continuous functions on the domain I = [0,1]. Define the addition operation $\oplus$ in U given by

$$(x \oplus y)(t) = x(t) + y(t) \quad \forall\, x, y \text{ of } U \quad \text{where } t \in [0,1],$$

and the scalar multiplication $\otimes$ in U be defined by

$$a \otimes x(t) = a \,.\, x(t) \qquad \forall\, a \in R.$$

If the population P be a finite collection of real valued continuous functions on the domain I = [0,1], then we can easily calculate the LM of P.

We observe that LM is a generalized concept of the concept of classical mean in the sense that the population data may/need not be real number data always. The definition of LM is simpler than that of MM because LM does not call for any optimization problem in its computation. LM can be directly computed. The same formula is true to define region mean (RM).

Our next thought is to define the immediate important basic measures like variance, and standard deviation with the notion of LM. The datatype of the population data could be any from the real world and LM will be no different from them. But whatever be the population, the kind of variance and standard deviation to be introduced below will always be real numbers only, no other type. For this we need to have that (P,d) forms a population space w.r.t a metric d suitably choosen by the concerned statistician by his best possible intuition and judgment.

Definition 10.2 Linear Variance (LV), Linear Standard Deviation (LSD), Region Variance (RV) and Region Standard Deviation (RSD)

Let (P, d) be a population space having its LM μ.

Then the 'linear variance (LV)' of the population P is defined by

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \{d(X_i, \mu)\}^2$$

and the 'linear standard deviation (LSD)' of P is defined by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{d(X_i,\mu)\}^2}$$

Obviously, whatever be the type of population, its LV and LSD are always real numbers but the LM μ will be of the analogous datatype of the population data. The analogous formula are true to define Region Variance (RV) and Region Standard Deviation (RSD) and hence not mentioned separately.

11 Conclusion

In this paper a number of new type of statistical measures are introduced, and consequently the existing universe of the domains of the statistical analysis have been expanded. Most of the work is to be viewed with philosophical eyes too. The populations considered are finite but not just limited to populations of real number data. Both R-populations and NR-populations are considered in this work. Infinite number real cases are there in our everyday life where the data constitute NR-population, not R-population. But in the existing huge volume of rich literature on Statistics do not have a serious attention on the data of NR-population. The major breakthrough done in this work is that most of the existing statistical measures (fundamental measures) which were undefined so far for NR-population have been now defined for NR-population too. A population is a collection of data which forms a multiset in general, and hence it can be viewed as a bag of Yager's model [30]. The theory of multisets have been extended mathematically and applied in the theory of Statistics. The notion of 'population' is characterized in few new directions, a number of new properties of populations are developed and modeled in mathematical ways to make them suitable for better statistical analysis. All Statistical measures are categorized in two groups : rigid measures and soft measures. A serious drawback and failure of the existing notion of mean is unearthed and a generalized notion of mean called by MM or MC has been introduced, having the potential for application in NR-population too.

There is not much study exists in the vast literature on the subject 'Statistics' about the congestion of data in a population (be it R-population or NR-population). In this work, the notion of congestion, and the notion of density of a population have been introduced which are in fact a kind of soft measures.

Another important new and highly useful kind of soft statistical measure called by 'nucleus' is introduced. Several examples and propositions are presented to understand the potential of these new measures. In fact, Nucleus is one kind of fuzzy mode. Nucleus is neither mean nor median nor mode. But it carries a significant information about the population (R or NR). In a population there may exist no nucleus, or only one nucleus or many nuclei. Two methods of nucleus-computing are presented with hypothetical examples. One is statistical Computing Technique and the other is Fuzzy Computing Technique. It is not necessary that the two methods will give the same results. Also, considering the serious drawback of the existing notion of fuzzy numbers as pointed out in [5] in

details with the help of many examples and with sufficient justification, we shall prefer to use the redefined version of fuzzy numbers [5] here. In mathematical analysis [10,12], we have studied the notion of 'limit point'. It is an example of nucleus. 'Limit points' exist for an infinite set, whereas in our notion nuclei exist even for a finite population, by definition. There are infinite number of data centered around a 'limit point', and there are a large number of data centered around a nucleus. Every neighborhood of a limit point (even any small neighborhood with negligible positive measure) is densely populated because of the residence of infinite number of data inside it. In the notion of nucleus, it is not the same and depends upon the value of 'large'. A nucleus may cease to be so if the 'large' is made larger enough. A population may or may not have a nucleus like an infinite set may or may not have a limit point. It is obvious that nucleus will play a great role to the statisticians in information processing. Whenever we want to know data congestion in a population, we need to visualize the population data in the form of points or dots on a suitable base. Whenever we think of density of a physical quantity, we immediately think of homogeneity or heterogeneity. A philosophically analogous concept is introduced in the theory of Statistics. We introduce the notion of a new kind of mean called by LM, and then the corresponding variance LV and standard deviation LSD. A generalized kind of the notion of classical mean is done with a different philosophy on the platform of region algebra, theory of objects [6] by defining the measures like linear standard deviation (LSD), linear variance (LV), region standard deviation (RSD), and region variance (RV) are defined. Planning for future work, an almost equally potential generalization of the existing measures can be done if we consider rough metric[4] d of rough set theory [32,33] instead of the classical metric d of classical mathematical analysis [10,12]. The notion of rough metric spaces have been introduced in [4]. It has been proved in [4] that the notion of rough metric spaces and of classical metric spaces are different. It is expected that the rough distance r(x,y) will play a different role in statistical analysis in many cases (not absolutely in all cases) than the classical distance d(x,y). Our future work will be on rough [32,33] statistical measures. With the generalized notions of mean, variance, standard deviation, nucleus, etc. introduced in this work, our future work will be to generalize the existing notion of 'Expectation' of a stochastic variate, to generalize the classical measures like Median, Mode, Covariance, Correlation Coefficients, etc. to enlarge the universe of the domains of statistical analysis further. Although the present work is based on finite population, but we shall make attempt in our future work to extend all the definitions and results of this work to the cases of infinite populations or big data [8].

References

[1] A. Kaufmann, Introduction to the Theory of Fuzzy Subsets, Academic Press, New York, 1975.

[2] Bouchon-Meunier, B., Yager, R. R. and Zadeh, L. A., Fuzzy Logic and Soft Computing, World Scientific: Singapore, 1995.
[3] Biswas, R., An application of Yager's Bag Theory in Multicriteria Based Decision Making Problems, Int. Jou. of Int. System. 14(12) (1999) 1231-1238.
[4] Biswas R., Fixed point theorem in rough metric spaces, Found. Comp. & Decision Sciences. Vol. 24 (1999) 13-20.
[5] Biswas, R., Fuzzy Numbers Redefined, Information Vol.15(4) 2012, pp.1369-1380.
[6] Biswas,R., Region Algebra, Theory of Objects & Theory of Numbers, International Journal of Algebra, Vol. 6(8), 2012, 1371 – 1417.
[7] Biswas, R., Decoding the 'Progress' of Decision Making Process in the Human/Animal Cognition Systems while Evaluating the Membership Value μ(x), Issues in Intuitionistic Fuzzy Sets and Generalized Nets, Vol.10 (2013) page 21-53 (ISBN 978-83-61551-08-9), published by Warsaw School of Information Technology, Warsaw, Poland.
[8] Biswas, R., Theory of Solid Matrices & Solid Latrices, Introducing New Data Structures MA, MT : for Big Data, International Journal of Algebra, Vol.7, 2013, No. 16, 767–789.
[9] Chakraborty, K., Biswas, R., Nanda, S., On Yager's theory of bags and fuzzy bags, Int. Jou. Comp. and Artificial Intelligence. 18(1) (1999) 1-17.
[10] Copson, E.T., Metric Spaces, Cambridge University Press (1968).
[11] Dubois and Prade, Fuzzy Sets & Systems : Theory and Applications, Academic Press, New York, 1990.
[12] G.F.Simmons, Intro. to Topology and Modern Analysis, McGraw Hill, New York, 1963.
[13] I. N. Herstein, Topics in Algebra, Wiley Eastern Limited, New Delhi, 2001.
[14] James J. Buckley, Fuzzy Statistics, Springer Berlin, Heidelberg, 2004
[15] J. Rice, Mathematical Statistics and Data Analysis, 2nd edition, Duxbury Press, California, 1995.
[16] Lake, J., Sets, Fuzzy Sets, Multisets and Functions, Jou. of London Maths. Society, Vol.12(2)1976, 323-326.
[17] Li, B., Fuzzy Bags and Applcations, Fuzzy Sets and Systems, Vol.34 (1990) 61-71.
[18] Li, B., Peizhang, W. and Xihui, L., Fuzzy Bags with Set-valued Statistics, Computer Math. Applic., Vol.15 (1988) 811-818.
[19] Lech Polkowski, Andrzej Skowron, Rough Sets in Knowledge Discovery, Case Studies, and Software Systems, Physica Verlag, Heidelberg, New York, 1998
[20] Meyer, R. and McRobbie, M., Multisets and Relevant Implications – I, Australian Journal of Philosophy, Vol.60. (1982) 107-139.
[21] Meyer, R. and McRobbie, M., Multisets and Relevant Implications - II, Australian Journal of Philosophy, Vol.60. (1982) 265-281.
[22] Novak, V., Fuzzy Sets and Their Applications, Adam Hilger, 1986.

[23] Roger K. Blashfield, The Growth of Cluster Analysis: Tryon, Ward, and Johnson, Multivariate Behavioral Research, Vol.15(4) 1980, 439-458.
[24] Tremblay, J.P. and Manohar, R., Discrete Mathematical Structures with Applications to Computer Science, McGraw Hill Int. Ed., 1987
[25] Tryon, Cluster Analysis, McGraw-Hill Publishers, New York, 1939.
[26] W.D.Blizard, Multiset Theory, Notre Dame Jour. of Formal Logic, Vol.30(1989) 36-66.
[27] W. D. Blizard, Real-valued Multisets and Fuzzy Sets, Fuzzy Sets and Systems, Vol.33 (1989) 77-79.
[28] W.D.Blizard, The Development of Multiset Theory, Modern Logic,Vol.1(1991) 319-352.
[29] Yager, R. R. and Zadeh, L. A., Fuzzy Sets, Neural Networks and Soft Computing, Van Nostrand Reinhold: New York, 1994.
[30] Yager, R. R., On the Theory of Bags, Int. Jour. of Gen. Sys. Vol.13(1) (1986), p 23-37.
[31] Zadeh, L.A., Fuzzy Sets, Inform. and Control, Vol.8 (1965) 338-353.
[32] Z. Pawlak, Rough Sets, Int. Jou. of Inf. Comp. Sc., Vol.11(1982) 341-356.
[33] Zdzisław Pawlak, Rough Sets : Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

Received: December 2, 2013