1.1 Introduction:-

The purpose of this chapter is to clarify the basic concepts of Fuzzy sets theory, Information theory, and “fuzzy entropy”. Uncertainty and fuzziness are the basic nature of human thinking and of many real world objectives. Fuzziness is found in our decision, in our language and in the way we process information. The main use of information is to remove uncertainty and fuzziness. In fact, we measure information supplied by the amount of probabilistic uncertainty removed in an experiment and the measure of uncertainty removed is also called as a measure of information while measure of fuzziness is the measure of vagueness and ambiguity of uncertainties.

By the nineteen sixties it became evident in mathematical systems research that the rigorous treatment based on Aristotelian logic is not appropriate in analyzing real systems. Fuzzy sets were defined by Zadeh [109] to free the mathematical model from the law of the excluded middle. Formally, the characteristic function \( \mu_A(x) \) describing the membership of element \( x \) in the set \( A \) was generalized: in classical mathematics the characteristic function takes either the value 0 or 1; in the case of fuzzy sets the characteristic function may take any value from the real interval \([0,1]\).

The concept of fuzziness has been applied to apparently all phenomena already formalized in systems research: Statistics, Information theory, Clustering and Decision analysis, Medical and Socio-economic predictions, Image processing, etc. This overwhelming success, seen in introspect, is not surprising for the following simple reasons. The mathematical ideas applied in systems research had well known for the workers involved and their applications had prevailed the planning and analysis of information processing systems. Thus both their theoretical clarity and practical relevance had been firmly established. The fundamental but essentially mathematical generalization of these formal ideas posed many challenging questions within the conceptually and methodologically well-known (mathematical) framework.

Within the extremely large field of theories and applications developed from the concept of fuzziness, there has been a relatively small area of dealing with the fuzziness of concepts. The most important questions in this area are: How should we calculate a numerical description of particularly fuzzy quantifiers like “very”, “more or less”, “rather,” for such categories as “short,”
“old,” “many,” and for statements connecting such ideas, like “much older than”?. How should we apply different operations defined on fuzzy sets to formal logic and to conceptual categories?

The concept of fuzziness was made a scientific one in mathematical systems theory by Zadeh’s [109] definition of fuzzy sets. This advantages a “framework which provides a natural way of dealing with problems in which the source of imprecision is the absence of sharply defined criteria of class membership”. The introduction of fuzzy sets was motivated by the fuzziness of concepts, i.e., that “More often than not, the classes of objects encountered in the real physical world do not have precisely defined criteria of membership”.

1.1.1 Fuzzy Models:

Fuzzy sets are a generalization of conventional set theory that was introduced by Zadeh [109] as a mathematical way to represent vagueness in everyday life. The basic idea of fuzzy sets is easy to grasp. Suppose, as you approach a red light, you must advise a driving student when to apply the brakes. Would you say, "Begin braking 74 feet from the crosswalk"? Or would your advice be more like, "Apply the brakes pretty soon"? The latter, of course; the former instruction is too precise to be implemented. This illustrates that precision may be quite useless, while vague directions can be interpreted and acted upon. Everyday language is one example of ways vagueness is used and propagated. Children quickly learn how to interpret and implement fuzzy instructions (“go to bed about 10”). We all assimilate and use (act on) fuzzy data, vague rules, and imprecise information, just as we are able to make decisions about situations which seem to be governed by an element of chance. Accordingly, computational models of real systems should also be able to recognize, represent, manipulate, interpret, and use (act on) both fuzzy and statistical uncertainties.

The Process and progress of knowledge unfolds into two stages: an attempt to know the character of the world and a subsequent attempt to know the character of the knowledge itself. The second reflective stage arises from the failures of the first; it generates an enquiry into the possibility of knowledge and into the limits of that possibility. It is in this second stage of enquiry that we find ourselves today. As a result, our concerns with knowledge, perception of problems and attempts at solutions are of a different order than in the past. We want to know not only specific facts or truths but what we cannot know, what we do and do not know, and how we know at all. Our problems have shifted from questions of how to cope with the world (how
provide ourselves with food, shelter, and so on, to questions of how to cope with knowledge (and ignorance) itself. Ours has been called an “information society,” and a major portion of our economy is devoted to the handling, processing, selecting, storing, disseminating, protecting, collecting, analyzing, and sorting of information, our best tool for this being, of course, the computer.

Our problems are seen in terms of decision, management, and prediction; solutions are seen in terms of faster access to more information and of increased aid in analyzing uncertainty, taken together constitute the ground of many of our problems today: complexity. As we become aware of how much we know and how much we do not, as information and uncertainty themselves become the focus of our concern, we begin to see our problems as centering on the issue of complexity.

How do we manage to cope with complexity as well as we do, and how could we manage to cope better? The answer seems to lie in the notion of simplifying complexity by making a satisfactory trade-off or compromise between the information available to us and the amount of uncertainty we allow. One option is to increase the amount of allowable uncertainty by sacrificing some of the precise information in favor of a vague but more robust summary. For instance, instead of describing the weather today in terms of the exact percentage of cloud cover (which would be much too complex), we could just say that it is sunny, which is more uncertain and less precise but more useful. In fact, it is important to realize that the imprecision or vagueness that is characteristic of natural language does not necessarily imply a loss of accuracy or meaningfulness. It is for instance, generally more meaningful to give travel directions in terms of city blocks than in terms of inches, although the former is much less precise than the latter. It is also more accurate to say that it is usually warm in the summer than to say that it is usually 72 degrees in the summer. In order for a term such as sunny to accomplish the desired introduction of vagueness, however, we cannot use it to mean precisely 0% cloud cover. Its meaning is not totally arbitrary, however; a cloud cover of 100% is not sunny and neither, in fact, is a cloud cover of 80%. We can accept certain intermediate states, such as 10 or 20% cloud cover, as sunny. But where do we draw the line? If, for instance, any cloud cover of 25% or less is considered sunny, does this mean that a cloud cover of 26% is not? This is clearly unacceptable;

...
since 1% of cloud cover hardly seems like a distinguishing character between sunny and not sunny. We could therefore add a qualification that any amount of cloud cover greater than 1% will also be labeled as sunny. We can see, however, that this definition eventually leads us to accept all degrees of cloud cover as sunny, no matter how gloomy the weather looks! In order to resolve this paradox, the term sunny may introduce vagueness by allowing some sort of gradual transition from degrees of cloud cover that are considered to be sunny and those that are not. This is, in fact precisely the basic concept of fuzzy set, a concept that is both simple and intuitively pleasing and that forms, in essence, a generalization of the classical or crisp set.

Fuzzy interpretations of data structures are a very natural and intuitively plausible way to formulate and solve various problems. Conventional (crisp) sets contain objects that satisfy precise properties required for membership. The set of numbers $H$ from 6 to 8 is crisp; we write $H = \{r \in \mathbb{R} | 6 \leq r \leq 8\}$. Equivalently, $H$ is described by its membership (or characteristic, or indicator) function $\mu_H: \mathbb{R} \to \{0, 1\}$ defined as

$$\mu_H(r) = \begin{cases} 1 & 6 \leq r \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

The crisp set $H$ and the graph of $\mu_H$ are shown in the left half of Fig. 1.1.1(a). Every real number $r$ either is in $H$ or is not. Since $\mu_H$ maps all real numbers $r \in \mathbb{R}$ onto the two points $(0, 1)$, crisp sets correspond to two-valued logic: is or isn't, on or off, black or white, 1 or 0. In logic, values of $\mu_H$ are called truth values with reference to the question, "Is $r$ in $H$?" The answer is yes if and only if $\mu_H(r) = 1$; otherwise, no.

Fig. 1.1.1 (a): Membership functions for hard and fuzzy subsets of $\mathbb{R}$. 
Consider next the set $F$ of real numbers that are close to 7. Since the property "close to 7" is fuzzy, there is not a unique membership function for $F$. Rather, the modeler must decide, based on the potential application and properties desired for $F$, what $\mu_F$ should be. Properties that might seem plausible for this $F$ include (i) normality ($\mu_F(7) = 1$), (ii) monotonicity (the closer $r$ is to 7, the closer $\mu_F(r)$ is to 1, and conversely) and (iii) symmetry (numbers equally far left and right of 7 should have equal memberships). Given these intuitive constraints, either of the functions shown in the right half of Fig. 1.1.1 (a) might be a useful representation of $F$. $\mu_{F1}$ is discrete (the staircase graph), while $\mu_{F2}$ is continuous but not smooth (the triangle graph). One can easily construct a MF for $F$ so that every number has some positive membership in $F$, but we would not expect numbers "far from 7," 20 000 987 for example, to have much! One of the biggest differences between crisp and fuzzy sets is that the former always have unique MFs, whereas every fuzzy set has an infinite number of MFs that may represent it. This is at once both a weakness and strength; uniqueness is sacrificed, but this gives a concomitant gain in terms of flexibility, enabling fuzzy models to be "adjusted" for maximum utility in a given situation.

1.1.2 Fuzzy Sets Theory:

Let $X$ be a space of objects and $x$ be a generic element of $X$. A classical set $A$, $A \subseteq X$, is defined as a collection of elements or objects $x \in X$, such that each element ($x$) can either belong or not to the set $A$. By defining a characteristic (or membership) function for each element $x$ in $X$, we can represent a classical set $A$ by a set of ordered pairs $(x,0)$ or $(x,1)$, which indicates $x \notin A$ or $x \in A$, respectively. Unlike the aforementioned conventional set, a fuzzy set expresses the degree to which an element belongs to a set. Hence the membership function of a fuzzy set is allowed to have values between 0 and 1, which denote the degree of membership of an element in the given set.

1.1.3 Fuzzy sets and membership functions:

If $X$ is a collection of objects denoted generically by $x$, then a fuzzy set $A$ in $X$ is defined as a set of ordered pairs $A = \{(x, \mu_A(x))/x \in X\}$, where $\mu_A(x)$ is called the membership function (or MF for short) for the fuzzy set $A$. The MF maps each element of $X$ to a membership grade (or membership value) between 0 and 1 (included). Obviously, the definition of a fuzzy set is a simple extension of the definition of a classical (crisp) set in which the characteristic function is permitted to have any values between 0 and 1. If the value of the
membership function is restricted to either 0 or 1, then \( A \) is reduced to a classical set. For clarity, we shall also refer to classical sets as ordinary sets, crisp sets, non-fuzzy sets, or just sets. Usually \( X \) is referred to as the universe of discourse, or simply the universe, and it may consist of discrete (ordered or non-ordered) objects or it can be a continuous space. This can be clarified by the following examples.

1.1.4 **Fuzzy sets with a discrete non-ordered universe:**

Let \( X = \{\text{San Francisco, Boston, Los Angeles}\} \) be the set of cities one may choose to live in. The fuzzy set \( A = "\text{desirable city to live in}" \) may be described as follows: \( A = \{\text{(San Francisco, 0.9), (Boston, 0.8), (Los Angeles, 0.6)}\} \). Apparently the universe of discourse \( X \) is discrete and it contains non-ordered objects - in this case, three big cities in the United States. As one can see, the foregoing membership grades listed above are quite subjective; anyone can come up with three different but legitimate values to reflect his or her preference.

1.1.5 **Fuzzy sets with a discrete ordered universe:**

Let \( X = \{0, 1, 2, 3, 4, 5, 6\} \) be the set of numbers of children a family may choose to have. Then the fuzzy set \( B = "\text{desirable number of children in a family}" \) may be described as follows: \( B = \{(0, 0.1), (1, 0.3), (2, 0.7), (3, 1), (4, 0.7), (5, 0.3), (6, 0.1)\} \). Here we have a discrete ordered universe \( X \); the MF for the fuzzy set \( B \) is shown in Fig. 1.1.5(a).

![Fig: 1.1.5 (a)](image1)

![Fig: 1.1.5 (b)](image2)

Again, the membership grades of this fuzzy set are obviously subjective measures.
1.1.6 Fuzzy sets with a continuous universe:

Let \( X = \mathcal{R}^+ \) be the set of possible ages for human beings. Then the fuzzy set \( C = \text{"about 50 years old"} \) may be expressed as \( C = \{(x, \mu_c(x)/x \in X)\} \), where

\[
\mu_c(x) = \frac{1}{1 + \left(\frac{x-50}{10}\right)^4}
\]

This is illustrated in Figure 1.1.5(b). From the preceding examples, it is obvious that the construction of a fuzzy set depends on two things: the identification of a suitable universe of discourse and the specification of an appropriate membership function. The specification of membership functions is subjective, which means that the membership functions specified for the same concept by different persons may vary considerably. This subjectivity comes from individual differences in perceiving or expressing abstract concepts and has little to do with randomness. Therefore, the subjectivity and non-randomness of fuzzy sets is the primary difference between the study of fuzzy sets and probability theory, which deals with objective treatment of random phenomena.

In practice, when the universe of discourse \( X \) is a continuous space, we usually partition it into several fuzzy sets whose MFs cover \( X \) in a more or less uniform manner. These fuzzy sets, which usually carry names that conform to adjectives appearing in our daily linguistic usage, such as "large," "medium," or "small," are called linguistic values or linguistic labels. Thus, the universe of discourse \( X \) is often called the linguistic variable.

1.2 Some nomenclature used in the literature:-

1.2.1 Support of a Fuzzy Set:

The support of a fuzzy set \( A \) is the set of all points \( x \) in \( X \) such that \( \mu_A(x) > 0 \).

1.2.2 Core of a Fuzzy Set:

The core of a fuzzy set \( A \) is the set of all points \( x \) in \( X \) such that \( \mu_A(x) = 1 \).

1.2.3 Normality of a Fuzzy Set:

A fuzzy set \( A \) is normal if its core is nonempty. In other words, we can always find at least a point \( x \in X \) such that \( \mu_A(x) = 1 \).
1.2.4 Crossover Points:

A crossover point of a fuzzy set $A$ is a point $x \in X$ at which $\mu_A(x) = 0.5$.

1.2.5 Fuzzy Singleton:

A fuzzy set whose support is a single point in $x$ with $\mu_A(x) = 1$ is called a fuzzy singleton.

1.2.6 $\alpha$-Cut, Strong $\alpha$-Cut:

The $\alpha$-cut or $\alpha$-level set of a fuzzy set $A$ is a crisp set defined by $A_\alpha = \{x / \mu_A(x) \geq \alpha\}$. Strong $\alpha$-cut or strong $\alpha$-level set are defined similarly $A'_\alpha = \{x / \mu_A(x) > \alpha\}$. Using this notation, we can express the support and core of a fuzzy set $A$ as support $(A) = A'_0$ and core $A = A_1$.

1.2.7 Convexity:

A fuzzy set $A$ is convex if and only if for any $x_1, x_2 \in X$ and any $\lambda \in [0, 1],$

$$\mu_A(\lambda x_1 + (1-\lambda) x_2) \geq \min \{\mu_A(x_1), \mu_A(x_2)\}.$$ 

Alternatively, $A$ is convex if all its $\alpha$-level sets are convex. It is to be noted that the definition of convexity of a fuzzy set is not as strict as the common definition of convexity of a function.

1.2.8 Linguistic variables and linguistic values:

Suppose that $X = \text{"age."}$ Then we can define fuzzy sets "young," "middle aged" and "old" that are characterized by MFs. Just as a variable can assume various values, a linguistic variable "age" can assume different linguistic values, such as "young," "middle aged" and "old" in this case. If "age" assumes the value of "young," then we have the expression "age is young," and so forth for the other values.

1.2.9 Fuzzy numbers:

A fuzzy number $A$ is a fuzzy set in the real line that satisfies the conditions for normality and convexity. Most fuzzy sets used in the literature satisfy the conditions for normality and convexity, so fuzzy numbers are the most basic type of fuzzy sets.

Union, intersection, and complement are the most basic operations on classical sets. On the basis of these three operations, a number of identities can be established. Corresponding to the ordinary set operations of union, intersection and complement, fuzzy sets have similar operations, which were initially defined in Zadeh's seminal paper [109]. Before introducing these
three fuzzy set operations, first we shall define the notion of containment, which plays a central role in both ordinary and fuzzy sets. This definition of containment is, of course, a natural extension of the case for ordinary sets.

1.2.10 Containment or Subset:

Fuzzy set A is contained in fuzzy set B (or, equivalently, A is a subset of B, or A is smaller than or equal to B, A \( \subseteq B \)) if and only if \( \mu_A(x) \leq \mu_B(x) \) for all x.

1.2.11 Union (disjunction):

The union of two fuzzy sets A and B is a fuzzy set C, written as

\[ C = A \cup B \text{ or } C = A \lor B, \]

whose MF is related to those of A and B by

\[ \mu_C(x) = \max(\mu_A(x), \mu_B(x)). \]

1.2.12 Intersection (conjunction):

The intersection of two fuzzy sets A and B is a fuzzy set C, written as \( C = A \cap B \text{ or } C = A \land B \), whose MF is related to those of A and B by

\[ \mu_C(x) = \min(\mu_A(x), \mu_B(x)). \]

1.2.13 Complement (negation):

The complement of fuzzy set A, denoted by \( \overline{A} \) or NOT A, is defined as

\[ \mu_{\overline{A}}(x) = 1 - \mu_A(x). \]

Since, the operations introduced above perform exactly as the corresponding operations for ordinary sets if the values of the membership functions are restricted to either 0 or 1. However, it is understood that these functions are not the only possible generalizations of the crisp set operations. For each of the aforementioned three set operations, several different classes of functions with desirable properties have been proposed subsequently in the literature (e.g. algebraic sum for union and product for intersection). In general, union and intersection of two fuzzy sets can be defined through T-conorm (or S-norm) and T-norm operators respectively. These two operators are functions \( S, T: [0,1] \times [0,1] \rightarrow [0,1] \) satisfying some convenient boundary, monotonicity, commutativity and associativity properties. As pointed out by Zadeh
[109], a more intuitive but equivalent definition of union is the, “smallest” fuzzy set containing both A and B. Alternatively, if D is any fuzzy set that contains both A and B, then it also contains A ∪ B. Analogously, the intersection of A and B is the "largest" fuzzy set which is contained in both A and B. The two fundamental (Aristotelian) laws of crisp set theory are:

(a) **Law of Contradiction:**

\[ A ∪ \overline{A} = X \text{(i.e., a set and its complement must comprise the universe of discourse)}, \]

(b) **Law of Excluded Middle:**

\[ A ∩ \overline{A} = \emptyset \text{(i.e., an object can either be in its set or its complement; it cannot simultaneously be in both)}. \]

It can be easily seen that for every fuzzy set that is non-crisp (i.e., whose membership function does not only assume values 0 and 1) both laws are broken (i.e., for fuzzy sets \( A ∪ \overline{A} \neq X \) and \( A ∩ \overline{A} \neq \emptyset \). Indeed if \( \forall x ∈ A \) such that

\[ μ_A(x) = α, 0 < α < 1; μ_{A∪\overline{A}}(x) = \max\{α, 1 − α\} \neq 1 \]

and

\[ μ_{A∩\overline{A}}(x) = \min\{α, 1 − α\} \neq 0. \]

### 1.3 Interpreting the Membership Function:

The first point to note is that, like \( P(A) \), the probability of a set A, fuzzy set theory does not tell us how to specify \( μ_A(x) \), the membership function of a fuzzy set A. The second point to note is that whereas there is a logical requirement that \( P(A) ∈ [0, 1] \), the fact that \( μ_A(x) ∈ [0, 1] \) is simply a convenience of scaling. The third point to note is that whereas \( P(A) \) can be interpreted as a two-sided bet (which in principle can be settled when A reveals itself), \( μ_A(x) \) reflects an individual’s view of the extent to which \( x ∈ A \); thus \( μ_A(x) \) cannot be made operational in the same sense as \( P(A) \). Finally, it is not a requirement that \( \sum x μ_A(x) \) be 1, and thus \( μ_A(x) \) as a function of x cannot be interpreted as a probability or, for that matter, as a conditional probability, as was done by Loginov [73] and also by Barrett and Woodall [19]. How then can we interpret the membership function \( μ_A(x) \)? Because \( μ_A(x) \), as a function of x, reflects the extent to which \( x ∈ A \) [i.e., \( μ_A(x) \) is an indicator of how likely it is that \( x ∈ A \)], we may view \( μ_A(x) \) as the likelihood of x for a fixed (i.e., specified) A. Even though the interpretation of a likelihood is almost always derived from a probability model, the likelihood is not a probability (in particular, it does not obey the addition rule) and in statistical inference, the likelihood
function reflects the relative degrees of support that a fixed observation provides to several hypotheses. Furthermore, the specification of likelihood is subjective. Thus our interpretation of the membership function is that it is a likelihood function with $\mathcal{A}$ taking the role of a fixed observation and the values of $x$ taking the role of the hypotheses.

To statisticians specializing in inference, our interpretation of the membership function as a likelihood will appear to be unconventional. This is because in the context of inference, the likelihood entails a fixed observation and a varying parameter. However, our structure for the likelihood is a consequence of the notion of the likelihood from a more philosophical viewpoint, and what we have proposed is in keeping with the foundational notion of likelihood. Basu[20].

The foregoing points are best illustrated via the following example involving two fuzzy sets $A$ and $B$, where

$$A = \{x; x \in X \text{and } x \text{ is } "medium"\}$$

And

$$B = \{x; x \in X \text{ and is } "small"\}$$

as before, $X = \{1, 2, 3, ..., 10\}$.

Suppose that an assessor assigns the membership functions $\mu_A(x)$ and $\mu_B(x)$ given in the table below.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\mu_A(x)$</th>
<th>$\mu_B(x)$</th>
<th>$\mu_A(x) + \mu_B(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>.2</td>
<td>.8</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>.5</td>
<td>.5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>.8</td>
<td>.3</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.1</td>
<td>1.1</td>
</tr>
<tr>
<td>6</td>
<td>.8</td>
<td>0</td>
<td>.8</td>
</tr>
<tr>
<td>7</td>
<td>.5</td>
<td>0</td>
<td>.5</td>
</tr>
<tr>
<td>8</td>
<td>.2</td>
<td>0</td>
<td>.2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| Col. Sums | 4          | 3.7        | 7.7                     |
Clearly, $\sum x \mu_A(x)$ and $\sum x \mu_B(x)$ are not 1, nor is it true that $\mu_A(x) + \mu_B(x)$ is necessarily 1. A plot of $\mu_A(x)$ and $\mu_B(x)$ as a function of $x$, with $\mu_A(x)$ and $\mu_B(x)$ viewed as likelihoods is shown in Figure 1.3.1(a). The plots reflect the extent to which an $x$ belongs to the sets $A$ and $B$.

### 1.4 Probability: A Calculus For The uncertainty Of Outcomes:

The underlying set-theoretic premise for considering probability and its calculus is an experiment, $\mathcal{E}$, which is yet to be performed. Let $x$ denote a generic outcome of $\mathcal{E}$, and let $\Sigma$ denote the set of all conceived outcomes of $\mathcal{E}$; thus $x \in \Sigma$. It is important to note that the probability theory does not tell one how to specify $\Sigma$; this choice is subjective and is up to the user. For convenience, we assume that $\Sigma$ is a countable set. Let $\mathcal{F}$ denote a set whose members are subsets of $\Sigma$; that is, $\mathcal{F}$ is a family of sets. However, $\mathcal{F}$ is such that it contains $\Sigma$ and $\emptyset$, where $\emptyset$ is the null set. Furthermore, $\mathcal{F}$ is closed under unions and intersections; that is, if $A, B \in \mathcal{F}$, then $(A \cup B)$ and $(A \cap B) \in \mathcal{F}$. The subsets of $\Sigma$ are called events, and in probability theory it is presumed that the events are well defined or “sharp” (also known as “crisp”); that is, there is no ambiguity in declaring whether any outcome $x$ of $\Sigma$ belongs to $A$ or to its complement $A^c$. In contrast, with fuzzy sets there is ambiguity in classifying an $x$ in a subset $A$ or $A^c$, because $A$ is not sharply defined. If the outcome of $\mathcal{E}$, say $x$, is such that $x \in A$, then we say that event $A$ has occurred. Because $\mathcal{E}$ is yet to be performed, we are uncertain about the occurrence of any particular $x$. Consequently, we are also uncertain about the occurrence of event $A$. We describe this uncertainty by a number, $P(A)$, where $0 \leq P(A) \leq 1$; $P(A)$ is the probability of event $A$, or
the probability measure of the set $A$. There are several interpretations of $P(A)$; the one that is
germane to our interest here is that $P(A)$ is a two-sided bet (or wager) on the occurrence of event $A$. Specifically, $P(A)$ is the amount that one is willing to stake out in exchange for a dollar should event $A$ occur or, equivalently, $(1 - P(A))$ is the amount staked in exchange for a dollar should event $A$ not occur. Furthermore, the individual specifying $P(A)$ is required to be indifferent between betting on $A$ or $A^c$. The two-sided bet will be settled when $E$ is performed and $\omega$ is observed, so that the disposition of $A$ is known. An advantage of the foregoing interpretation of $P(A)$ is that probability can be made “operational” via the mechanism of betting. This interpretation of probability is a basis for a personalistic (or a subjectivistic) theory of probability. It is important to note that probability theory does not tell us how to arrive at a particular $P(A)$, nor does it in its purely abstract form even attempt to interpret $P(A)$. Many probabilists would declare that the assignment of initial probabilities is a job for a statistician, though some would say that the role of a statistician is to help clients formulate their prior knowledge, because it is the client who knows.

The calculus (or the algebra) of probability tells one how the various uncertainties (i.e., the initial probabilities) combine or cohere. In particular, if $P(B)$ denotes the quantification of uncertainty of another event $B$, then

a). $$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$ where

b). $$P(A \cap B) = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ \frac{P(A/B)P(B)}{P(A)} & \text{otherwise} \end{cases}$$

The quantity $P(A/B)$ is called the conditional probability of $A$ were $B$ to occur. Like $P(A)$, $P(A/B)$ should lie between 0 and 1; it represents the amount that one is willing to stake on the event $A$ should the event $B$ occur but under the proviso that all bets on $A$ will be called off should $B$ not occur. It is crucial to bear in mind that $P(A/B)$ is a bet in the subjunctive mood; this is because the disposition of $B$ is unknown when $P(A/B)$ is specified. Finally, ignoring the relevance of a conditioning event, events $A$ and $B$ are said to be mutually independent if $P(A/B) = P(A)$. The calculus given earlier has an axiomatic foundation based on behavioristic considerations.

Thus, to summarize, a foundation for the theory of probability is based on the following ingredients:
a) A well-defined set $\mathcal{S}$ and subsets of $\mathcal{S}$.

b) An adherence to the “law of the excluded middle,” the essential import of which is that any outcome $\omega$ of $\mathcal{E}$ belongs to a set $A$ or to a set $A^c$, but not to both.

c) A calculus based on behaviorist axioms involving numbers between 0 and 1 that can be made operational once $\mathcal{E}$ is performed and its outcome observed.

1.4.1 Probability Measures of Fuzzy Events:

In probability theory [101], an event, $A$, is a member of $\sigma$-field $\mathcal{A}$, of subsets of a sample space $\Omega$. A probability measure, $P$, is a normed measure over a measurable space $(\Omega, \mathcal{A})$; that is, $P$ is a real valued function which assigns to every $A \in \mathcal{A}$; a probability $P(A)$, such that

a) $P(A) \geq 0$ for all $A \in \mathcal{A}$;

b) $P(\Omega) = 1$; and

c) $P$ is countably additive, i.e., if $\{A_i\}$ is any collection of disjoint events then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \tag{1.4.1}$$

The notion of an event and its probability constitute the most basic concepts of probability theory. As defined above, an event is a precisely specified collection of points in the sample space. By contrast, in every day experience one frequently encounters situations in which an “event” is a fuzzy rather than sharply defined collection of points. For example, the ill defined events “it is a warm day,” “$x$ is approximately equal to 5,” “in twenty tosses of a coin there are several more heads than tails,” are fuzzy because of imprecision of the meaning of the underlined words.

By using the concepts of a fuzzy set [109], the notion of an event and its probability can be extended in a natural fashion to fuzzy events of the type explained above. It is possible that such an extension may eventually significantly enlarge the domain of applicability of probability theory, especially in those fields in which fuzziness is a pervasive phenomenon.

Let us assume that for simplicity that $\Omega$ is an Euclidean n-space $R^n$. Thus our probability space will be assumed to be a triplet $(R^n, \mathcal{A}, P)$, where $\mathcal{A}$ is $\sigma$-field of Borel sets in $R^n$ and $P$ is a probability measure over $R^n$. A point in $R^n$ will be denoted by $x$.

Let $A \in \mathcal{A}$, then the probability of $A$ can be expressed as


\[ P(A) = \int_A dP \quad (1.4.2) \]

Or equivalently

\[ P(A) = \int_{R^n} \mu_A(x) dP = E(\mu_A). \quad (1.4.3) \]

Where \( \mu_A \) denotes the characteristic function of \( A (\mu_A(x) = 0 \text{ or } 1) \). And \( E(\mu_A) \) is the expectation of \( \mu_A \).

The equation (1.4.3) equates the probability of an event \( A \) with the expectation of the characteristic function of \( A \). It is this equation that can readily be generalized to fuzzy events through the use of the concept of fuzzy set.

**Definition 1.4.1:** Let \((R^n, \alpha, P)\) be a probability space in which \( \alpha \) is a \( \sigma \)-field of Borel sets in \( R^n \) and \( P \) is a probability measure over \( R^n \). Then fuzzy event in \( R^n \) is a fuzzy set \( A \) in \( R^n \) whose membership function, \( \mu_A: R^n \to [0, 1] \) is Borel measurable. The probability of a fuzzy event \( A \) is defined by the Lebesgue-Stieltjes integral

\[ P(A) = \int_{R^n} \mu_A(x) dP = E(\mu_A). \quad (1.4.4) \]

Thus as in (1.4.3), the probability of a fuzzy event is the expectation of its membership function. The existence of the Lebesgue-Stieltjes integral is insured by the assumption that \( \mu_A \) is Borel measurable.

The above definition of a fuzzy event and its probability form a basis for generalizing within the framework of the theory of fuzzy sets a member of the concepts and results of probability theory, information theory and related fields.

### 1.5 Information theory:-

It is a branch of probability and statistics with extensive potential applications to communication system. Like several other branches of mathematics, information theory has a physical origin. It was initiated by communication scientists C.E. Shannon [87], who were studying the statistical structure of electrical communication equipments. The subject followed by a flood of research papers speculating upon the possible applications to a broad spectrum of research areas, such as pure mathematics, semantics, physics, management, thermodynamics, botany, econometrics, operations research, psychology, epidemiological studies, disease management and related disciplines.
The Mathematical Theory of Communication is the early work of R. V. L. Hartley on the mathematics of information transmission that is recognized. R. A. Fisher introduced notion i.e. Fisher information in 1925 which is closely related to Claude Shannon’s notion of entropy. What follows is not intended as a general introduction to information theory through two outstanding contributions to the mathematical theory of communications in 1948 and 1949. Despite several hasty generalization which produces thousands research papers, one thing became evident; this scientific theory has stimulated the interest of thousands of scientists around the world.

1.5.1 Shannon’s Information Theory:

Claude E. Shannon’s “A Mathematical Theory of Communication” [87] is considered as the “Magna carta” of the Information Age. Shannon’s discovery of the fundamental laws of data comprehension and transmission marks the birth of “Information Theory”.

Information theory started out as an engineering project. Shannon’s simple goal was to find a way to clear up noisy telephone connections. Today, there would be no internet without Shannon’s theory. Every new modem upgrade, every compressed file, which includes any in (.gif) or (.jpeg) format, owes something to information theory of Shannon. Even the everyday compact disc would not be possible without error connection based on information theory. To solve the “noise” problem in communications, Shannon developed a new concept, the “channel” and its associated concepts “the channel capacity” and the “redundancy”.

Shannon and Weaver [87] suppose a set of possible events whose probabilities of occurrence are $(p_1, p_2, ..., p_n)$. These probabilities are known but that is all we know concerning which event will occur. Then it is asked: “Can we find a measure of how much ‘choice’ is involved in the selection of the event or how uncertain we are of the outcome?” If there is such a measure, say $H(p_1, p_2, ..., p_n)$ it is reasonable to require of it the following properties:

(i) It should be continuous in the probabilities $(p_i)$.

(ii) If all the $(p_i)$ are equal, $p_i = 1/n$ then $H$ should be monotonic increasing function of $(n)$. With equally likely events, there is more choice, or uncertainty, when there are more possible events.

(iii) If a choice be broken down into two successive choices, the original $H$ should be the weighted sum of the individual values of $H$. 
function was recognized as “Entropy” as in Boltzmann’s famous $H$ theorem in statistical mechanics.

1.5.2 Information Function:

Let $E_i$ be the $ith$ event with probability of occurrence $p_i$, the information function may be defined as

$$ h(p_i) = - \log(p_i) \quad (1.5.1). $$

1.5.3 Shannon’s Entropy

Let $X$ be a discrete random variable taking on a finite number of possible values

$$ X = (x_1, x_2, \ldots, x_n) $$

happening with probabilities $P = (p_1, p_2, \ldots, p_n)$, $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$, $i = 1, 2, \ldots, n$

we denote

$$ H(P) = H(p_1, p_2, \ldots, p_n) = - \sum_{i=1}^{n} p_i \log p_i \quad (1.5.3). $$

Generally, the base of logarithm is taken ‘2’ and it is assume $\log 0 = 0$. When the logarithm is taken as a base ‘2’ the unit of information is called a ‘bit.’ When the natural logarithm is taken, the resulting unit is called a ‘nit’. If the logarithm is taken with base 10, the unit of information is known as ‘Hartley’.

The information measure (1.5.3) satisfies the following properties.

(1) **Non-negativity:**

$$ H(p_1, p_2, \ldots, p_n) \geq 0 $$

The entropy is always non-negative.
(2) **Symmetry:**

\[ H(p_1, p_2, \ldots, p_n) = H(p_{k(1)}, p_{k(2)}, \ldots, p_{k(n)}) \quad \forall (p_1, p_2, \ldots, p_n) \in P \]

where \((k(1), k(2), \ldots, k(n))\) is an arbitrary permutation on \((1, 2, \ldots, n)\) \(H(p_1, p_2, \ldots, p_n)\) is a symmetric function on every \(p_i, i=1,2,\ldots,n\)

(3) **Normality:**

\[ H\left(\frac{1}{2}, \frac{1}{2}\right) = 1 \] The entropy becomes unity for two equally probable events.

(4) **Expansibility:**

\[ H_n (p_1, p_2, \ldots, p_n) = H_{n+1} (0, p_1, p_2, p_3, \ldots, p_n) = H_{n+1} (p_1, p_2, p_n, 0, p_{i+1}, \ldots, p_n) = \ldots \]

\[ = H_{n+1} (p_1, p_2, \ldots, p_n, 0) \]

(5) **Recursively:**

\[ H_n (p_1, p_2, \ldots, p_n) = H_{n-1} (p_1 + p_2, \ldots, p_n) + (p_1 + p_2) H_2 \left( \frac{P_1}{p_1 + p_2}, \frac{P_1}{p_1 + p_2} \right) \]

where \(p_i \geq 0\) with \(p_1 + p_2 > 0\)

(6) **Decisively:**

\[ H_2 (1, 0) = H_2 (0, 1) = 0 \]

If one of the events is sure to occur then the entropy is zero in the scheme.

(7) **Maximality**

\[ H (p_1, p_2, \ldots, p_n) \leq H \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right) = \log n \]

The entropy is maximum when all the events have equal probabilities.

(8) **Additivity:**

\[ H_{nl} (PQ) = H_{nl} (p_1 q_1, p_1 q_2, \ldots, p_1 q_i, p_2 q_2, \ldots, p_2 q_i, \ldots, p_n q_1, p_n q_2, \ldots, p_n q_i) = H_n (p_1, p_2, \ldots, p_n) + H_l (q_1, q_2, \ldots, q_i) \]
For all \((p_1, p_2, \ldots, p_n) \in P\) and for all \((q_1, q_2, \ldots, q_l) \in Q\).

If the two experiments are independent then the entropy contained in the experiment is equal to the entropy in the first experiment plus entropy in the second experiment.

\[ (9) \text{ Strong Additivity:} \]

\[
H_{nl}(PQ) = H_{nl}(p_1q_1, p_2q_2, \ldots, p_nq_n) = H_n(p_1, p_2, \ldots, p_n) + \sum_{i=1}^{n} p_i H_q(q_{i1}, q_{i2}, \ldots, q_{in}) \text{ for all } (p_1, p_2, \ldots, p_n) \in P \text{ and for all } (q_1, q_2, \ldots, q_l) \in Q \text{ and } q_{il} \text{ are conditional probabilities i.e., entropy contained in the two experiments is equal to the entropy in the first plus the conditional entropy in the second experiment given that the first experiment has occurred.}
\]

The Shannon’s entropy (1.5.3) was characterized by Shannon assuming a set of postulates. There exists several other characterization of the measure (1.5.3) using different set of postulates.

1.6 Coding theorems:-

The elements of a finite set of \(n\) input symbols \(X = (x_1, x_2, \ldots, x_n)\) be encoded using alphabet of \(D\) symbols. The number of symbols in a codeword is called the length of the codeword. It becomes clear that some restriction must be placed on the assignment of codeword’s. One of the restrictions may be that the sequence may be decoded accurately. A code is uniquely decipherable if every finite sequence of code character corresponds to at most one message. In other words, we can say uniquely decipherability is to require that no code be prefix of another codeword. We mean by prefix that a sequence ‘A” of code character is prefix of a sequence ‘B’, if ‘B’ may be written as ‘AC’ for some sequence ‘C’.

A code having the property that no codeword is prefix of another codeword is said to be instantaneous code. Kraft [65] proved that instantaneous/uniquely decipherable code with lengths \((l_1, l_2, \ldots, l_n)\) is possible iff

\[
\sum_{i=1}^{n} D^{-l_i} \leq 1 \quad (1.6.1)
\]

where \(D\) is the size of the code alphabet. Also if
is the average codeword length where \( p_i \) is the probability of the \( i^{th} \) input symbol to a noiseless channel then for a code which satisfy (1.6.1), the following inequality holds

\[
L \geq \frac{H(P)}{\log D}, \quad (1.6.3)
\]

by suitable encoding the message, the average code length can be arbitrarily close to \( H(P) \).

Shannon’s [87] and Renyi’s [84] entropies have been studied by several research workers. The study has been carried out from essentially two different points of view. The first is an axiomatic approach and the second is a pragmatic approach. However, these approaches have little connection with the coding theorem of information theory.

Campbell [28] defined a codeword length of order as

\[
L(t) = \frac{1}{t} \log \left( \sum_{i=1}^{a} p_i D^{t_i} \right), \quad -1 < t < \infty, \quad t \neq 0, \quad (1.6.4)
\]

and developed a noiseless coding theorem for Renyi’s [84] entropy of order \( \alpha \) which is quite similar to the noiseless coding theorem for Shannon’s [87] entropy.

By means of prefix code Gurdial and Pessoa [45], Sharma et al [89], Bernard and Sharma [24], Autar and Soni [8], Autar and khan [9] Beig and Zaheerudin [7], Singh, Kumar and Tuteja[92], etc. have established coding theorems for various information measures.

1.6.1 Theorem: A necessary and sufficient condition for the existence of a instantaneous code \( S(x_i) \) such that the length of each word \( S(x_i) \) should \( l_i, i=1,2,\ldots,n \) is that the Kraft inequality [65].

\[
\sum_{i=1}^{n} D^{-l_i} \leq 1 \quad (1.6.5)
\]

should hold, where \( D \) is the number of symbols in the code alphabet.

Proof: Necessary part:

First suppose that there exists a code \( S(x_i) \) with the word length \( l_i, i=1,2,\ldots,n \).
Define \( m = \max \{ l_i, i=1,2,\ldots,n \} \) and let \( u_j, j=1,2,\ldots,n \) be the number of codeword’s with length \( j \) (some \( u_j \) may be zero). Thus the number of codeword’s with only one letter cannot be larger than \( D \)

\[
u_1 \leq D
\]  
(1.6.6)

The number of codeword’s of length 2, can use only of the remaining \((D - u_i)\) symbols in their first place, because of prefix property four of our codes, while any of the \( D \) symbols can be used in the second place, thus

\[
u_2 \leq (D-u_1) = D^2 - u_1 D
\]  
(1.6.7).

Similarly,

\[
u_3 \leq [(D-u_1)(D-u_2)]D = D^3 - u_1 D^2 - u_2 D
\]  
(1.6.8).

Finally, if \( m \) is the maximum length of the encoded words, one concludes that

\[
u_m \leq D^m - u_1 D^{m-1} - u_2 D^{m-2} - \ldots - u_{m-1} D
\]  
(1.6.9).

Dividing (1.6.9) by \( D^{-m} \), we get

\[
0 \leq 1 - u_1 D^{-1} - u_2 D^{-2} - \ldots - u_{m-1} D^{1-m} - u_m D^{-m}
\]  
(1.6.10).

Or

\[
\sum_{i=1}^{n} u_i D^{-i} \leq 1.
\]  
(1.6.11).

It may not be obvious that this condition is identical with (1.6.5) but note that \( m \geq l_i \), \( i = 1,2,\ldots,n \) and \( \sum_{i=1}^{n} u_i D^{-i} \leq 1 \) means the sum of ‘the members of all sequences of length \( i \) multiplied by \( D^{-i} \), where the summation extends from 1 to \( m \). The left hand side of (1.6.11) can be written as

\[
\sum_{i=1}^{n} u_i D^{-i} = \frac{1}{D^{u_1}} + \frac{1}{D^{u_2}} + \frac{1}{D^{u_3}} + \ldots + \frac{1}{D^{u_m}}
\]  
(1.6.12).
Each bracketed expression corresponds to message \( x_i \), and thereof the total number of term is \( n \).

\[
\begin{array}{c}
\underbrace{1, \ldots, 1}^{u_1}, \underbrace{2, \ldots, 2}^{u_2}, \ldots, \underbrace{m, \ldots, m}^{u_m}
\end{array}
\]

\[ u_1 + u_2 + \ldots + u_m = n \]

The term in \( u_k \) corresponds to the encoded messages of length \( K \). These terms can be considered as \( \sum_{i=1}^{n} D^{-l_i} \) when the summation takes place over all those terms with \( l_i = k \).

Therefore, by a simple re-assignment of terms, we may equivalently write

\[
\sum_{i=1}^{n} u_i D^{-i} = \sum_{i=1}^{n} D^{-l_i}
\]

(1.6.13)

Thus

\[
\sum_{i=1}^{n} u_i D^{-i} = \sum_{i=1}^{n} D^{-l_i} \leq 1
\]

The desired set of positive integers \([l_1, l_2, \ldots, l_n]\) must satisfy the inequality (1.6.5). This proves the necessity requirement of the theorem.

**Sufficient Part:**

Suppose now, that inequality (1.6.5) is satisfied for \([l_1, l_2, \ldots, l_n]\) then every summand of the left hand side of (1.6.5) being non negative, the partial sums are also at most 1.

\[
u_1 D^{-1} \leq 1, \quad \text{or} \quad u_1 \leq 1
\]

\[
u_1 D^{-1} + u_2 D^{-2} \leq 1, \quad \text{or} \quad u_2 \leq D^2 - u_1 D
\]

\[
\ldots
\]

\[
u_1 D^{-1} + u_2 D^{-2} + \ldots + u_n D^{-n} \leq 1,
\]
or 
\[ u_n \leq D^n - u_1 D^{n-1} - u_2 D^{n-2} - \ldots - u_{n-1} D \]

but these are exactly the conditions that we have to satisfy in order to guarantee that no encoded message can be obtained from any other by the addition of a sequence of letters of the encoding alphabet, thereof, which implies the existence of the instantaneous code.

**Remark:** For binary case the Kraft inequality tells us that the length \( l_i \) must satisfy the equation

\[ \sum_{i=1}^{n} 2^{-l_i} \leq 1 \]  
(1.6.14)

where the summation is over all the words of the block code.

**1.6.1 Lemma:** For a probability distribution \( P = (p_1, p_2, \ldots, p_n) \), \( p_i > 0 \), \( \sum_{i=1}^{n} p_i = 1 \) and incomplete distribution \( Q = (q_1, q_2, \ldots, q_n) \), \( q_i > 0 \), \( \sum_{i=1}^{n} q_i \leq 1 \) the following inequality holds

\[ -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} q_i \log q_i \]  
(1.6.15)

Before proving Lemma (1.6.1) we state the following lemma.

**1.6.2 Lemma:** If \( \psi \) is differentiable concave function in \((a, b)\), then for all \( x_i \in (a, b) \), \( i = 1, 2, \ldots, n \) and for all \( q_1, q_2, \ldots, q_n \), \( q_i > 0 \), \( \sum_{i=1}^{n} q_i \leq 1 \), \( i = 1, 2, \ldots, n \), the inequality holds

\[ \psi \left( \sum_{i=1}^{n} q_i x_i \right) \geq \sum_{i=1}^{n} q_i \psi (x_i). \]

Define the function

\[ L(X) = \begin{cases} -x \log x & \text{for } x \in (0, \infty) \\ 0 & \text{for } x = 0 \end{cases} \]

It is differentiable concave function of \( x \) on \([0, \infty)\) and continuous at 0 (from right), as
\[ \frac{\partial^2}{\partial x^2} (x \log x) > 0, \lim_{x \to 0} x \log x = 0 \log 0 = 0 \]

Putting \( x_i = \frac{p_i}{q_i}, i = 1, 2, \ldots, n \) in lemma 1.6.2, we get

\[ \sum_{i=1}^{n} q_i L \left( \frac{p_i}{q_i} \right) \leq L \left( \sum_{i=1}^{n} q_i \frac{p_i}{q_i} \right) \]

\[ L \left( \sum_{i=1}^{n} p_i \right) = L (1) = 0 \]

Thus

\[ 0 \geq - \sum_{i=1}^{n} q_i \frac{p_i}{q_i} \log \frac{p_i}{q_i} \]

\[ = - \sum_{i=1}^{n} p_i \left( \log p_i - \log q_i \right) \]

\[ = \sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} p_i \log q_i \]

or

\[ = - \sum_{i=1}^{n} p_i \log p_i \leq - \sum_{i=1}^{n} p_i \log q_i \]

**1.6.2 Theorem:** Let \( \{X\} \) be a discrete message source, without memory, and \( x_i \) be any message of this source with probability of transmission \( p_i \). If the \( \{X\} \) ensemble is encoded in a sequences of uniquely decipherable character taken from the alphabet \( (a_1, a_2, \ldots, a_n) \) then

\[ L = \sum_{i=1}^{n} p_i l_i \geq \frac{H(P)}{\log D} \quad (1.6.16) \]

**Proof:** The Condition \( L \geq \frac{H(P)}{\log D} \) is equivalent to

\[ \log D \sum_{i=1}^{n} p_i l_i \geq - \sum_{i=1}^{n} p_i \log p_i, \]
since \( p_i \log D = p_i \log D^L = -p_i \log D^{-l_i} \), the above condition may be written as

\[-\sum_{i=1}^{n} p_i D^{-l_i} \geq -\sum_{i=1}^{n} p_i \log p_i\]

we define \( q_i = -\frac{D^{-l_i}}{\sum_{i=1}^{n} D^{-l_i}} \), then \( q_i \)'s add to unity and Lemma 1.6.1 yields

\[-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log \left( \frac{D^{-l_i}}{\sum_{i=1}^{n} D^{-l_i}} \right) \]

(1.6.17)

with equality iff \( p_i = \frac{D^{-l_i}}{\sum_{i=1}^{n} D^{-l_i}} ; \forall i = 1, 2, \ldots, n. \)

Hence by (1.6.17).

\[
H(P) \leq -\sum_{i=1}^{n} p_i \log D^{-l_i} + \sum_{i=1}^{n} p_i \log \left( \frac{D^{-l_i}}{\sum_{i=1}^{n} D^{-l_i}} \right)
\]

\[
H(P) \leq L \log D + \log \left( \sum_{i=1}^{n} D^{-l_i} \right)
\]

With equality iff \( p_i = \frac{D^{-l_i}}{\sum_{i=1}^{n} D^{-l_i}} \) \( \forall i = 1, 2, \ldots, n. \)

By theorem 1.6.1, \( \sum_{i=1}^{n} D^{-l_i} \leq 1 \) which gives

\[
\log \left( \sum_{i=1}^{n} D^{-l_i} \right) \leq 0.
\]

Therefore,

\[ H(P) \leq \log D \]
\[ L \leq \frac{H(p)}{\log D} \]

### 1.6.3 Theorem:

Given a random variable \( X = (x_1, x_2, ..., x_n) \) having probability distribution \( P = (p_1, p_2, ..., p_n) \), where \( p_i \geq 0 \), with entropy (uncertainty) \( H(P) \), there exists a base \( D \), instantaneous code for \( X \), whose average code word length \( L = \sum_{i=1}^{n} l_i p_i \) satisfies

\[ \frac{H(p)}{\log D} \leq L \leq \frac{H(p)}{\log D} + 1 \quad (1.6.18) \]

**Proof:**

In general we cannot hope to construct an absolutely optimal code for a given set of probability \( P = (p_1, p_2, ..., p_n) \), since if we choose \( l_i \) to satisfy \( p_i = D^{-l_i} \) then \( l_i = \frac{-\log p_i}{\log D} \) may not be an integer. However we can do the next best thing and select the integer \( l_i \) such that

\[ \frac{-\log p_i}{\log D} \leq l_i = \frac{-\log p_i}{\log D} + 1, \quad i = 1, 2, ..., n \quad (1.6.19) \]

We claim that and instantaneous code can be constructed with word lengths \( l_i, l_2, ..., l_n \). To prove this we must show that \( \sum_{i=1}^{n} D^{-l_i} \leq 1 \).

For the left hand inequality of (1.6.19) it follows that

\[ \log p_i \geq l_i \log D \]

or

\[ p_i \geq D^{-l_i} \]

Thus

\[ \sum_{i=1}^{n} D^{-l_i} \leq \sum_{i=1}^{n} p_i \]

\[ \sum_{i=1}^{n} D^{-l_i} \leq 1; \quad \sum_{i=1}^{n} p_i = 1 \]
To estimate the average codeword length, we multiply $1.6.19$ by $p_i$ and sum over $i = 1, 2, \ldots, n$, to obtain

$$- \sum_{i=1}^{n} p_i \frac{\log p_i}{\log D} \leq \sum_{i=1}^{n} p_i l_i < - \sum_{i=1}^{n} p_i \frac{\log p_i}{\log D} + \sum_{i=1}^{n} p_i$$

$$\frac{H(p)}{\log D} \leq L \leq \frac{H(p)}{\log D} + 1.$$

### 1.7 Fuzzy Entropy:

The measure of uncertainty is adopted as a measure of information. Hence, the measure of fuzziness is known as fuzzy information measures. The measure of a quantity of fuzzy information gained from a fuzzy set or fuzzy system is known as fuzzy entropy.

A fuzzy subset ‘A’ in $U$ (universe of discourse) is characterized by a membership function $\mu_A : U \rightarrow [0,1]$ which represents the grade of membership of $x \in U$ in A as follows:

$$\mu_x(A) = \begin{cases} 
0, & \text{if } x \not\in A \text{ and there is no ambiguity} \\
1, & \text{if } x \in A \text{ and there is no ambiguity} \\
0.5, & \text{if maximum ambiguity, i.e., } x \in A \text{ or } x \not\in A
\end{cases}$$

In fact $\mu_A(x)$ associated with each $x \in U$, a grade of membership in the set ‘A’, when $\mu_A(x)$ is values in $\{0,1\}$, it is the characteristic function of a crisp (i.e. non-fuzzy) set.

A fuzzy set $A^*$ is called a sharpened version of A if the following conditions are satisfied:

$$\mu_{A^*}(x_i) \leq \mu_A(x_i) \text{, if } \mu_A(x_i) \leq 0.5; \quad \forall i$$

and

$$\mu_{A^*}(x_i) \geq \mu_A(x_i) \text{, if } \mu_A(x_i) \leq 0.5; \quad \forall i.$$

Since $\mu_A(x)$ and $1 - \mu_A(x)$ gives the same degree of fuzziness, therefore, corresponding to the entropy due to Shannon [87]. De Luca and Termini [33] suggested the following measure of fuzzy entropy:

$$H(A) = - \sum_{i=1}^{n} \left[ \mu_A(x_i) \log \mu_A(x_i) + (1 - \mu_A(x_i)) \log (1 - \mu_A(x_i)) \right]. \quad (1.7.1)$$
De Luca and Termini [33] introduced a set of four properties and these properties are widely accepted as a criterion for defining any new fuzzy entropy. In fuzzy set theory, the entropy is a measure of fuzziness which expresses the amount of average ambiguity/difficulty in making a decision whether an element belongs to a set or not. So, a measure of average fuzziness $H(A)$ in a fuzzy set should have at least the following properties to be valid fuzzy entropy:

i) **Sharpness**: $H(A)$ is minimum if and only if $A$ is a crisp sets, i.e. $\mu_A(x) = 0$ or $1; \forall x$.

ii) **Maximality**: $H(A)$ is maximum if and only if $A$ is most fuzzy set, i.e. $\mu_A(x) = 0.5; \forall x$.

iii) **Resolution**: $H(A) \geq H(A^*)$, where $A^*$ is sharpened version of $A$.

iv) **Symmetry**: $H(A) = H(A^c)$, where $A^c$ is the complement of $A$ i.e. $\mu_A(x_i) = 1 - \mu_A(x_i)$.

### 1.7.1 Entropy of a Fuzzy Event:

Let $x$ be a random variable which takes the values $\{x_1, x_2, ..., x_n\}$ with respective probabilities $p_1, p_2, ..., p_n$. Then, the entropy of the distribution $P = \{p_1, p_2, ..., p_n\}$ is given by

$$H(x) = - \sum_{i=1}^{n} p_i \log p_i \quad \text{(1.7.2)}$$

This definition suggests that entropy of fuzzy event, $A$, of the finite set $\{x_1, x_2, ..., x_n\}$ with respect to a probability distribution $P = \{p_1, p_2, ..., p_n\}$ be defined as follow

$$H^p(A) = - \sum_{i=1}^{n} \mu_A(x_i) p_i \log p_i \quad \text{(1.7.3)}$$

Where $\mu_A$ is the membership function of $A$. Since (1.7.1) expresses the entropy of a distribution $P$, (1.7.2) represents the entropy of a fuzzy event $A$ with respect to the distribution $P$. Thus, (1.7.1) does not reduce to (1.7.2) when $A$ is non-fuzzy, unless $A$ is taken to be the whole space $\{x_1, x_2, ..., x_n\}$. Intuitively, $H^p(A)$ may be interpreted as the uncertainty associated with a fuzzy event.

### 1.8 Fuzzy Reliability:

System failure engineering is primarily concerned with the failure and related problems. Specifically, by system failure engineering we mean the technological area comprising all failure oriented or failure driven aspects. So, it may compass reliability, safety, security, and so on. If everything went well and met desired requirements, then there would be no dissatisfaction, no failure, and therefore there would be no system failure engineering. Unfortunately, this is not the
case. Actually, failure is a nearly unavoidable phenomenon with technological products and systems. One can observe various kind of failure in various circumstances: space shuttle explosion, nuclear reaction accident, airplane crash, chemical plant leak, bridge break and electrical network collapse. One can also observe defective screw, faulty VLSI chip, error us management decision, and so on. Failures can be frequent or rare. The causes of failure are diverse. They can be physical, human, logical and even financial. The effects of failure may be minor or disastrous, and various kinds of criteria and factors can be taken into account to define what a failure means: structure, performance, cost and even subjective intention. However, whatever failure is, if the effect of it tends to be critical, research on it becomes essential.

In conventional reliability theory [18], it is assumed that components and systems have only two abrupt states: good and bad. This implies that the success and failure are precisely defined and there is no intermediate state between them. That is, the failure or success criterion is binary. Even in the research of multi-stat systems [10], the failure or success criterion is also assumed to be binary. In other words, in conventional reliability theory and multi-stat systems, it is assumed that the system states can be binary defined in terms of some structure function (e.g., coherent structure function) of component states. Needless to say, this assumption is valid in extensive cases.

However, the above assumption may not be true in every case. In degradable computing systems the attribute of performance degradation is prominent and should be taken into account in the failure or success criterion [29]. If we treat quality as a body of performance indices (static or dynamic), it is easy to see that quality can be factor of the failure or success criterion. This builds a bridge linking quality control and failure research. Further, it has been argued that other factors like cost, purchasability, etc., should also be taken into account in the definition of failure or success in some cases [86, 96]. After all, besides the structural factors, others like performance, quality, cost, etc., can make contributions to the failure or success criterion. This leads us to a general definition of failure or success.

Let \( A = \{a_1, a_2, \ldots, a_n\} \) be a set of factors of concern. Let

\[
x_{S_i}: a_i \rightarrow [0, 1]
\]

\[
x_{F_i}: a_i \rightarrow [0, 1]
\]
We call \( \{x_{si}\} \) success factor variables, and \( \{x_{Fi}\} \) failure factor variable. Let

\[
\mu_S = \mu_S(x_{s1}, x_{s2}, \ldots, x_{sn}) : [0, 1]^n \rightarrow [0, 1]
\]
\[
\mu_F = \mu_S(x_{F1}, x_{F2}, \ldots, x_{Fn}) : [0, 1]^n \rightarrow [0, 1]
\]

we call \( \mu_S \) (system) success variable or success membership function, and \( \mu_F \) (system) failure variable or failure variable function. Then system success \( S \) and system failure \( F \) are defined as fuzzy sets.

\[
S = \{(x_1, x_2, \ldots, x_n), \mu_S(x_1, x_2, \ldots, x_n)x_i \in [0, 1] \}
\]
\[
F = \{(x_1, x_2, \ldots, x_n), \mu_S(x_1, x_2, \ldots, x_n)x_i \in [0, 1] \}
\]

Since success and failure factor variables are defined on \( A \), we can also define \( S \) and \( F \) directly on \( A \), that is,

\[
S = \{(x_1, x_2, \ldots, x_n), \mu_S(x_1, x_2, \ldots, x_n)x_i \in A \}
\]
\[
F = \{(x_1, x_2, \ldots, x_n), \mu_S(x_1, x_2, \ldots, x_n)x_i \in A \}
\]

The generality of the above definition can be easily justified. In conventional reliability theory [18], we treat \( q_i \) as the \( i \)th component in a system, and \( x_{si} \) and \( x_{Fi} \) represent its states (0 or 1). Then \( \mu_S \) and \( \mu_F \) coincide with the corresponding system structure function. In a degradable computing system [29], \( a_1, a_2, \ldots, a_n \) can represent the system (non-fuzzy) states and the corresponding success (failure) factor variables represent the relative performance indices. Then \( \mu_S \) and \( \mu_F \) can be accordingly determined. For a software system, \( a_i \) can represent the \( i \)th module, and \( x_{si}(x_{Fi}) \) represents its quality index. Then \( \mu_S \) can be interpreted as a system quality variable. Alternatively, we can treat \( \{a_i\} \) as a set of quality factors such as correctness, reliability, efficiency, integrity, usability, maintainability, flexibility, portability, reusability, and so on. The factors in turn determine the quality variable \( \mu_S \). Anyway, defining failure and success as fuzzy sets enable them to be widely interpreted.
1.9 Measure and probability:-

1.9.1 Field and sigma (σ) field:

Let \( \Omega \) be a space of elements \( X \). A non empty class, \( \mathcal{R} \) of sets of \( \Omega \), closed under complementation and finite union is called a field. i.e. \( \mathcal{R} \) is a field if it satisfies the following axioms:-

i. \( \mathcal{R} \) is non empty.

ii. If \( A \in \mathcal{R} \) then \( A' \in \mathcal{R} \) where \( A' \) is the complement of \( A \) relative to \( \Omega \).

iii. If \( A_1, A_2, \ldots, A_n \in \mathcal{R} \) then \( \bigcup_{i=1}^{n} A_i \in \mathcal{R} \) if axiom (iii) is replaced by the axiom.

iv. If \( A_1, A_2, \ldots, A_n \in \mathcal{R} \) then \( \bigcup_{i=1}^{n} A_i \in \mathcal{R} \). then \( \mathcal{R} \) is called the \( \sigma \)-field.

Remark 1.9.1: It can be easily verified that the null set \( \phi \), the space and the countable intersection of sets of field also belongs to \( \mathcal{R} \).

1.9.2 Measurable set and measurable space:

The subset belonging to the \( \sigma \)-field \( \mathcal{R} \) are called measurable set.

The doublet \( (\Omega, \mathcal{R}) \) is called measurable space.

1.9.3 Measure and measure space:

The measurable space \( (\Omega, \mathcal{R}) \) indicates that this is the structure upon which a measure can be defined.

A real valued function \( \mu \) defined on \( (\Omega, \mathcal{R}) \) is called a measure if it satisfies the following axioms.

(i) \( \mu(\phi) = 0 \), where \( \phi \) is non-empty set.

(ii) \( \mu(A) \geq 0 \), for all \( A \in \mathcal{R} \).

(iii) if \( A_1, A_2, \ldots \) are disjoint measurable sets, then \( \mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i) \)

Property (iii) is called \( \sigma \)-Additivity.
Remarks:

i) A set of $\mu$-measure zero is said to be a $\mu$-null set and relations valid outside a $\mu$-null set are said to be valid almost everywhere $\mu$.

ii) If $\mu(\Omega)$ is finite then $\mu$ is said to be finite measure.

iii) A measure is said to be a $\sigma$-finite if the space $\Omega$ can be partitioned into a countable numbers of sets in $\mathcal{R}$, for each of which the value of $\mu$ is finite.

iv) The triplet $(\Omega, \mathcal{R}, \mu)$ is called measure space.

v) $\mu$ is called the finite measure on $(\Omega, \mathcal{R})$, if an addition the above axioms (i), (ii) and (iii), we have $\mu(\Omega)=1$, where $\Omega$ is the space of elementary events or sample space. A probability measure is usually denoted by $p$.

A probability space is the triplet $(\Omega, \mathcal{R}, p)$, formed by a sample space $\Omega$, a $\sigma$-field $\mathcal{R}$ defined on $\Omega$ and a probability measure $p$ defined on $(\Omega, \mathcal{R})$. All measure sets $A \in \mathcal{R}$ are called events.

Thus, with every event $A$ consisting of one or more outcomes of an experiment, we associated a numerical quantity, called the probability of $A$ denoted by $P(A)$ which will measure the chance that event $A$ will occur, we take $0 \leq P(A) \leq 1$.

1.9.4 Function:

If $X$ and $Y$ be two non empty sets, then a function $f$ from the set $X$ into set $Y$ is a correspondence (mapping) such that for each element of $X$, there exists only one element $Y$. This correspondence is generally denoted as $f: X \rightarrow Y$. if $x \in X$ and $y \in Y$ then $y$ is said to be a function of $x$ and we write $y=f(x)$.

1.9.5 Measurable function and random variable:

A real valued function $f(\cdot)$ defined on $\Omega$, the sample space is said to be an $\mathcal{R}$ measurable function or simply measurable function if for every real number $r$, $\{X: f(x) \leq r\} \in \mathcal{R}$.
If \((\Omega, \mathcal{F}, P)\) is a probability space, then a \(\mathcal{F}\) measurable function \(f(\cdot)\) is called a random variable.

### 1.10 Some Mathematical functions and Inequalities:

#### 1.10.1 Convex Function:

A real valued function \(f(x)\) defined on \((a, b)\) is said to be convex function if for every \(\alpha\) such that \(0 \leq \alpha \leq 1\) and for any two points \(x_1\) and \(x_2\) such that \(a < x_1 < x_2 < b\), we have

\[
f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)
\]

(1.10.1)

If we put \(\alpha = \frac{1}{2}\), then (1.10.1) reduces to

\[
f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f(x_1) + f(x_2)}{2}
\]

(1.10.2) which is taken as the definition of convexity.

**Remark 1.10.1:**

If \(f''(x) \geq 0\), then \(f(x)\) is convex function.

#### 1.10.2 Strictly Convex Function:

A real valued function \(f(x)\) defined on \((a, b)\) is said to be strictly convex function if for every \(\alpha\) such that \(0 < \alpha < 1\) and for any two points \(x_1\) and \(x_2\) in \((a, b)\) we have

\[
f[\alpha x_1 + (1-\alpha)x_2] < \alpha f(x_1) + (1-\alpha)f(x_2)
\]

(1.10.3)

**Remark 1.10.2:** If \(f''(x) > 0\), then \(f(x)\) is strictly convex function.

#### 1.10.3 Joint Convexity:

Let \(f(0, \infty) \rightarrow \mathbb{R}\) be a convex, then \(C_f(p, q)\) is jointly convex in \(p\) and \(q\), where \(p, q \in \mathbb{R}_+^n\).

#### 1.10.4 Concave Function:

A function \(f(x)\) is said to be concave if \(-f(x)\) is convex.

**Remark 1.10.3:** If \(f''(x) \leq 0\), then \(f(x)\) is concave function.
1.10.5 Strictly Concave Function:

A function \( f(x) \) is said to be strictly concave if \( -f(x) \) is strictly convex.

**Remark 1.10.4:** \( f''(x) < 0 \), then \( f(x) \) is strictly concave function.

1.10.6 Log - Concave Function:

A function \( f(x) \) is said to be log - concave if every \( \delta \),

\[
\frac{1}{2} \ln f(x - \delta) + \frac{1}{2} \ln f(x + \delta) \leq \ln f(x)
\]

If a density is log – concave, we can always assume that it is log – concave because densities are defined up to a set of measure zero.

1.10.7 Increasing Function:

Let \( I \) be an open interval contained in the domain of a real function. The function \( f(x) \) is an increasing function on \( I \) if \( x_1 < x_2 \) in \( I \), implies

\( f(x_1) \leq f(x_2) \).

1.10.8 Decreasing Function:

Let \( I \) be an open interval contained in the domain of a real function. The function \( f(x) \) is a decreasing function on \( I \) if \( x_1 < x_2 \) in \( I \), implies

\( f(x_1) \geq f(x_2) \).

1.10.9 Maximum of a Function:

A function \( f(x) \) is said to have a maximum value in an interval \( I \) at \( x \), if \( f(x) \geq f(x) \) for all \( x \) in \( I \).

1.10.10 Minimum of a Function:

A function \( f(x) \) is said to have a minimum value in an interval \( I \) at \( x \), if \( f(x) \leq f(x) \) for all \( x \) in \( I \).

The following theorems give the working rule for finding the points of local maxima or points of local minima.
1.10.11 Some inequalities:

i) Jensen’s inequality:

If $X$ is a random variable such that $E(X) = \mu$ exists and $f(x)$ is a convex function, then

$$E[f(X)] \geq f(E(X))$$

with equality iff the random variable $X$ has a degenerate distribution at $\mu$.

The following important concept is due to Csiszar and Korner [32].

Let: $f : (0, \infty) \to \mathbb{R}$ be a convex function. Then for any $p, q \in \mathbb{R}^n$ with

$$p_n = \sum_{i=1}^{n} p_i > 0, \quad Q_n = \sum_{i=1}^{n} q_i > 0,$$

we have the inequality

$$C_f(p, q) \geq Q_n f\left(\frac{p_n}{q_n}\right).$$

The equality sign holds iff

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} = \frac{p_3}{q_3} = \ldots = \frac{p_n}{q_n}.$$

In particular, for all we have

$$C_f(p, q) \geq f(1)$$

With equality iff $P = Q$.

ii) Holder’s Inequality:

If $x_i, y_i > 0, \ i = 1, 2, \ldots, n$ and $\frac{1}{p} + \frac{1}{q} = 1, \ p > 1, \ q > 1$, then the following inequality holds

$$\sum_{i=1}^{n} x_i y_i \leq \left[\sum_{i=1}^{n} x_i^p\right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} y_i^q\right]^{\frac{1}{q}}.$$

iii) Chebychev’s Inequality:

If $X$ is a random variable with mean $\mu$ and variance $\sigma^2$, then for any positive number $k$
\[ P\{\left| X - \mu \right| \geq k\sigma \} \leq \frac{1}{k^2} \]

Or

\[ P\{\left| X - \mu \right| < k\sigma \} \geq 1 - \frac{1}{k^2} \]

iv) **Bienaymé - Chebychev’s Inequality:**

Let \( g(x) \) be a non-negative function of a random variable \( X \), then for any \( k > 0 \),

\[ P\{g(x) \geq k\} \leq \frac{E[g(x)]}{k} \]

v) **Markov’s Inequality:**

If we take \( g(x) = |x| \) in inequality (iv), then

\[ P\{|x| \geq k\} \leq \frac{E[|x|]}{k} \]

which is Markov’s Inequality.

Taking, \( g(x) = |x|^r \) and replacing \( k \) by \( k'^r \) in inequality (iv), we get a more generalized form of Markov’s inequality.

\[ P\{|x|^r \geq k'^r\} \leq \frac{E[|x|^r]}{k'^r} \]

vi) **Log Sum Inequality:**

For non-negative numbers \( a_1, a_2, \ldots, a_n \) and \( b_1, b_2, \ldots, b_n \) the log sum inequality is given as

\[ \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} \frac{a_i}{\sum b_i} \sum_{i=1}^{n} a_i \log \frac{a_i}{\sum b_i} \]

with equality, iff \( \frac{a_i}{b_i} = k \) where \( k \) is a constant.