CHAPTER - I

MEANING AND SCOPE OF THE INFORMATION

Introduction :

In every-day life we get information by various means such as radio, television, newspapers, books and others. Literally, the term 'information' is used to mean some change in the state of mind or in the previous knowledge about some event. The amount of change in the knowledge that occurs upon getting the information (equivalent to observing the random variable X from the point of view of obtaining inferences about unknown parameter θ) is measured through a function that is called the entropy.

This chapter contains four sections. The section 1.1 deals with basic concepts of information through communication process and terminologies used in the same process are defined. Then the motivation to the formula of the measure of information is used. Based upon some desirable properties of a measure of information it shows that the Shannon's entropy is a suitable measure to measure the information. In the section 1.2 the Shannon's entropy is defined and some properties of this entropy are given. The section 1.3 deals with the principle of the maximum information in which under certain conditions the unknown distribution which maximises the entropy is found out. The section 1.4 deals with other measure of information in which a function which follows certain properties of entropy is considered as a 'measure of information'. In this section we have discussed the Hartley's entropy, the Renyi entropy, the generalised entropies. The Kullaback and Leibler information measure and the Fisher's information measure as measure of information.

1.1 Basic Concepts :

In every-day life various means of transmission of information are available. Some agencies like Radio, news-papers, books, television and telegraph transmit the information.

We know that radio-broad cast in which the information is transmitted in the form of waves and radio transmits these waves into the sound. News-papers gives various sorts of information of day-to-day event happenings in the world. The knowledge that we get from the books is also an information about the scientific and other developments. The television-broadcasts sound waves and picture waves are transmitted through the T.V. sets. It is quite evident that through television much more information can be given as compared to radio.

Suppose A wants to pass some news (say I) by a device suitable to him to 'B', 'C' and 'D'. Previously B knows all about 'I', 'C' a little about 'I' and 'D' does not know anything about 'I'. After getting the news 'I', 'B' does not aware of 'I', 'C' aware of 'I' but little, and 'D' is greatly surprised.

Observe that 'I' is not at all informative to 'B' while it is little informative to 'C' and is very much informative to 'D'. Further it depends upon how 'A' passes this news. In general, the mind (that of 'A') affects another mind (may be of 'B', 'C', 'D', [illegible]). This procedure of one mind affecting other mind is called a communication procedure.

A communication model means a device by which the message from the source is given to the receiver. Here, consider simple communication model :

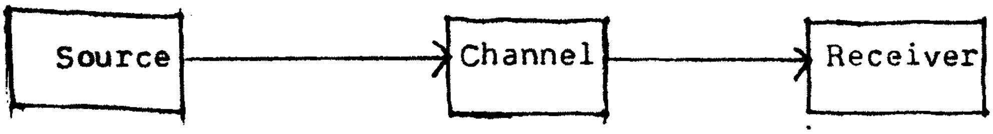

Fig.I

But practically there are a number of sources and receivers', hence clear-cut transmission of information is not possible. For example, in telephone system we can hear many sounds, therefore, sometimes we cannot understand what the message is.

Suppose that there is a noisy street and your friend says something which you cannot hear at all or you hear differently. In this way proper transmission of the information may not be possible; or even a great interruption may occur. To avoid such disturbances, the transmission of the information must be done efficiently. The communication model is not deterministic nature but it is a probabilistic one. In the telegraph system, the message from the message-sheet is transformed into the different language by which it can be transmitted. The process of transforming a message in the different language which is suitable for transmission is called coding; such coded message is transmitted to the receiver. In this system, this coded message is transformed into the original form. It is called decoding. Suppose the message 'A' is transmitted by a source. Let p(A) be the probability of receiving the message 'A' from the source. Let p(A\A) be the conditional probability of getting the message 'A' at receiver through channel given that the message passed to the receiver is 'A'. The probability of transmitting the message 'A' and also getting the message 'A' by receiver is given as :

$$p(A) \quad p(A|A)$$

By adding encoder and decoder in the previous model (Fig.I), the efficiency of the model can be increased. The improved model is given as below :

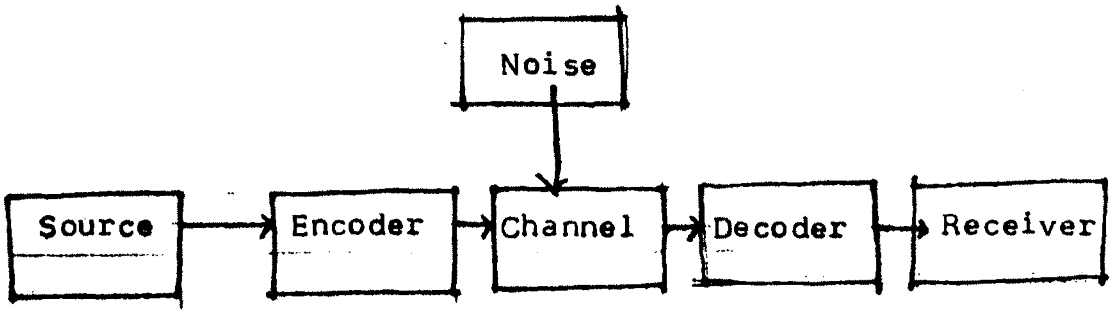

Fig.II

In this model the efficiency of the transmission could be improved.

Now, we describe different terms in the model.

1) Source : The source is defined as the agency which gives messages or sequence of symbols of the given alphabets.

2) Encoder : It is a device which transforms a message into the specific codes and the codes can be transmitted.

3) Channel : The media in which an encoded message is transmitted to the decoder.

4) Decoder : A decoder is a device which converts coded message into the original form and passes it on to receiver.

5) Noise : The interruption, disturbance, resistance that occurs in the channel which affects the coded message during the course of its transmission from encoder to decoder is called as noise e.g. i) In the newspapers misprints or improper printing can be considered as noise (here, printed matter in the news-paper serves as channel). ii) Consider an example of a transport company, suppose some goods are to be transported by trucks, the bad condition of road and truck will decrease the efficiency of transportation (here, the road or truck is considered as channel).

Suppose the result of an experiment is awaited. And the experiment results in one of k mutually exclusive and exhaustive outcomes; $E_1, E_2, \ldots, E_k$. Let 'A', 'B' and 'C' are three persons who are interested in the result of the experiment with their belief (prior knowledge about the result of the same experiment). Let $p_A(E_r)$ be the probability of the outcome E_r according to the prior knowledge of 'A', similarly $p_B(E_r)$, $p_C(E_r)$ be defined.

Further let, $p_A(E_1) = p_A(E_2) = \ldots = P_A(E_k)$ i.e. according to 'A', all possible outcomes are equally likely and let,

$$p_B(E_1) > p_B(E_2) > \ldots > p_B(E_k)$$

i.e. according to 'B' E_r is more likely than E_{r+1} $(r = 1,2,\ldots k-1)$. Let

$$p_C(E_1) < p_C(E_2) < \ldots < P_C\ (E_k)$$

i.e. E_{r+1} is likely than E_r.

Suppose the experiment results in the outcome E_1. Note that 'B' is not at all surprised of the result, while 'C' is very much surprised. And 'A' says the event has occurred as per his guess. Basically, the 'information' that an event E_1 gives to the persons depends on their prior knowledge about that specific event E_1. If $p_A(E_1) = 1$, the event E_1 is a sure event according to the prior knowledge of the person 'A'. Further, the occurrence of event E_1 does not give any additional information as such, so that his state of knowledge about the experiment would not change after the occurrence of the event E_1.

Suppose the experiment E results in the outcomes of the form (x_i, y_j) $i = 1,2,\ldots m$; $j = 1,2,\ldots n$.

Further let,

$$p_{ij} = p\ [X = x_i,\ Y = y_j]$$

and

$$p_i = \sum_{i=1}^{n} p_{ij} = p[\ X = x_i\ ];$$

$$q_{ji} = p_{ij} \backslash p_{i.}$$

the conditional probability of the event $[Y{=}y_j \mid X = x_i]$. Let p_{ij} $i{=}1,2,\ldots,m;\ j{=}1,2,\ldots n$ be the prior distribution of a person 'A' in the experiment. Thus, $p_{i.}$ is the belief in the partial result $X = x_i$ of the experiment and q_{ji} is the distribution about the secondary result $Y = y_j$ of the experiment given the primary result $X = x_i$. The information that the person 'A' gathers from the outcome $[\ X{=}x_i\ \ Y{=}y_j\ ]$ must be the addition of the information gathered from :

1) The partial observation $X = x_i$ and

2) The secondary observations $[\ Y{=}y_i$ given $X{=}x_j\ ]$

We shall call this property as additivity of information. If $f(p_{ij})$ be a suitable measure of the 'amount of information contained' in the outcome $[\ X = x_i,\ \ Y = y_j\ ]$. The $f(\cdot)$ is a decreasing function on the interval (0,1) with

$$f(1) = 0 \qquad (1.1.1)$$

and

$$f(p_{ij}) = f(p_i) + f(q_{ji}) \qquad (1.1.2)$$

Where,

$$p_{ij} = p_i . q_{ji}$$

The solution of function 'f' satisfying the relation (1.1.2) is of the form $f(x) = k \log x$. Further, this function satisfies properties (1.1.2) and (1.1.1). And k must be arbitrary and it is chosen as -1 for our convenience. Thus, a suitable measure to measure the amount of information contained in the event E which occurs with probability p is $-\log_2 p$. The "amount of information" contained in the experiment $\mathcal{E}$ which results in one of the mutually exclusive, exhaustive outcomes E_1, E_2, ... with $p_i = p(E_i)$ is defined to be

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

and to be denoted by $H(p_1, p_2, \ldots, p_n)$.

This can be interpreted as 'the average information contained the experiment $\mathcal{E}$'. The unit of information is termed as 'bit'.

1.2. The Shannon's Entropy :

In this section we will define the Shannon's entropy and prove some of its properties.

Consider, the telegraphic system, there n messages on the message-sheet are $x_1, x_2, \ldots, x_n$ with transmission probabilities $p_1, p_2, \ldots, p_n$ (the source selects particular message x_k with probability p_k). The amount of information (I_k) associated with x_k is given by :

$$I_k = - \log_2 p_k \qquad (1.2.1)$$

Then, the average information (I) per message is given by

$$H_n = - \sum_{k=1}^{n} p_k \log_2 p_k \qquad (1.2.2)$$

The average information per message is called the entropy.

Shannon (1948) defined the entropy as given below :

Definition 1.2.1 : Consider the discrete probability distribution p_i such that

$$\sum_{i=1}^{n} p_i = 1. \quad i = 1,2,\ldots,n$$

The shannon's entropy $H_n(p_1,p_2,\ldots,p_n)$ is defined as

$$H(p) = H_n(p_1,p_2,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Interpretations :

1) It is the expected value of $-\log_2 p(x)$, where X is a. r. v. such that $p(X) = p_i$

2) It is considered as "the measure of uncertainty contained in the experiment before the result of experiment has occurred.

3) If Θ is a parameter (in many standard experimental setup for example, i) Binomial random variable ii) poisson random variable, the probability of different events or function of Θ is called the parameter $p(E_i) = p_i(\Theta)$ for $i=1,2,\ldots,n$) which is unknown, $H(\Theta)$ is considered as "the amount of missing information in Θ".

Example 1.2.1 : Consider two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, where $\mathcal{E}_1$ has two events A_1 and A_2 with probabilities p_{11} and p_{12}. There also two events of the experiment $\mathcal{E}_2$ which are B_1 and B_2 with probabilities p_{21} and p_{22} respectively. Let

$$p_i = (p_{i1}, p_{i2}) \quad i = 1,2$$

If

$$p_{11} = p \quad p_{12} = 1 - p = q$$

and

$$p_{21} = q \,, \quad p_{22} = p$$

The amount of information missing in the both experiments is same according to the $H(P_1) = H(P_2)$.

$$[\ H(p_{11},p_{12}) = -p\ \log_2 p - (1-p)\ \log_2(1-p) = H(p_{21},p_{22})$$

In particular,

$$p_{11} = p_{12} = 1/2,\ p_{21} = 1/4,\ p_{22} = 3/4$$

Then we have

$$H(p_{21},p_{22}) < H(p_{11},p_{22})$$

$$\text{i.e.}\quad 2 - \log_2^3 \quad < \quad \log_2^2$$

In the following, we use the term 'the amount of information contained in' and the 'amount of information missing in' with the same meaning. An appropriate term is used as per context.

Properties : The Shannon's entropy satisfies following properties :

1) Symmetry :

$$H_n(p_1,p_2,\ldots,p_n) = H_n(p_{k(1)},p_{k(2)},\ldots p_{k(n)}) \qquad (1.2.3)$$

Where, k is arbitrary permutation on $(1,2,\ldots,n)$.
This means the amount of information does not change when order of event is changed.

2) Normality : $H_2(1/2,\ 1/2) = 1$ (1.2.4)

If experiment has equally likely two cases the amount of information in the experiment is unity.

3) Expansibility :

$$H_{n+1}(p_1, p_2, \ldots, p_n, 0) = H_n(p_1, p_2, \ldots p_n) \tag{1.2.5}$$

If the additional outcome added in the experiment with zero probability, the amount of information in the experiment would not change.

4) Decisivity : $H_2(1,0) = H_2(0,1) = 0$ (1.2.6)

The amount of information corresponds to the sure event is zero.

5) Recursivity :

$$H_n(p_1, p_2, \ldots, p_n) = H_{n-1}(p_1 + p_2,\ p_3,\ \ldots,\ p_n) + \\ + (p_1+p_2)H_2(p_1/(p_1+p_2), p_2/(p_1+p_2)) \tag{1.2.7}$$

Where, $p_1 + p_2 > 0$.

If an experiment E' is derived from the experiment E by clubbing different events of the experiment E. The amount of information missing in E' will increase.

<u>Proof</u> : Let, $p = p_1 + p_2$ $\quad q = p_2/(p_1 + p_2)$

$1 - q = p_1/(p_1 + p_2)$, $p_2 = pq$.

i.e. $p_1 = p(1-q)$

$$H_n(p(1-q), pq, p_3,\ldots,p_n) = -p(1-q) \log_2 p(1-q) - $$

$$- pq \log_2 pq + H(p_3,p_4,\ldots,p_n)$$

$$H_n(p(1-q), pq, p_3,\ldots,p_n)$$

$$= -p(1-q)\log_2 p-p(1-q)\log_2(1-q)-pq \log_2 p -$$

$$-pq \log_2 q + H(p_3,p_4,\ldots,p_n).$$

$$= -p \log_2 p - \sum_{k=1}^{n} p_k \log_2 p_k - p[(1-q)\log_2(1-q) + + q \log_2 q]$$

$$= H_{n-1} (p,p_3,\ldots,p_n) + pH_2(1-q,q)$$

$$= H_{n-1}(p,p_3\ldots,p_n)+(p_1+p_2)H_2(p_1/(p_1+p_2), p_2/(p_1+p_2))$$

6) Maximality :

For any probability distribution

$p_i \geqslant 0$ $(i=1,2,\ldots n)$ and $\sum_{i=1}^{n} p_i = 1.$

$$H_n(p_1,p_2,\ldots,p_n) \leqslant H_n(1/n,1/n,\ldots 1/n) \qquad (1.2.8)$$

This shows that the amount of information missing in any distribution is less than the amount of information missing in uniform distribution. i.e. the entropy maximizes for the uniform distribution.

Consider the joint experiment A and B and their outcomes are not independent. Then the entropy of combined experiment is defined as

$$H_{mn}(A \otimes B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log p_{ij} \qquad (1.2.9)$$

$$i = 1,2,\ldots n$$

$$j = 1,2,\ldots m$$

Define :

$$\left.\begin{array}{l} p_i = \sum_{j=1}^{m} p_{ij},\ q_j = \sum_{i=1}^{n} p_{ij} \\ \\ p(i/j) = p_{ij} / q_j\ ,\quad p(j/i) = p_{ij}/p_i \end{array}\right\} \qquad (1.2.10)$$

Now, we shall give some definitions.

Definition 1.2.2 : The conditional entropy calculated for experiment B under assumption that event a_i of experiment A is happened.

$$\text{i.e. } H_m(B \mid ai) = -\sum_{j=1}^{m} p(j \mid i) \log_2 p(j \mid i) \qquad (1.2.11)$$

Definition 1.2.3 : The conditional entropy of the experiment B given A is

$$H_m(B/A) = \sum_{i=1}^{n} p_i H(B|ai) \qquad (1.2.12)$$

Similarly, we have

$$H_n(A/b_j) = -\sum_{i=1}^{n} p(i/j) \log_2 p(i/j) \qquad (1.2.13)$$

$$H_n(A|B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} q_j p(i|j) \log_2 p(i/j) \qquad (1.2.14)$$

Proposition 1.2.1 :

$$H_{mn}(A \otimes B) = H_n(A) + H_m(B|A)$$

$$= H_m(B) + H_n(A/B) \qquad (1.2.15)$$

Proof :

$$H_{mn}(A \otimes B) =$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log_2 p_{ij}$$

$$H_{mn}\ (A \otimes B)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(i|j)q_j \ \log_2 p(i|j)q_j \ -\sum_{i}\sum_{j} p(i|j)q_j \log \ q_j$$

$$= \sum_{j}\sum_{i} q_j p(i/j) \ \log_2 p(i/j) - \sum_{j} q_j \ \log_2 q_j$$

$$= \ H(A/B) + H_m(B)$$

$$= \ H_m(B) + H(A|B)$$

Where,

$$p_{ij} = p(i/j)q_j$$

Similarly,

$$H_{mn}(A \otimes B) = H_n(A) + H_m \ (B|A)$$

<u>Definition 1.2.4</u> : For any two experiments

$$H_n(A) - H_n(A|B) = H_m(B) - H_m(B|A) \qquad (1.2.16)$$

is known as the 'information balance'.

Let us consider the experiments A and B are independent, then

$$p_{ij} = p_{i.} q_j$$

$$H_m(B|A) = H_m(B) \quad \text{and} \quad H_n(A|B) = H_n(A)$$

Therefore,

$$H_{mn}(A \otimes B) = H_n(A) + H_m(B) \qquad (1.2.17)$$

Proposition 1.2.2 : For any two experiments A and B we have

$$H_m(B|A) \lessapprox H_m(B) \qquad (1.2.17)$$

Proof : We will use here Jensen's inequality for concave function. Let g(x) be real valued function which is concave defined on the interval $[r, s]$. Let $x_1, x_2, \ldots, x_n \in [r, s]$ and c_i such that $\sum_{i=1}^{n} c_i = 1$.

$$\text{i.e.} \sum_{i=1}^{n} c_i g(x_i) \lessapprox g\left(\sum_{i=1}^{n} c_i x_i\right) \qquad (1.2.18)$$

In the equation (1.2.18) put $r = 0$ $S = 1$,
$g(x) = -x \log_2 x$ $c_i = p_i$ $x_i = q(j|i)$

$$-\sum_{i=1}^{n} p_i q(j|i) \log_2 q(j|i) \lessapprox -\sum_{i=1}^{n} p_i q(j|i) \log_2\left(\sum_{i=1}^{n} p_i q(j|i)\right)$$

$$= -q_j \log q_j$$

$$\text{i.e.} \sum_{j=1}^{m} \sum_{i=1}^{n} p_i q(j|i) \log_2 q(j|i) < -\sum_{j=1}^{m} q_j \log_2 q_j$$

i.e. $H_m(B|A) \lessapprox H_m(B)$.

Hence, the proof.

1.3 Principle of Maximum Information :

The Shannon's entropy is a natural and most useful measure of information and discrete uniform distribution has maximum entropy. According to Laplace's 'Principle of insufficient Reason'; two events are to be assigned equal probabilities, if there is no reason to think otherwise. Thus, when nothing is known except that there are n possible outcomes according to Laplece's principle. the probability of an outcome will be 1/n. This is a distribution having more uncertainty than any other distribution.

However, when some thing about the probability distribution is known under Some constraints one would like to maximize the entropy. Thus, when partial information is available (say through constraints) we must use the distribution which has maximum Shannon's entropy subject to the given constraints. Thus the only unbiased assignment that one can do, otherwise it would to measure some more information being known. Now we shall prove here theorem.

Theorem 1.3.1 : Let X be a random variable having density f(x) such that $x \in (a,b)$. Let $k_1, k_2, \ldots$ be integrable functions on (a,b) satisfying the condition

$$\int_a^b k_i(x)f(x)dx = m_i \quad i = 1,2... \tag{1.3.1}$$

Where, m_i are known constants.

The density of the form

$$f(x) = \exp[a_0 + a_1k_1(x)+a_2k_2(x)+...] \tag{1.3.2}$$

maximises the entropy. The constants a_0, a_1,... are determined such that f(x) satisfies the conditions i) It must be density function (ii) (1.3.1).

<u>Proof</u> : Let f(x) be of the form

$$f(x) = \exp\ [a_0+a_1k_1(x) + a_2k_2(x)+ ...]$$

Let g(x) be another density function satisfying condition (1.3.1)

Therefore,

$$\int g(x)\ \log_e [g(x)/f(x)]\ dx \geq 0$$

$$\text{i.e.}\quad \int g(x)\ \log_e g(x)\ dx \geq \int g(x)\ \log_e f(x)\ dx$$

$$-\int g(x)\ \log_e g(x)dx \leq -\int g(x)\ \log_e f(x)\ dx$$

$$= -\int g(x)[a_0+a_1k_i(x)+...]dx$$

$$= -\ (a_0+a_1m_1+a_2m_2+\ ...) \tag{1.3.4}$$

This shows that

$$-(a_0 + a_1 m_1 + a_2 m_2 + \ldots)$$

is fixed upperbound for $-\int g(x) \log_e g(x)\, dx$.

If g(x) be density function and satisfies the condition (1.3.1). The upperbound is obtained when g(x) is chosen of the form of (1.3.2) with finding suitable constants $a_0, a_1, \ldots$ (Kagan and Linnik, Rao)

Mathematically, the maximum entropy distribution has the important property that no possibility is ignored; it assigns positive weight to every situation that is not absolutely excluded by the given information. The density function which maximizes entropy is found out by using Lagrangian multiplier method.

Now we shall give following example.

Example 1.3.1 : Let the support of X be (0,10)

$$\int_0^{10} f(x)\, dx = 1$$

Maximizes the function

$$H_n = -\int_0^{10} f(x) \log_e f(x)\, dx$$

Subject to condition

$$E(x) = 6$$

Using Lagrangian multiplier method, we have

$$H_n - \alpha \int_0^{10} f(x)dx - \beta \int_0^{10} xf(x)\ dx$$

$$= \int_0^{10} f(x)[\log(1/f(x) - \alpha - \beta x]\ dx$$

$$= \int_0^{10} f\ (x)\ \log_e\ [\exp\ (-\alpha-\beta x)/f(x)]\ dx$$

$$\leq \int_0^{10} f(x)\ \log_e\ [(\exp\ (\alpha-\beta x)/f(x))-1]\ dx$$

because if $x \neq 1.$, $\log x \leq x-1$

Equality holds if and only if

$$f(x) = \exp\ (-\ \alpha\ -\ \beta x) \quad .$$

Where α and β are Lagrangian multipliers.

To determine α and β using equation

$$\int_0^{10} f(x)\ dx = 1$$

$$\int_0^{10} \exp\ (-\alpha-\beta x)\ dx = 1$$

$$\int_0^{10} \exp\ (-\beta x) dx = \exp\ (\alpha)$$

$$\alpha = \log_e \int_0^{10} \exp\ (-\beta x)\ dx$$

Using the condition

$$\int_{o}^{1o} xf(x)dx = 6$$

$$\int_{o}^{1o} x \exp(-\beta x)dx \Big/ \int_{o}^{1o} \exp(-\beta x)dx = 6$$

$$\int_{o}^{1o} x \exp(-\beta x)dx = 6 \int_{o}^{1o} \exp(-\alpha x)dx$$

$$\text{i.e. } -(4\beta+1) \exp(-10\beta) = (6\beta-1)$$

$$\exp(-10\beta) = (1-6\beta) / (4\beta+1)$$

Therefore,

$$\beta = -0.123$$

$$f(x) = \exp(-\beta x) \Big/ \int_{o}^{1o} \exp(-\beta x)$$

$$= \beta \exp(-\beta x) / (1 - \exp(-10\beta)$$

$$\text{i.e. } f(x) = 0.0508006 \exp(0.123x)$$

Therefore,

$$-\int_{o}^{1o} f(x) \log_e f(x)dx = -0.738 - \log_e 0.050806$$

In the above example, the density function f(x) which maximizes function H_n under E(x) = 6 is given as

$$f(x) = 0.0508006 \exp(0.123x)$$

i.e. the f(x) has nature like equation (1.3.2).

In the following table we shall give supports of random variable X, restriction and corresponding density functions.

Table - 1

Set of values of r.v.	Restrictions	Density function corresponding to maximum entropy
(0 1)	-	$f(x) \doteq 1$
(0 1)	$E \log_e x = g_1$ $E \log_e(1-x)=g_2$	$f(x)=x^{m-1}(1-x)^{n-1}/ \beta(m\ n)$
(0 ∞)	$Ex = g_1$	$f(x) = a \exp(-ax)$
(0 ∞)	$Ex = g_1$ $E \log x = g_2$	$f(x)=a^p \exp(-ax)x^{p-1}/\Gamma p$
(-∞ oo)	$Ex = g_1$ $Ex^2 = g_2$	$f(x)=\exp[-(x-\mu)^2/2\sigma^2]/\sigma\sqrt{2\pi}$
(-∞ ∞)	$E\ x = g_1$	$f(x) = (a/2)\exp(-a\lvert x\rvert)$

From above it is clear that all density functions (given above) have same nature as (1.3.2) (exponetial type).

1.4 Other Measures of Information :

In the previous section we discussed Shannon's entropy. This is a logarithmic function which satisfies certain properties such as symmetry, expansibility and subadditivity or additivity. These are 'essential' and 'natural' properties of an entropy as a measure of information.

There are some other logarithmic functions which also have above properties; considered as measure of information. Some specific measure can be chosen depending on the problem of interest. In the following we discuss some of the measures of information.

I) Hartley's Entropy :

Hartley (1928) first introduced a measure of information. His intention was to store information by some instrument. Wherein one needs to consider how much information can be stored in each store. For this suppose a storage unit (such as knob) has n possible states, then r such storage units put together provide 'n^r' states. Thus duplication of the storage units can be used as a strong criterion to increase the storage capacity.

Moreover, maintaining through r storage units each of n states is easier than that of maintaining through a single storage unit of n states.

As the number of states depends on the number of storage units exponentially, for fixed n, the storage capacity (total number of states) depends on r, the duplication of storage units. Thus basically, the storage capacity can be measured on the basis of the values of r that is

$$C = \log_2 N \qquad (1.4.1)$$

the capacity of storage which can store N distinguishable states.

Here C is called the Hartley's entropy.

Remark : Hartley's measure can be viewed as when all N states are equally probable.

Hence,

$$C = H_N\ (1/N,\ \ldots,\ 1/N) \geq H_n(p_1,p_2,\ldots,p_n) \qquad (1.4.2)$$

Following is a characterisation of the Hartley's entropy.

Corollary 1.4.1 : If and only if an entropy is weakly subadditive, additive, symmetric, normalised and intensive, then it is Hartley's entropy (J.Aczel, B. Forte and C.T. Ng (1974)).

The linear combination of the Shannon's entropy and the Hartley's entropy is again an entropy. Let $H_n(p_1,p_2,\dots,p_n)$ be the Shannon's entropy and C be Hartley's entropy, then,

$$H\ (p_1,p_2,\dots,p_n) = a\ H_n(p_1,p_2,\dots,p_n) + b.C \quad (1.4.3)$$

(Aczel et al (1974))

II) The Renyi's Entropy :

Here, we will define the entropy which depends on its order. The order α (constant) entropy is considered as Renyi entropy.

Definition 1.4.1 : The Renyi entropy of order $\alpha \neq 1$ of the probability distribution $(p_1,p_2,\dots,p_n) \in O_n$ i=1,2,..n is defined as

$$_\alpha H_n(p_1,p_2,\dots,p_n) = (1/1-\alpha)\log_2 \sum_{i=1}^{n} p_i^\alpha \quad (1.4.4)$$

With $0^\alpha = 0$ Where $O_n = \left\{(p_1,p_2,\dots p_n) \middle| \sum_{i=1}^{n} p_i = 1 \text{ i.e. } i=1,2,\dots,n\right\}$

Where α is real.

Remarks : 1) When $\alpha \longrightarrow 1$ the Renyi entropy tends to the Shannon entropy.

i.e.

$$\lim_{\alpha\to 1} \alpha\ H_n(p_1,p_2,\dots,p_n)$$

$$= \lim_{\alpha\to 1} [(1/1-\alpha)\log_2 \sum_{i=1}^{n} p_i^\alpha]$$

$$= \lim_{\alpha\to 1} [\sum_{i=1}^{n} p_i^\alpha \log_e p_i \ / \ (-\sum_{i=1}^{n} p_i^\alpha \log_e 2) \ ]$$

$$= -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$= H_n(p_1,p_2,\dots,p_n)$$

2) If $\alpha = 0$ the Renyi entropy tends to the Hartley entropy.

3) The Renyi entropies are symmettic, normalized, expansible, decisive, additive, non-negative, measurable. If $\alpha \geqslant 0$ they are also maximal, bounded and monotonic, if $\alpha > 0$ they are continous small for small probabilities. They are also subadditive for $\alpha = 0$ and $\alpha = 1$.

The Renyi entropies are not recursive.

III) The Generalized Entropies :

In the previous article we have discussed the entropies of order α. Now we will define the class of entropies of degree $\alpha \neq 1$.

Definition 1.4.2 : The entropies of degree $\alpha \neq 1$ are given as :

$$H_n^{\alpha}(p_1,p_2,\dots,p_n) = (2^{1-\alpha}-1)^{-1}\Big(\sum_{i=1}^{n} p_i^{\alpha} - 1\Big).\alpha \neq 1 \qquad (1.4.5)$$

Where $p_1, p_2, \dots, p_n \in O_n$, $n = 2,3,\dots$

With $0^{\alpha} = 0$, where $O_n = \Big\{(p_1,p_2,\dots,p_n)/\sum_{i=1}^{n} p_i=1,\ p_i \geq 0,\ i=1,2,..n\Big\}$

Remark : If $\alpha \to 1$ the generalized entropies tends to the Shannon entropy.

$$\lim_{\alpha\to 1} H_n^{\alpha}(p_1,p_2,\dots,p_n) = \lim_{\alpha\to 1}(2^{1-\alpha}-1)^{-1}\Big(\sum_{i=1}^{n} p_i^{\alpha} - 1\Big)$$

$$= \lim_{\alpha\to 1}\Big(\sum_{i=1}^{n} p_i^{\alpha}\log p_i\Big)/(-2^{(1-\alpha)}\log 2)$$

$$= -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$= H_n(p_1,p_2,\dots p_n)$$

In the following we give the properties of generalized entropies without proof.

Theorem 1.4.1 : The entropies H_n^{α}; $O_n \to R$ $(n=2,3\dots)$ of degree α are symmetric, normalised, expansible, decisive, non-negative measurable and have sum properties.

The entropies non-negative degree ($\alpha \geqslant 0$) are also maximal, bounded and monotonic and of positive degree are continuous and small for small probabilities. Those of degree $\alpha \geqslant 1$ are also subadditive. These H_n^α entropies also follow the following properties.

i) Additivity of degree α :

$$H_{mn}^\alpha (p_1 q_1, p_2 q_2, \ldots p_m q_1, p_m q_2 \ldots p_n q_n)$$
$$= H_m^\alpha (p_1, p_2, \ldots, p_m) + H_n^\alpha (q_1, q_2, \ldots, q_n) +$$
$$(2^{1-\alpha}-1) H_m^\alpha (p_1, p_2, \ldots, p_m) H_n^\alpha (q_1, q_2, \ldots, q_n) \tag{1.4.6}$$

for all $(p_1, p_2, \ldots, p_m) \in O_m \quad (q_, , q_2, \ldots, q_n) \in O_n$

ii) Strong additivity :

$$H_{mn}^\alpha (p_1 q_{11}, p_1 q_{12}, \ldots, p_1 q_{1n} \ldots, p_m q_{m1}, p_m q_{m2} \ldots p_m q_{m2})$$
$$= H_m^\alpha (p_1, p_2, \ldots p_m) + \sum_{j=1}^{m} p_j^\alpha H_n^{\ \alpha} (q_{j1}, q_{j2} \ldots q_{jn}) \tag{1.4.7}$$

for all $(p_1, p_2, \ldots, p_m) \in O_m, (q_{j1}, q_{j2}, \ldots q_{jm}) \in O_n$

$j = 1, 2, \ldots m \quad m = 2, 3, \ldots \quad n = 2, 3, \ldots$

Recursivity of degree α

$$H_n^\alpha (p_1, p_2, \ldots, p_n) = H_{n-1}^\alpha (p_1+p_2, p_3, \ldots, p_n) +$$
$$+ (p_1+p_2)^\alpha H_2^\alpha [p_1/(p_1+p_2), p_2/(p_1+p_2)] \tag{1.4.8}$$

for all $(p_1,p_2,\dots,p_n) \in O_n \quad n = 2,3,\dots.$

with $p_1 + p_2 > 0$

Note that meaning of properties has been given in the previous section (1.2).

IV) The Kullback and Leibler Information :

Measure :

The Kullback and Leibler information measure is used for the purpose of testing statistical hypotheses. Let $(S \ \mathbb{F} \ p_i)$ i=1,2 be measure spaces, where p_1,p_2 are absolutely continuous probability measures and dominated by measure λ. Let X be a random variable and $H_i(x)$ i=1,2 be hypotheses that X has distribution p_i, i= 1,2.

According to Bayes' theorem

$$p(H_i| x)=[p(H_i)f_i(x)]/[p(H_1)f_1(x)+p(H_2)f_2(x)] \qquad (1.4.9)$$

Where $f_i(x)$ is Radon Nikodym derivative.

$$f_i(x) = dp_i/ d\lambda$$

$$\log_e[f_1(x)/f_2(x)] = \log_e[p(H_1| x)/p(H_2| x)] -$$

$$- \log_e[p(H_1)/p(H_2)] \quad [\lambda] \ (1.4.10)$$

Where $p(H_i)$ is the prior probability and $p(H_i|x)$ is the conditional or posterior probability. And

$$\log_e(f_1(x)\ /\ f_2(x))$$

is defined to be the information.

i.e. It is the information that the observation x contains for discrimination in favour of H_1 against H_2. Further, the mean infomation for discrimination in favour of H_1 against H_2 given that $x \in A$ from p_1 is given as

$$\begin{aligned} I(1:2;A) &= \int_A \log_e [(f_1(x)/f_2(x)) dp_1(x)] / p_1(A) \\ &= \int_A \log_2(f_1(x)/f_2(x)) f_1(x)\ d\lambda\ /\ p_1(A) \\ &= 0 \qquad\qquad \text{if } p_1(A) = 0 \qquad\qquad (1.4.11) \end{aligned}$$

Where, $x \in A$ $\in$FF $dp_1(x) = f_1(x) d\lambda(x)$.

Let A be the entire sample space S, I(1:2;S) i.e. the mean information for discrimination in favour of H_1 against H_2 per observation from p_1 is given as

$$\begin{aligned} I(1:2;,S) = I(1:2) &= \int \log_e(f_1(x)/f_2(x)) dp_1(x) \qquad (1.4.12) \\ &= \int \log_e(f_1(x)\ /\ f_2(x)) f_1(x)\ d\lambda\ (x) \\ &= \int \log_e\ [p(H_1|x)\ /\ p(H_2|x)]\ dp_1\ - \\ &\quad - \log_e\ [\ p(H_1)\ /\ p(H_2)\ ]. \end{aligned}$$

I(1:2) is the difference between the mean value with respect to p_1 of the logarithm of posterior probability of the hypotheses and the logarithm of the prior probability.

I(1:2) is also called as the information of p_1 with respect to p_2.

Also the mean information for discrimination H_2 against H_1 is defined as

$$I(2:1) = \int f_2(x)\ \log_e(f_2(x)/f_1(x))\ d\lambda(x)$$

$$= -\int f_2(x)\log_e(f_1(x)/f_2(x))\ d\lambda(x) \qquad (1.4.13)$$

The necessary condition for I(1:2) and I(2.:1) be finite is

$$p_1 \equiv P_2$$

<u>Example 1.4.1</u> : $f_1(x) \longrightarrow U(0,1)$, $f_2(x) \longrightarrow U(0\ \ 2)$

$$I(1:2) = \int_0^1 \log_e 2\ d_x + \int_1^2 0\cdot\log_e 0\ dx.$$

$$= \log_e{}^2$$

$$I(2:1) = \int_0^1 \log_e 1/2\ dx + \int_1^2 \log\ (1/2/0).1/2\ dx$$

$$= \log_e 1/2 + \infty \quad (\text{using } \log_e\ (C/0) = \infty)$$

(Where C is constant)

$$= \infty$$

i.e. I(1:2) is finite but I(2:1) is infinite.

Example 1. 4.2 :

$$\text{for } f_1(x) \longrightarrow U(0 \quad 2)$$

$$f_2(x) \longrightarrow U(1, 3)$$

$$I(1:2) = \int_0^1 1/2.\log(1/2/0)dx + \int_2^1 1/2\log_e[1/2/1/2]dx$$

$$+ \int_2^3 \log(0/1/2)\ 0\ dx$$

$$= \infty + 0 + 0$$

$$= \infty$$

Similarly,

$$I\ (2:1) = \infty$$

From above examples it is clear that although $p_1 \equiv p_2$ the I(1:2) and I (2:1) may be infinite.

In general, we can observe that

$$I\ (1:2) \neq I\ (2:1)$$

and hence, I(1:2) or I(2:1) cannot be used as a measure of divergence between H_1 and H_2 .

Jeffery (1946, 1948, p.158) introduce the measure of divergence between H_1 and H_2 which is defined as follows :

Definition 1.4.3 :

$$J(1\ \ 2) = \int [f_1(x)-f_2(x)]\ \log_e(f_1(x)/f_2(x))\ d\lambda(x)$$

$$= I(1:2) + I(2:1)$$

$$= \int \log_e\ [p\ (H_1|\ x)/p(H_2|x)]dp_1(x) -$$

$$- \int \log_e[p(H_1/x)/p(H_2/x)]\ dp_2(x) \qquad (1.4.14)$$

This measure of divergence is symmetric. This is not a metric, since traingular inequality property is not satisfied.

Now we will consider some properties of k - L information measure.

Theorem 1.4.2 : Let X and Y be independent random variables under H_i, i = 1,2 then

$$I(1:2;X,Y) = I(1:2;X) + I\ (1:2,Y) \qquad (1.4.15)$$

Proof :

$$I(1:2\ X,Y)= \int f(x\ y)\log_e[f_1(x\ y)/f_2(x\ y)]\ d\lambda\ (x\ y)$$

$$= \int g_1(x)h_1(y)\log_e[g_1(x)h_1(Y)/g_2(x)h_2(y)]d\mu(x),\ d\nu(Y)$$

$$= \int g_1(x)\log_e\ [g_1(x)/g_2(x)]d\mu(x) +$$

$$+ \int h_1(y)\log \frac{h_1(Y)}{h_2(y)} d\nu(y)$$

Where, $f_i(x\ y) = g_i(x)h_i(Y) \quad i = 1,2$

$$d\lambda(x\ Y) = d\mu(x).\ d\nu(y), \quad \int f_i(x)\,d\mu(x) = 1.$$

$$\int h_1(y)\ d\nu\ (y) = 1.$$

Theorem : 1.4.3 : If X and Y are not independent

$$I(1:2;\ X,Y) = I(1:2;X) + I(1:2;\ Y/X)$$

$$= I(1:2;Y) + I(1:2;\ X/Y) \qquad (1.4.16)$$

$I(1:2;\ Y|X=x)$ is the conditional information of Y for the discrimination in favour of H_1 against H_2 when $X = x$ under H_1. i.e. $I\ (1:2;\ Y|X)$ is the mean value of the conditional discrimination information undeŕ H_1.

Theorem 1.4.4 : The I(1:2) is positive, i.e. $I(1:2) \geq 0$. If equality holds if and only if $f_1(x) = f_2(x)$ a.s. $[\lambda]$.

Proof : Let us define $g(x) = f_1(x)/f_2(x) \qquad (1.4.17)$

$$I(1:2) = \int f_1(x)\log_e[f_1(x)/f_2(x)]\,d\lambda(x)$$

$$= \int f_2(x)\ g(x)\ \log_e[f_2(x).g(x)/f_2(x)]\,d\lambda(x)$$

$$= \int g(x)\ \log_e g(x)\ d\ p_2(x).$$

Where $d\ p_2(x) = f_2(x)d\lambda(x)$.

Define function $\Psi(t) = t\log_e t$ with $0 < g(x) < \infty[\lambda]$

Expanding by Taylor's series,

5748

A

$$\psi(g(x)) = \psi(1) + [g(x)-1]\psi'(1)+1/2[g(x)-1]^2\psi''(h(x))[\lambda] \qquad (1.4.17)$$

i.e. $g(x) < h(x) < 1$ and $0 < h(x) < \infty$.

With $\psi(1) = 0$, $\psi'(1) = 1$

$$\int g(x)\, dp_2(x) = \int f_1(x)\, d\lambda(x) = 1 \qquad (1.4.18)$$

and $$\int \psi(g(x))\, dp_2(x) = 1/2\int [g(x)-1]^2\psi''(h(x))\, dp_2(x) \qquad (1.4.19)$$

Where, $\psi''(1) = 1/t > 0$ for $t > 0$

i.e.

$$\int g(x)\log_e g(x)\, dp_2 = \int f_1(x) \log_e [f_1(x)/f_2(x)]\, d\lambda(x) \geq 0$$

If $f_1(x) = f_2(x)$ a.s. $[\lambda]$ (1.4.20)

Therefore $I(1:2) = 0$

Theorem 1.4.4 tells us that, the mean discrimination information obtained from statistical observations is positive And also there is no discrimination information if the distributions of the observations are same under both hypotheses.

V) The Fisher's Measure of Information :

This measure is used in the problems of estimation of unknown parameter θ. It is used in obtaining minimum variance unbaised estimate of statistic 'T' i.e. Cramer-Rao - inequality.

Let X be a random variable with sample space S and having probability density function $f(x,\Theta)$ with respect to a σ- finite measure μ and $f(x,\Theta)$ is differentiable with respect to Θ. Then, for any measurable set A,

$$(d/d\Theta) \int_A f(x\ \Theta)\ d\mu = \int_A (d/d\,\Theta)\ [f(x, \Theta)]\ d\mu \qquad (1.4.21)$$

It can be observed that

$$\begin{aligned} E\ [(d/d\Theta)[\log_e f]] &= 0 \text{ and } E\left\{-(d^2/d\Theta^2)[\log_e f]\right\} \\ &= E\left\{(d/d\Theta)[\log_e f]\right\}^2 \\ &= E\ [\ f'(x\ \Theta)/f(x\ \Theta)\ ]^2 \\ &= V\ [(d/d\Theta)(\log_e f)] \qquad (1.4.22) \end{aligned}$$

<u>Definition 1.4.4</u> : The Fisher's information measure on Θ contained in the random variable X, is defined to be

$$E\ [\ (d/d\Theta)\ (\log_e f)\ ]^2$$

and is denoted by $I(\Theta)$.

The properties of this measure are given as below:

1) Let $I_1(\Theta)$ and $I_2(\Theta)$ are information contained in two independent variables X and Y. Let $I(\Theta)$ be the information contained in joint (X Y) then,

$$I(\Theta) = I_1(\Theta) + I_2 (\Theta) \qquad (1.4.23)$$

2) Let $X_1, X_2, \ldots, X_n$ be identically distributed random variables and $I(\Theta)$ be the information contained in each variable. Then, the information contained in $(X_1, X_2, \ldots, X_n)$ is $nI(\Theta)$.

3) Let X be a random sample, $X = (x_1, x_2, \ldots x_n)$ independently, identically distributed with $f(x\ \Theta)$. And T be a measurable function of X with a density function $\psi(\cdot, \Theta)$ with respect to a σ- finite measure ν. And it is differentiable with respect to Θ. Let $I_T(\Theta)$ be the information contained in the statistic T about Θ,

$$I_T(\Theta) = E\ [\psi'(T\ \Theta) / \psi(T, \Theta)\ ]^2 \qquad (1.4.24)$$

Observe that

a) $E\ [f'(x,\Theta)/f(x,\Theta) | \ T=t\ ] = \psi'(t,\Theta)/\psi(t,\Theta)$

(Rao (1973))

b) $I(\Theta) \geqslant I_T(\Theta)$ (1.4.25)

The result (b) is equivalent to

$$I(\Theta) \geqslant I_T(\Theta) \quad (x_1, x_2, \ldots x_n)$$

That is the information contained in a sample is greater than or equal to any statistic T. It will be shown in the Section (2.1). If equality in b holds, the statistic T is sufficient statistic.

Remark : If Θ is vector, the result is similar to b can be established in terms of the Fisher information matrix.