Chapter one

1- Random Variables

A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous. All random variables have a ***cumulative distribution function***. It is a function giving the probability that the random variable *X* is less than or equal to *x*, for every value *x*.

1-1 Discrete Random Variables

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,..... If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the number of defective light bulbs in a box of ten. The ***probability distribution*** of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

When the sample space Ω has a finite number of equally likely outcomes, so that the discrete uniform probability law applies. Then, the probability of any event *x* is given by:

$$P(A) = \frac{Number\ of\ elements\ of\ x}{Number\ of\ elements\ of\ \Omega}$$

This distribution may also be described by the ***probability histogram.*** Suppose a random variable *X* may take *k* different values, with the probability that $X = x_i$ defined to be $P(X = x_i) = P_i$. The probabilities P_i must satisfy the following:

1- $0 < P_i < 1$ for each i

2- $P_1 + P_2 + \cdots + P_k = 1$ or,

$$\sum_{i=1}^{k} P_i = 1$$

Example

Suppose a variable X can take the values 1, 2, 3, or 4. The probabilities associated with each outcome are described by the following table:

Outcome	1	2	3	4
Probability	0.1	0.3	0.4	0.2

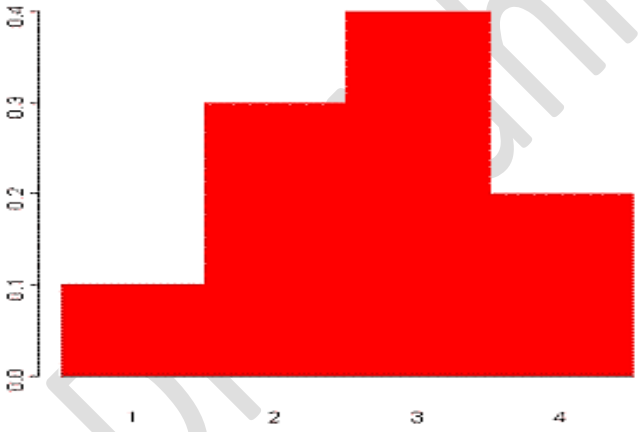


Figure 1: probability distribution

The cumulative distribution function for the above probability distribution is calculated as follows:

The probability that X is less than or equal to 1 is 0.1,

the probability that X is less than or equal to 2 is 0.1+0.3 = 0.4,

the probability that X is less than or equal to 3 is 0.1+0.3+0.4 = 0.8, and

the probability that X is less than or equal to 4 is 0.1+0.3+0.4+0.2 = 1.

H.W: Having a text of (ABCAABDCAA). Calculate the probability of each letter, plot the probability distribution and the cumulative distribution.

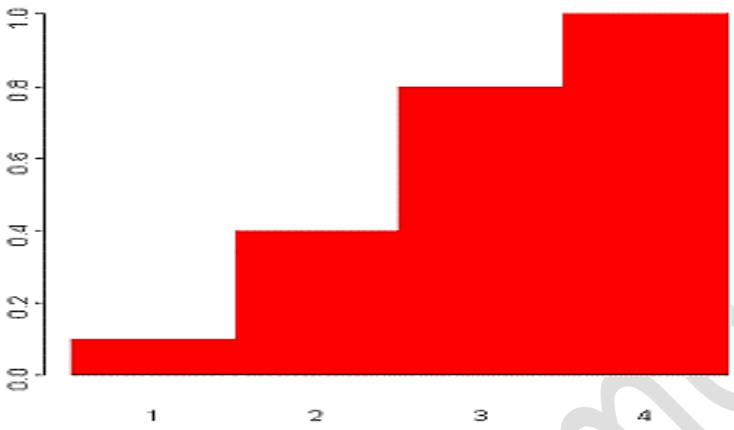


Figure 2: cumulative distribution

1-2 Continuous Random Variables

A ***continuous random variable*** is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight and the amount of sugar in an orange. A continuous random variable is not defined at specific values. Instead, it is defined over an *interval* of values, and is represented by the ***area under a curve***. The curve, which represents a function *p(x)*, must satisfy the following:

1: *The curve has no negative values ($p(x) \geq 0$ for all x)*

2: *The total area under the curve is equal to 1.*

A curve meeting these requirements is known as a ***density curve***. If any interval of numbers of equal width has an equal probability, then the curve describing the distribution is a rectangle, with constant height across the interval and 0 height elsewhere, these curves are known as uniform distributions.

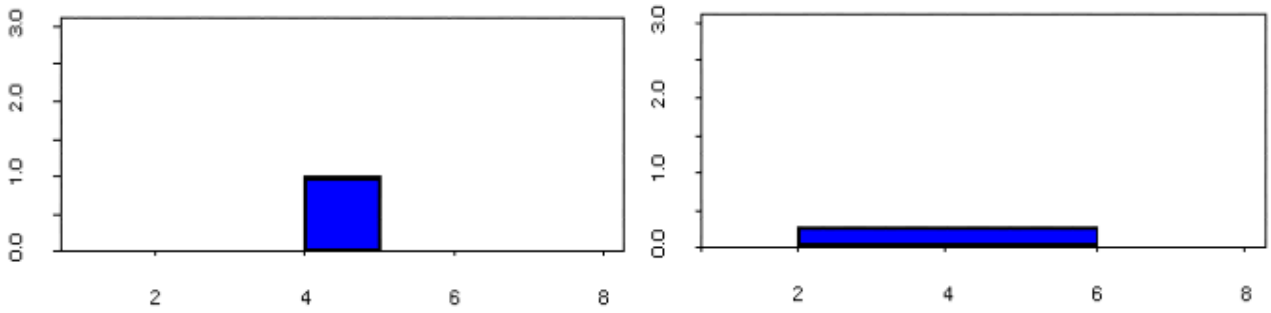


Figure 3: Uniform distribution

Another type of distribution is the normal distribution having a bell-shaped density curve described by its mean μ and standard deviation σ. The height of a normal density curve at a given point x is given by:

$$h = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2}$$

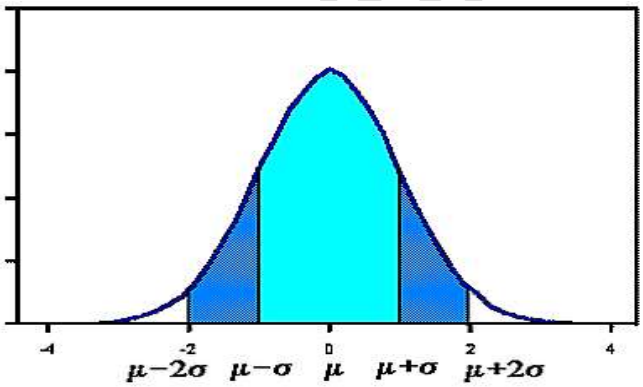


Figure 4: The Standard normal curve

2- Joint Probability:

Joint probability is the probability of event Y occurring at the same time event X occurs. Its notation is $P(X \cap Y) or\ P(X, Y)$, which reads; the joint probability of X and Y.

$$P(X, Y) = P(X) \times P(Y)$$

If X and Y are discrete random variables, then $f(x,y)$ must satisfy:

$$0 \le f(x,y) \le 1 \text{ and,}$$

$$\sum_x \sum_y f(x,y) = 1$$

If X and Y are continuous random variables, then $f(x,y)$ must satisfy:

$$f(x,y) \ge 0 \quad \text{and,}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) = 1$$

Example:

For discrete random variable, if the probability of rolling a four on one die is $P(X)$ and if the probability of rolling a four on second die is $P(Y)$. Find $P(X,Y)$.

Solution:

We have $P(X) = P(Y) = 1/6$

$$P(X,Y) = P(X) \times P(Y) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = 0.0277 = 2.8\%$$

3- Conditional Probabilities:

It is happened when there are dependent events. We have to use the symbol "|" to mean "given":

- P(B|A) means "Event B **given** Event A has occurred".
- P(B|A) is also called the "Conditional Probability" of B given A has occurred .
- And we write it as

$$P(A \mid B) = \frac{number\ of\ elements\ of\ A\ and\ B}{number\ of\ elements\ of\ B}$$

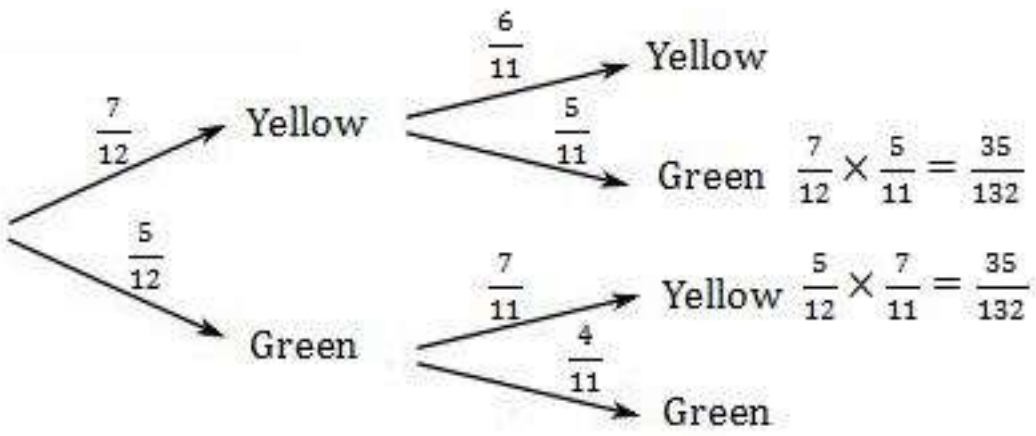
Or

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Where $P(B) > 0$

Example: A box contains 5 green pencils and 7 yellow pencils. Two pencils are chosen at random from the box without replacement. What is the probability they are different colors?

Solution: Using a tree diagram:



4- Bayes' Theorem

Bayes' theorem: an equation that allows us to manipulate conditional probabilities. For two events, A and B, Bayes' theorem lets us to go from p(B|A) to p(A|B).

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad P(A) \neq 0$$

$$P(A \cap B) = P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

$$P(A \mid B) = P(B \mid A) \times \frac{P(A)}{P(B)} \qquad P(B) \neq 0$$

Example:

If $P(X = 0) = 0.2, P(X = 1) = 0.3, P(X = 2) = 0.5, P(Y = 0) = 0.4$ *and* $P(Y = 1) = 0.6$. Determine $P(X = 0 \mid Y = 0)$, $P(X = 1 \mid Y = 0)$

5- Independence of Two Variables:

The concept of independent random variables is very similar to independent events. If two events A and B are independent, we have *P(A,B)=P(A)P(B)=P(A∩B)*. For example, let's say you wanted to know the average weight of a bag of sugar so you randomly sample 50 bags from various grocery stores. You wouldn't expect the weight of one bag to affect another, so the variables are independent.

6- Venn's Diagram:

A Venn diagram is a diagram that shows *all* possible logical relations between a finite collections of different sets. These diagrams depict elements as points in the plane, and sets as regions inside closed curves. A Venn diagram consists of multiple overlapping closed curves, usually circles, each representing a set. The points inside a curve labelled *S* represent elements of the set *S*, while points outside the boundary represent elements not in the set *S*. Fig. 5 shows the set $A = \{1, 2, 3\}, B = \{4, 5\}$ *and* $U = \{1, 2, 3, 4, 5, 6\}$.

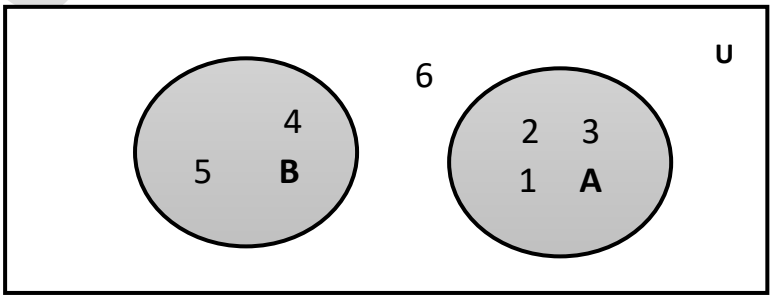


Figure 5: An example of Venn's Diagrams

Example:

From the adjoining Venn diagram of Fig. 6, find the following sets:

A, B, C, X, A', B', C-A, B-C, A ∪ B, A ∩ B

$(B \cup C)'$….

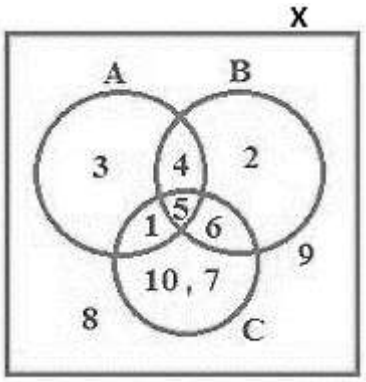


Figure 6:Venn's Diagram

Solution:

$A = \{1,3,4,5\}, B = \{2,4,5,6\}, C = \{1,5,6,7,10\}$

$X = \{1,2,3,4,5,6,7,8,9,10\}$

$A' = \{2,6,7,8,9,10\}$,

$B' = \{1,3,7,8,9,10\}$,

$(C - A) = \{3,4,6,7,10\}$,

$(B - C) = \{1,2,4,7,10\}$,

$(A \cup B) = \{1,2,3,4,5,6\}$,

$(A \cap B) = \{4,5\}$,

$(B \cup C)' = \{3,8,9\}$……..

7- Model of information transmission system

Transmitting a message from a transmitter to a receiver can be sketched as in Fig. 7:

The components of information system as described by Shannon are:

1. An information source is a device which randomly delivers symbols from an alphabet. As an example, a PC (Personal Computer) connected to internet is an information source which produces binary digits from the binary alphabet {0, 1}.
2. A source encoder allows one to represent the data source more compactly by eliminating redundancy: it aims to reduce the data rate.

3. A channel encoder adds redundancy to protect the transmitted signal against transmission errors.

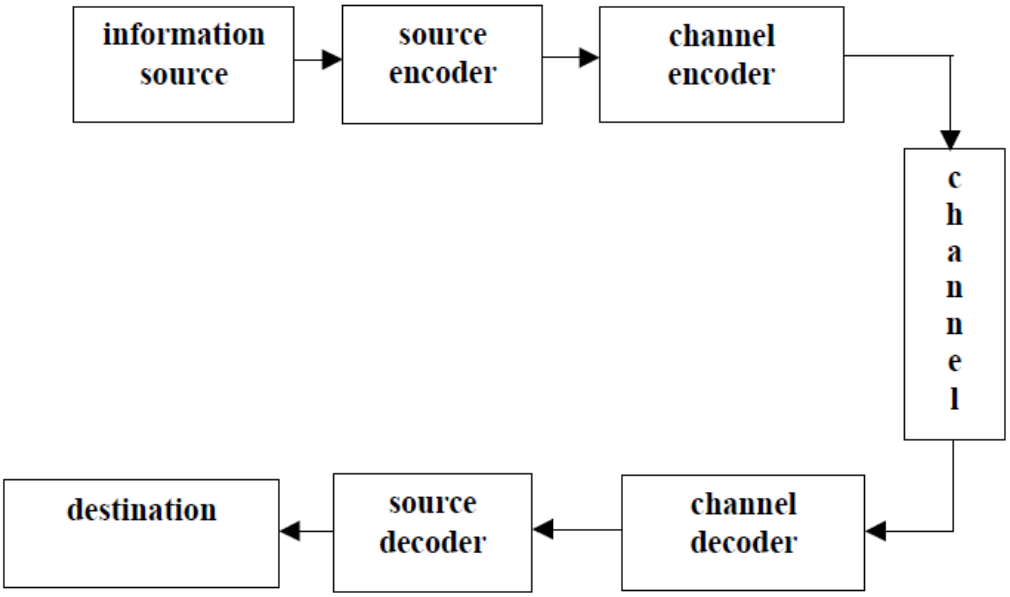


Figure 7: Shannon paradigm

4. A channel is a system which links a transmitter to a receiver. It includes signaling equipment and pair of copper wires or coaxial cable or optical fiber, among other possibilities.
5. The rest of blocks is the receiver end, each block has inverse processing to the corresponding transmitted end.

8- Self- information:

In information theory, **self-information** is a measure of the information content associated with the *outcome* of a random variable. It is expressed in a unit of information, for example bits, nats, or hartleys, depending on the base of the logarithm used in its calculation.

A **bit** is the basic unit of information in computing and digital communications. A bit can have only one of two values, and may therefore be physically implemented with a two-state device. These values are most commonly represented as 0 and 1.

A **nat** is the **natural unit of information**, sometimes also **nit** or **nepit**, is a unit of information or entropy, based on natural logarithms and powers of *e*, rather than the powers of 2 and base 2 logarithms which define the bit. This unit is also known by its unit symbol, the nat.

The **hartley** (symbol **Hart**) is a unit of information defined by International Standard IEC 80000-13 of the International Electrotechnical Commission. One hartley is the information content of an event if the probability of that event occurring is 1/10. It is therefore equal to the information contained in one decimal digit (or dit).

1 Hart ≈ 3.322 Sh ≈ 2.303 nat.

The amount of self-information contained in a probabilistic event depends only on the probability of that event: the smaller its probability, the larger the self-information associated with receiving the information that the event indeed occurred as shown in Fig.8.

i- Information is zero if $P(x_i) = 1$ (certain event)

ii- Information increase as $P(x_i)$ decrease to zero

iii- Information is a +ve quantity

The log function satisfies all previous three points hence:

$$I(x_i) = \log_a \frac{1}{P(x_i)} = -\log_a P(x_i)$$

Where $I(x_i)$ is self information of (x_i) and if:

i- If "a" =2 , then $I(x_i)$ has the unit of bits

ii- If "a"= e = 2.71828, then $I(x_i)$ has the unit of nats

iii- If "a"= 10, then $I(x_i)$ has the unit of hartly

Recall that $\log_a x = \frac{lnx}{lna}$

Example 1:

A fair die is thrown, find the amount of information gained if you are told that 4 will appear.

Solution:

$$P(1) = P(2) = \cdots \ldots \ldots . = P(6) = \frac{1}{6}$$

$$I(4) = -\log_2\left(\frac{1}{6}\right) = \frac{\ln(\frac{1}{6})}{ln2} = 2.5849 \ \ bits$$

Example 2:

A biased coin has P(Head)=0.3. Find the amount of information gained if you are told that a tail will appear.

Solution:

$$P(tail) = 1 - P(Head) = 1 - 0.3 = 0.7$$

$$I(tail) = -\log_2(0.7) = -\frac{ln0.7}{ln2} = 0.5145 \ \ bits$$

HW

A communication system source emits the following information with their corresponding probabilities as follows: A=1/2, B=1/4, C=1/8. Calculate the information conveyed by each source outputs.

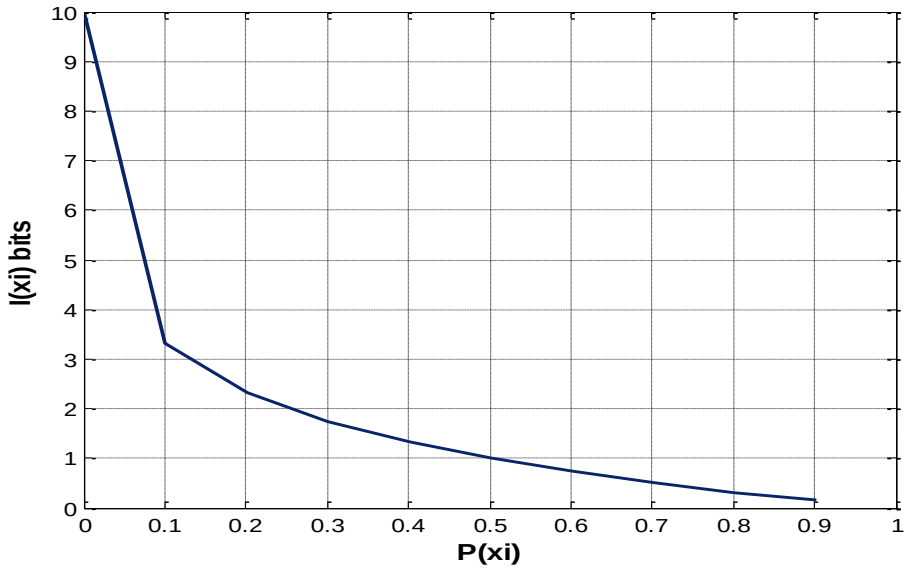


Figure 8: Relation between probability and self-information

9- Average information (entropy):

In information theory, **entropy** is the average amount of information contained in each message received. Here, *message* stands for an event, sample or character drawn from a distribution or data stream. Entropy thus characterizes our uncertainty about our source of information.

9-1 Source Entropy:

If the source produces not equiprobable messages then $I(x_i), i = 1, 2, \dots\dots, n$ are different. Then the statistical average of $I(x_i)$ over i will give the average amount of uncertainty associated with source X. This average is called source entropy and denoted by $H(X)$, given by:

$$H(X) = \sum_{i=1}^{n} P(x_i)\, I(x_i)$$

$$\therefore \quad H(X) = -\sum_{i=1}^{n} P(x_i) \log_a P(x_i)$$

Example:
Find the entropy of the source producing the following messages:

$$Px_1 = 0.25, \qquad Px_2 = 0.1, \;\; Px_3 = 0.15, \;\; and \; Px_4 = 0.5$$

Solution:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_a P(x_i)$$

$$= -\frac{[0.25ln0.25 + 0.1ln0.1 + 0.15ln0.15 + 0.5ln0.5]}{ln2}$$

$$H(X) = 1.7427 \; \frac{bits}{symbol}$$

9-2 Binary Source entropy:

In information theory, the **binary entropy function**, denoted or H(X) or $H_b(X)$, is defined as the entropy of a Bernoulli process with probability *p* of one of two values. Mathematically, the Bernoulli trial is modelled as a random variable *X* that can take on only two values: 0 and 1:

$$P(0_T) + P(1_T) = 1 \rightarrow P(1_T) = 1 - P(0_T)$$

We have:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_a P(x_i)$$

$$H_b(X) = -\sum_{i=1}^{2} P(x_i) \log_a P(x_i)$$

Then:

$$H_b(X) = -[P(0_T) \log_2 P(0_T) + (1 - P(0_T)) \; \log_2(1 - P(0_T))] \; bits/symbol$$

$$\text{If } P(0_T) = 0.2, then \; P(1_T) = 1 - 0.2 = 0.8, and \; put \; in \; above \; equation,$$

$$H_b(X) = -[0.2 \log_2(0.2) + 0.8 \log_2(0.8)] = 0.7$$

9-3 Maximum Source Entropy:

For binary source, if $P(0_T) = P(1_T) = 0.5$**,** then the entropy is**:**

$$H_b(X) = -[0.5 \log_2(0.5) + 0.5 \log_2(0.5)] = -\log_2\left(\frac{1}{2}\right) = \log_2(2) = 1 \; bit$$

Note that $H_b(X)$is maximum equal to 1(bit) if: $P(0_T) = P(1_T) = 0.5$**,** the entropy of binary source or any source having only two value is distributed as shown in Fig.9

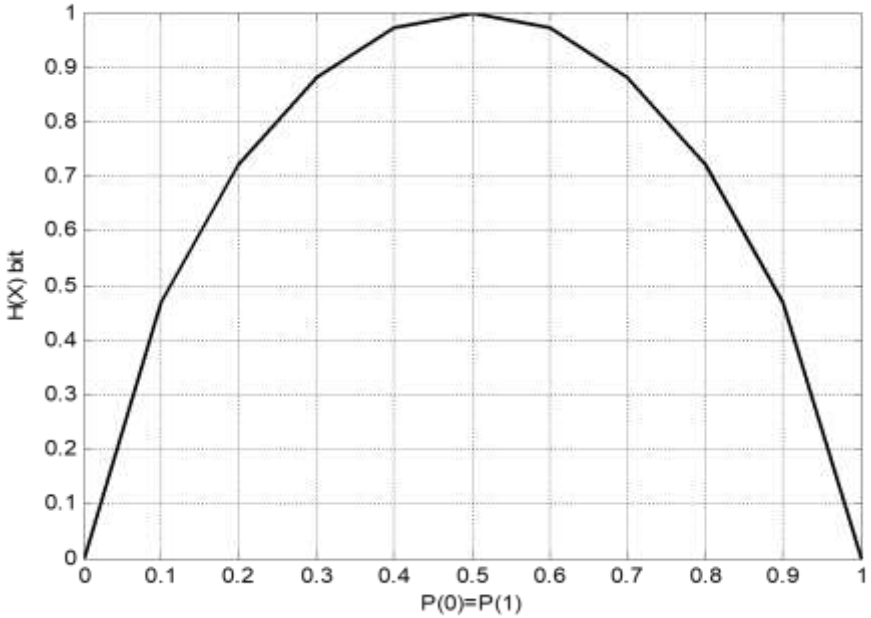


Figure 10: Entropy of binary source distribution

For any non-binary source, if all messages are equiprobable, then $P(x_i) = 1/n$ so that: $H(X) = H(X)_{max} = -[\frac{1}{n}\log_a\left(\frac{1}{n}\right)] \times n = -\log_a\left(\frac{1}{n}\right) = \log_a n \ \ bits/symbol,$ which is the maximum value of source entropy. Also, $H(X) = 0$ if one of the message has the probability of a certain event or *p(x) =1*.

9-4 Source Entropy Rate:

It is the average rate of amount of information produced per second.

$$R(X) = H(X) \times rate\ of\ producing\ the\ symbols \ \ = \frac{bits}{sec} = bps$$

The unit of H(X) is bits/symbol and the rate of producing the symbols is symbol/sec, so that the unit of R(X) is bits/sec.
Sometimes

$$R(X) = \frac{H(X)}{\bar{\tau}},$$

Where

$$\bar{\tau} = \sum_{i=1}^{n} \tau_i P(x_i)$$

$\bar{\tau}$ is the average time duration of symbols, τ_i is the time duration of the symbol x_i.

Example 1:

A source produces dots '.' And dashes '-' with P(dot)=0.65. If the time duration of dot is 200ms and that for a dash is 800ms. Find the average source entropy rate.

Solution:

$P(dash) = 1 - P(dot) = 1 - 0.65 = 0.35$

$H(X) = -[0.65\log_2(0.65) + 0.35\log_2(0.35)] = 0.934\ bits/symbol$

$\bar{\tau} = 0.2 \times 0.65 + 0.8 \times 0.35 = 0.41\ sec$

$$R(X) = \frac{H(X)}{\bar{\tau}} = \frac{0.34}{0.41} = 2.278\ bps$$

Example 2:

A discrete source emits one of five symbols once every millisecond. The symbol probabilities are 1/2, 1/4, 1/8, 1/16 and 1/16 respectively. Calculate the information rate.

Solution:

$$\mathrm{H} = \sum_{\mathrm{i}=1}^{5} \mathrm{P_i} \log_2 \frac{1}{\mathrm{p_i}}$$

$$\mathrm{H} = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{16}\log_2 16 + \frac{1}{16}\log_2 16$$

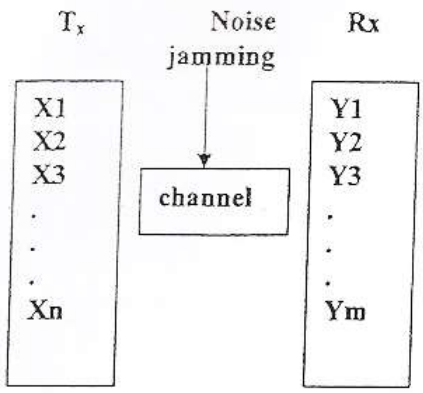
H=0.5+0.5+0.375+0.25+0.25 = 1.875 bit/symbol

$$\mathrm{R} = \frac{\mathrm{H}}{\tau} = \frac{1.875}{10^{-3}} = 1.875\ \mathrm{kbps}$$

HW:

A source produces dots and dashes; the probability of the dot is twice the probability of the dash. The duration of the dot is 10msec and the duration of the dash is set to three times the duration of the dot. Calculate the source entropy rate.

10- Mutual information for noisy channel:



Consider the set of symbols $x_1, x_2, \ldots, x_n$, the transmitter T_x my produce. The receiver R_x may receive $y_1, y_2 \ldots\ldots\ldots y_m$. Theoretically, if the noise and jamming is neglected, then the set X=set Y. However and due to noise and jamming, there will be a conditional probability $P(y_j \mid x_i)$:

1- $P(x_i)$ to be what is so called the a priori probability of the symbol x_i, which is the prob of selecting x_i for transmission.

2- $P(y_j \mid x_i)$ to be what is called the aposteriori probability of the symbol x_i after the reception of y_j.

The amount of information that y_j provides about x_i is called the mutual information between x_i and y_i. This is given by:

$$I(x_i, y_j) = \log_2\left(\frac{aposterori\ prob}{apriori\ prob}\right) = \log_2\left(\frac{P(y_j \mid x_i)}{P(x_i)}\right)$$

Properties of $I(x_i, y_j)$:

1- It is symmetric, $I(x_i, y_j) = I(y_j, x_i)$.

2- $I(x_i, y_j) > 0$ if aposteriori probability > a priori probability, y_j provides +ve information about x_i.

3- $I(x_i, y_j) = 0$ if aposteriori probability = a priori probability, which is the case of statistical independence when y_j provides no information about x_i.

4- $I(x_i, y_j) < 0$ if aposteriori probability $<$ a priori probability, y_j provides -ve information about x_i, or y_j adds ambiguity.

$$\text{Also } I(x_i, y_j) = \log_2 \left(\frac{P(x_i|y_j)}{P(y_j)} \right)$$

Example:

Show that I(X, Y) is zero for extremely noisy channel.

Solution:

For extremely noisy channel, then $y_j$gives no information about x_i the receiver can't decide anything about x_i as if we transmit a deterministic signal x_i but the receiver receives noise like signal y_j that is completely has no correlation with x_i. Then x_i and y_j are statistically independent so that $P(x_i \mid y_j) = P(x_i) and\ P(y_j \mid x_i) = P(x_i)\ for\ all\ i\ and\ j, then$:

$$I(x_i, y_j) = \log_2 1 = 0 \ \ for\ all\ i\ \&\ j, then\ I(X, Y) = 0$$

10.1 Joint entropy:

In information theory, joint entropy is a measure of the uncertainty associated with a set of variables.

$$\boldsymbol{H(X,Y) = H(XY) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i, y_j) \log_2 P(x_i, y_j) \quad bits/symbol}$$

10.2 Conditional entropy:

In information theory, the conditional entropy quantifies the amount of information needed to describe the outcome of a random variable Y given that the value of another random variable X is known.

$$H(Y \mid X) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i, y_j)\log_2 P(y_j \mid x_i) \qquad bits/symbol$$

10.3 Marginal Entropies:

Marginal entropies is a term usually used to denote both source entropy H(X) defined as before and the receiver entropy H(Y) given by:

$$H(Y) = -\sum_{j=1}^{m} P(y_j)\log_2 P(y_j) \qquad \frac{bits}{symbol}$$

10.4 Transinformation (average mutual information):

It is the statistical average of all pair $I(x_i, y_j)$, $i = 1, 2, \ldots\ldots, n,\ j = 1, 2, \ldots\ldots, m.$

This is denoted by $I(X, Y)$ and is given by:

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} I(x_i, y_j)P(x_i, y_j)$$

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log_2\left(\frac{P(y_j \mid x_i)}{P(y_j)}\right)\frac{bits}{symbol}$$

or

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log_2\left(\frac{P(x_i \mid y_j)}{P(x_i)}\right)\ bits/symbol$$

10.5 Relationship between joint, conditional and transinformation:

$$H(Y \mid X) = H(X,Y) - H(X)$$
$$H(X \mid Y) = H(X,Y) - H(Y)$$

Where, the $H(X \mid Y)$ is the losses entropy.
Also we have:

$$I(X,Y) = H(X) - H(X \mid Y)$$
$$I(X,Y) = H(Y) - H(Y \mid X)$$

Example:

The joint probability of a system is given by:

$$P(X,Y) = \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} \begin{bmatrix} 0.5 & 0.25 \\ 0 & 0.125 \\ 0.0625 & 0.0625 \end{bmatrix}$$

Find:

1- Marginal entropies.

2- Joint entropy

3- Conditional entropies.

4- The transinformation.

1- $P(X) = \begin{bmatrix} x_1 & x_2 & x_3 \\ 0.75 & 0.125 & 0.125 \end{bmatrix}$ $P(Y) = \begin{bmatrix} y_1 & y_2 \\ 0.5625 & 0.4375 \end{bmatrix}$

$$H(X) = -[0.75\ln(0.75) + 2 \times 0.125\ln(0.125)]/ln2$$

$$= 1.06127\ bits/symbol$$

$$H(Y) = -[0.5625\ln(0.5625) + 0.4375\ln(0.4375)]/ln2$$

$$= 0.9887\ bits/symbol$$

2-

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i, y_j)\log_2 P(x_i, y_j)$$

$$H(X,Y)$$

$$= -\frac{[0.5\ln(0.5) + 0.25\ln(0.25) + 0.125\ln(0.125) + 2 \times 0.0625\ln(0.0625)]}{ln2}$$

$$= 1.875 \qquad bits/symbol$$

3- $H(Y \mid X) = H(X,Y) - H(X) = 1.875 - 1.06127 = 0.813\ \frac{bits}{symbol}$

$$H(X \mid Y) = H(X,Y) - H(Y) = 1.875 - 0.9887 = 0.886\ bits/symbol$$

4- $I(X,Y) = H(X) - H(X \mid Y) = 1.06127 - 0.8863 = 0.17497\ bits/symbol.$